Autonomous Cycle of Data Analysis Tasks based on Machine Learning Techniques for the Management of Small Agroindustrial Producers

Submitted in partial fulfillment of the requirements for

the degree of Doctor of in Engineering

Jairo FUENTES

Universidad EAFIT

Escuela de Ingeniería

Medellín, Colombia

03/03/2025

Director of dissertation research:

Prof. Jose Aguilar

Prof. Edwin Montoya

Examining committee:

Dissertation committee chairperson:

President of the Board of the Doctoral Program in Engineering:

Dissertation date: 2025

Abstract

In this research, we address the use of autonomous data analysis task cycles (ACODAT) in the agro-industrial production chain. ACODAT's main objective is to adopt a data-driven perspective to solve business challenges. In the context of ACODAT, to achieve a specific objective within the process under observation, different tasks related to data analysis must be performed simultaneously. These tasks are interconnected with each other, and each one plays a unique role in the cycle: (i) observe the process, (ii) analyze and interpret what happens in it, and (iii) make decisions to guide the process towards achieving its objectives. This integration of tasks in a closed cycle allows complex problems to be efficiently addressed.

Despite the existence of advances in the design and use of ACODAT in different domains, there are no related works on using ACODAT to automate industrial production processes. To address this need, our objective was to develop methodologies, models and approaches to automate production processes using data analysis tasks in the agro-industrial production chain. Based on this objective, several challenges were proposed, the first was to specify ACODATs for the Management of Agro-industrial Micro, Small and Medium Enterprises (MSMEs), the second was to define Fuzzy Cognitive Maps (FCM) for the Assessment of the level of Digital Transformation in MSMEs, and the third was to develop other novel AI techniques to improve the agro-industrial processes of MSMEs. Several research articles were developed during this thesis to reach its objective. To meet the first challenge, two articles were proposed. The first article specified several ACODATs for the agro-industrial sector, and the second article presented an ACODAT for the determination of the Coffee Production Process for MSMEs. For the second challenge, three articles were produced, the first article is on the evaluation of the level of digital transformation in MSMEs through FCMs based on experts, a second article on the explainability analysis of the previous evaluation model based on FCMs. Finally, the third article is on a model of evaluation of the digital transformation in MSMEs based on learned FCMs using data. Regarding the third challenge, three articles were proposed. The first article defines a Meta-Learning Architecture for ACODATs to automate coffee production. The second article presents a Meta-Learning Architecture for ACODAT in the context of agro-industrial production chains of MSMEs, and a third article on the analysis of quantum support vector machines in classification problems. In each article, the strategies/models were evaluated using various datasets. The results show the capacity of the developed methodologies and models for decision making for the intelligent management of the production chain using ACODAT.

In summary, the ACODATs allowed the automation of some of the processes in the production chains of agro-industrial MSMEs, the evaluation models based on FCM allowed the analysis of the level of Digital transformation in MSMEs and explained what actions to take to improve it, if applicable, while meta-learning and quantum computing allowed improvements and adaptations to be made in the machine learning models based on data of ACODATs.

Keywords: Artificial Intelligence, Machine Learning, Autonomous Cycles of Data Analysis tasks, Fuzzy Cognitive Maps, Quantum Computing, Meta-learning, Production and Automation Processes in MSMEs.

Resumen

En esta investigación abordamos el uso de ciclos de tareas de análisis autónomo de datos (ACODAT) en la cadena de producción agroindustrial. El objetivo principal de ACODAT es adoptar una perspectiva basada en datos para resolver desafíos de negocio. En el contexto de ACODAT, para lograr un objetivo específico dentro del proceso en observación, se deben realizar simultáneamente diferentes tareas relacionadas con el análisis de datos. Estas tareas están interconectadas entre sí, y cada una juega un papel único en el ciclo: (i) observar el proceso, (ii) analizar e interpretar lo que sucede en él, y (iii) tomar decisiones para guiar el proceso hacia el logro de sus objetivos. Esta integración de tareas en un ciclo cerrado permite abordar de manera eficiente problemas complejos.

A pesar de la existencia de avances en el diseño y uso de ACODAT en diferentes dominios, no existen trabajos relacionados sobre el uso de ACODAT para automatizar procesos de producción industrial. Para abordar esta necesidad, nuestro objetivo fue desarrollar metodologías, modelos y enfoques para automatizar procesos de producción utilizando tareas de análisis de datos en la cadena de producción agroindustrial. Con base en este objetivo, se propusieron varios retos, el primero fue especificar ACODATs para la Gestión de Micro, Pequeñas y Medianas Empresas (MIPYMES) Agroindustriales, el segundo fue definir Mapas Cognitivos Difusos (MCD) para la Evaluación del nivel de Transformación Digital en MIPYMES, y el tercero fue desarrollar otras técnicas novedosas de IA para mejorar los procesos agroindustriales de las MIPYMES. Durante esta tesis se desarrollaron varios artículos de investigación para alcanzar su objetivo. Para cumplir con el primer reto, se propusieron dos artículos. El primer artículo especificó varios ACODAT para el sector agroindustrial, y el segundo artículo presentó un ACODAT para la determinación del Proceso de Producción de Café para MIPYMES. Para el segundo reto, se produjeron tres artículos, el primer artículo es sobre la evaluación del nivel de transformación digital en MIPYMES a través de MCD basados en expertos, un segundo artículo sobre el análisis de explicabilidad del modelo de evaluación anterior basado en MCD. Finalmente, un tercer artículo es sobre un modelo de evaluación de la transformación digital en MIPYMES basado en MCD aprendidos utilizando datos. Respecto al tercer reto, se propusieron tres artículos. El primer artículo define una Arquitectura de Meta-Aprendizaje para ACODATs para automatizar la producción de café. El segundo artículo presenta una Arquitectura de Meta-Aprendizaje para ACODAT en el contexto de cadenas productivas agroindustriales de MIPYMES, y un tercer artículo sobre el análisis de máquinas de vectores de soporte cuántico en problemas de clasificación. En cada artículo se evaluaron las estrategias/modelos utilizando diversos conjuntos de datos. Los resultados muestran la capacidad de las metodologías y modelos desarrollados para la toma de decisiones para la gestión inteligente de la cadena productiva utilizando ACODAT.

En resumen, los ACODATs permitieron automatizar algunos de los procesos en las cadenas productivas de MIPYMES agroindustriales, los modelos de evaluación basados en FCM permitieron analizar el nivel de Transformación Digital en las MIPYMES y explicar qué acciones tomar para mejorarlo en caso de ser el caso, mientras que el meta-aprendizaje y la computación cuántica permitieron realizar mejoras y adaptaciones en los modelos de aprendizaje automático basados en datos de ACODATs.

Palabras clave: Inteligencia Artificial, Machine Learning, Ciclos Autónomos de Tareas de Análisis de Datos, Mapas Cognitivos Difusos, Computación Cuántica, Meta aprendizaje, Procesos de Producción y Automatización en MIPYMES.

Scientific contributions

Several scientific articles were generated and published during the development of this research project.

Articles published in journals indexed in SJR:

- Fuentes, J.: Aguilar, J, Montoya, E. Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs, Journal of Industrial Information Integration. Volume 44, 2025, 100788, ISSN 2452-414X, https://doi.org/10.1016/j.jii.2025.100788. **(Q1).**
- Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector. *Information,* 15(2), 2024. https://doi.org/10.3390/info15020086 **(Q2)**
- Fuentes, J.: Aguilar, J.; Montoya, E.; Hoyos, W.: Benito, D. Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps: Explainability Analysis on Fuzzy Cognitive Maps, *CLEI Electronic Journal*, 27(2), 2024 https://doi.org/10.19153/cleiej.27.2.2 **(Q4)**

Articles published in conference Proceedings

- Fuentes, J.; Aguilar, J.; E Montoya, E.; Hoyos, W. Evaluation of the Level of Digital Transformation in MSMEs Using Fuzzy Cognitive Maps Based on Experts," *XLIX Latin American Computer Conference (CLEI),* La Paz, Bolivia, https://doi.org/10.1109/CLEI60451.2023.10346161
- Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. Analysis of Quantum Support Vector Machines in Classification Problems, *50 Latin American Computer Conference (CLEI),* Bahia Blanca, Argentina, 2024. 10.1109/CLEI64178.2024.10700573

Articles currently submitted to high-impact journals (under review):

- J. Fuentes, J. Aguilar, W. Hoyos, "Assessment of the Digital Transformation Level in MSMEs through Trained Fuzzy Cognitive Maps", *Applied Soft Computing.*
- J. Fuentes, J. Aguilar, "A systematic literature review on the management of small Agroindustrial producers using artificial intelligence", *International Journal of Business Performance and Supply Chain Modelling*.
- J. Fuentes, J. Aguilar, "Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs." *International Journal of Data Science and Analytics.*
- J. Fuentes, R. Dos Santos, J. Aguilar, D. Shi, Meta-learning Architecture for ACODAT in the Context of Agro-Industrial Production Chains of MSMEs, *Information Processing in Agriculture.*
- R. Dos Santos, J. Aguilar, J. Fuentes, An Autonomous Meta-Learning Architecture based on Transfer Learning and Linked Data, *Expert Systems With Applications.*

Project context

This study was partially funded by Colombian Administrative Department of Science, Technology and Innovation - COLCIENCIAS (grant number 111572553478) (J. Aguilar) and Colombian Ministry of Science and Technology Bicentennial PhD Grant (J. Fuentes).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

I thank God who has guided me in this stage of my life and has given me the wisdom and intelligence to achieve this goal.

Sincere gratitude for the support, collaboration and contribution of all those wonderful people who exist in my life. To my mother Amelia Camargo, to my wife Rosa Afanador, the most important people in my life. My vital support throughout this process.

With deep esteem and recognition, I extend my most sincere gratitude to my tutor Dr. JOSE AGUILAR, his vision and energy for research were contagious throughout this process. It has been an honor to have worked under his supervision.

I would like to thank my co-directors Mauricio Toro and Edwin Montoya for their commitment, guidance and feedback in this process. Raul Toscano, William Hoyos and Rodrigo Garcia for supporting and contributing to my research.

I would like to gratefully acknowledge Colciencias and the U. EAFIT and its entire organization for their support throughout this period.

Content

Chapter 1. Introduction and research context

1.1 Problem statement and motivation.

The agro-industrial sector is characterized by producing food and non-food products, obtained through raw materials normally of biological origin, which it processes. Currently, there is a considerable number of companies in the agro-industrial sector that lack smart technology to improve their production processes, which could be used in their transformation, conservation and marketing processes, among others [1]. Specially, small agro-industrial producers make up a significant portion of companies and play a vital role in a country's economic and social advancement. However, it is crucial that they integrate technologies such as Artificial Intelligence (AI) to use their process data to build knowledge models that allow them to make decisions about their production chains to achieve high levels of performance and competitiveness.

The work seeks to provide MSMEs with a computational environment based on ACODATs deployed as services, so that they can be invoked during the processes of handling, preserving and transforming raw materials. Thus, the computational environment of autonomous cycles will be integrated into each agro-industrial process, to help it in its management with the adequate knowledge to improve the processing, storage, distribution and marketing activities of agro-industrial products. In this way, the system will allow agro-industrial companies to strengthen and evolve using AI techniques that allow companies to extract knowledge from their data repositories and other available sources, to improve the management of their production processes.

Therefore, this is a way to introduce digital transformation in MSMEs to take advantage of digital technologies in organizations through the digitalization of data and the automation of their operations. It is a way to promote the modernization of the MSME sector in the agro-industrial field, in our case, using ACODATs within companies, which allow them to innovate in their processes and products. This will enable generating added value to the product offering of agro-industrial companies, so that differentiation is the basic element of competitiveness. Specifically, *the objective is to develop ACODATs based on machine learning techniques for the management of agro-industrial MSMEs*.

ACODATs use machine learning techniques to develop knowledge models from their data, and other AI techniques to exploit expert knowledge. These models will allow decisions to be made that improve the management and innovation capabilities of companies and, consequently, their competitiveness. Some of these techniques used are FCMs, some based on experts and others on data, which were used to build a business digital transformation evaluation model. Predictive models using machine learning techniques and explainability analysis methods were also used to help MSMEs understand their results. Finally, the meta-learning paradigm was used to allow ACODATs to self-manage, and quantum computing to improve the quality of machine learning models.

1.2 Research objectives

1.2.1 General objective

Apply Artificial Intelligence (AI) to improve agro-industrial processes MSMEs

1.2.2 Specific objectives

Several specific objectives are derived from the general objective of the thesis:

- Analyze the Machine Learning Techniques used in the MSMEs Management.
- Specify Autonomous Cycles of Data Analysis Tasks for the Management of Agroindustrial MSMEs.
- Develop novel AI techniques to improve the agro-industrial processes of MSMEs.
- Implement a case study in a Colombian MSMEs

1.3. Contributions and research scope

This research focuses on the development of models, methodologies and computational techniques to support and improve the tasks of an agro-industrial production chain. Several contributions were made. The first contribution of this research was a systematic literature review (SLR) on the management of small agro-industrial producers using AI. In this review, we focused on three trends in technologies for the management of agro-industrial processes: i) Data Management, ii) Machine Learning and iii) and Cyber-Physical Systems. For each of these trends, we identified challenges and opportunities for future work.

Based on the information reported in the SLR, we set as the main objective to specify ACODATs. Then, we apply AI techniques on datasets related to the management of the agro-industrial production chain to generate diagnostic, predictive and prescriptive models with explanatory capacity. Next, ACODATs were developed that combine these diagnostic, predictive and prescriptive models to improve the processes of a production chain. In literature, there are few works that implement ACODATs in the production chain. In the ACODATs, each task interacts with the others and has different functions to achieve the objective of the cycle.

Particularly, three industrial automation processes were identified within the production chain, in which autonomous cycles were applied. The first cycle is responsible for identifying the type of input to be transformed, such as quantity, quality, time and cost, based on information from the organization and its context. The second cycle selects the technological level to be used in the transformation of the raw material, characterizing the processing platform of the plant. The last cycle identifies the main characteristics of the generated product, such as the quality and value of the product. Finally, the first autonomous cycle was applied in a coffee factory.

On the other hand, FCMs based on experts and data were implemented for the evaluation of the level of digital transformation in MSMEs. The main digital transformation variables used to define our FCMs were classified into five groups, based on the COBIT standard [3]. Our model was able to infer the level of digital transformation of the organizations studied. In addition, the explanatory capacity of the FCMs developed in this work was analyzed using different explainability methods, some generals and others specific to the FCMs. The explainability techniques (very different from each other) allowed analyzing in depth the behavior of the variables in the results obtained, something very important to understand how to improve the levels of digital transformation in organizations. Other innovative AI techniques used were quantum computing and the meta-learning paradigm. Particularly, Quantum Support Vector Machines (QSVM) were used in classification problems. The purpose of this research was to analyze the use QSVM for different types of classification problems. To do so, several kernels (linear and Gaussian) were considered. In turn, a comparison with Random Forest was made. Also, a detailed analysis was made for the binary classification of coffee beans. Finally, a Meta Learning architecture was designed to adapt ACODATs, capable of performing model transfer, data transfer or parameter transfer processes, or invoking synthetic data generation procedures, to define the new knowledge models of the tasks of new ACODATs. Different methodologies were used to develop the ACODATs or knowledge models, to take advantage of the knowledge of experts and the datasets of coffee agro-industrial companies. Specifically, for the specification of the ACODATs, MIDANO was used for the definition of autonomous data analysis task cycles [2]. For the conception of data-based knowledge models, CRISP-DM was used, a widely used standard methodology for data mining and data analysis projects [3]. Finally, for the development of the FCM-based models, the methodology proposed in [4] was used. Implementations of the different ACODAT models were made using specific scenarios of the coffee company "Café Galavis" located in the city of Cúcuta (Norte de Santander), which were validated using the appropriate metrics for each case. The deployment of the models in mobile, web and desktop applications is beyond the scope of this work. All the contributions made in this research are described in research articles. A total of nine (9) scientific articles were generated, of which three (4) are published and the other five (5) are under review. Each of the articles is briefly explained below. A first systematic review article of the literature on the management of small agro-industrial producers using AI was generated. The second research article corresponds to the design of ACODATs for the automation of the production chain of MSMEs in the agro-industrial sector. The third article presents an ACODAT to automate the Coffee Production Process in MSMEs. The fourth article describes an Evaluation Model of the Level of Digital Transformation in MSMEs through FCMs based on Experts. The fifth article performs an Explainability Analysis of the Evaluation Model Based on FCMs. The sixth article defines an Evaluation Model of the Level of Digital Transformation in MSMEs using FCMs based on machine learning mechanisms. The seventh article designs a General Meta-Learning Architecture for knowledge models of coffee production in MSMEs. The eighth article creates a Meta-Learning Architecture for ACODAT in the context of agro-industrial production chains of MSMEs. Finally, the last article analyzes QSVMs in classification problems.

1.4 Thesis organization

This thesis is presented as a collection of articles developed to meet each of the proposed objectives. Chapter 2 corresponds to the fulfillment of the first objective. Chapter 3 corresponds to the fulfillment of the second objective. Chapter 4 and 5 correspond to the fulfillment of the third objective, while the 4th objective is scattered between chapters 3, 4 and 5 as a case study. One article was generated for the first objective, two articles were generated for the second objective, 6 articles were generated for the third objective, while the 4th objective is present in 8 of the articles as a case study. The articles will be presented in each section with the title, Doi (if published), abstract and full text.

Chapter 2. A systematic literature review on the management of small agroindustrial producers using artificial intelligence

2.1 Motivation

In this chapter, we present an SLR that reviewed trends in AI in the agro-industrial sector, and particularly, the use of machine learning in its production processes. Among the most relevant trends, the use of Machine Learning to build knowledge models for diagnostic and prediction tasks, mainly, was identified. It was also identified that there are no ACODATs reported in the literature to automate the productive processes of this. With this article, we fulfill objective one. The full article can be found in Appendix A.

2.2 Identification of the article

J. Fuentes, J. Aguilar. “A systematic literature review on the management of small agroindustrial producers using artificial intelligence”, submitted to publication, *International Journal of Business Performance and Supply Chain Modelling*

2.3 Abstract

Small agro-industrial producers make up a significant portion of companies and play a vital role in a country's economic and social advancement. However, it is crucial that they integrate technology and become agro-industrial producers. Artificial Intelligence (AI) can use process data to build knowledge models that allow agro-industrial producers to make decisions about production chains to achieve high levels of performance and competitiveness. The objective of this work is to perform a systematic literature review on the use of AI for the management of small agro-industrial productions. Seven scientific digital libraries (ACM, Web of Sciences, Scopus, Google Scholar, IEEE Xplore, Sciences Direct and EBSCO) were used to investigate the use of AI for the management of small agribusinesses. A search methodology based on research questions, inclusion, and exclusion criteria, among other aspects, was used to select the most relevant articles. An analysis of sixty-two articles was performed to assess the strengths and limitations of using AI in Small Agribusiness Production Management, as well as to identify the primary challenges

and opportunities. In the systemic review of the literature, some studies based on AI for the management of small agro-industrial productions are evidenced. Among the limitations found are the low investment in new technologies, and the rejection of changes and innovative business models, very common aspects among Micro, Small & Medium Enterprises (MSMEs). The results of this systematic review offer significant insights for future studies on the application of AI in the production chain of small-scale agro-industrial producers.

2.4 Link to the full article

See Appendix A

Chapter 3. Autonomous cycles of data-analysis tasks for the analysis of the agroindustrial sector.

In this chapter, ACODATs are implemented for the automation of production chains in MSMEs in the agro-industrial sector. Section 3.1 presents an article that proposes an analysis of the production processes in MSMEs in the agro-industrial sector, and which ones can be automated using ACODATs, with the specification of the latter. Then, a particular case is presented in Section 3.2. for the coffee production process, where these cycles allow the analysis of critical data related to the production, processing and quality of the coffee. With these articles, we meet subobjectives two and four.

3.1 Autonomous cycles of data-analysis tasks for the automation of the production chain of MSMEs for the agroindustrial sector.

3.1.1 Motivation

Autonomous cycles of data analysis tasks are proposed for the automation of production chains, with the aim of improving the productivity of Micro, Small and Medium Enterprises (MSMEs) in the context of agroindustry. In autonomous cycles of data analysis tasks, each task interacts with the others and has different functions, to achieve the objective of the cycle. This article identifies three industrial automation processes within the production chain, in which autonomous cycles can be applied. The first cycle is responsible for identifying the type of input to be transformed, such as quantity, quality, time and cost, based on information from the organization and its context. The second cycle selects the technological level used in the transformation of the raw material, characterizing the processing platform of the plant. The last cycle identifies the level of specialization of the generated product, such as the quality and value of the product. Finally, the first autonomous cycle is applied to define the type of input to be transformed in a coffee factory. The whole article is in Appendix B.

3.1.2 Identification of the article

J. Fuentes; J. Aguilar; E. Montoya; A. Pinto. "Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector." Information, 2024, 15, 86. https://doi.org/10.3390/info15020086

3.1.3 Abstract

In this paper, we propose autonomous cycles of data analysis tasks for the automation of the production chains aimed to improve the productivity of Micro, Small and Medium Enterprises (MSMEs) in the context of agroindustry. In the autonomous cycles of data analysis tasks, each task interacts with the others and has different functions, in order to reach the goal of the cycle. In this article, we identify three industrial-automation processes within the production chain, in which autonomous cycles can be applied. The first cycle is responsible for identifying the type of input to be transformed—such as quantity, quality, time, and cost—based on information from the organization and its context. The second cycle selects the technological level used in the raw-material transformation, characterizing the platform of plant processing. The last cycle identifies the level of specialization of the generated product, such as the quality and value of the product. Finally, we apply the first autonomous cycle to define the type of input to be transformed in a coffee factory.

3.1.4 Link to the full article

https://www.mdpi.com/2078-2489/15/2/86

3.2 Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs

3.2.1 Motivation

Coffee production requires certain levels of efficiency to ensure that the quality of the bean, the roasting process and, in general, the coffee processing methods, achieve financial and environmental sustainability objectives. This requires monitoring and analysis of coffee bean characteristics and the roasting process, among other aspects, so that the actors in the MSME agro-industrial sector can know what happens in coffee production and can make better decisions to improve it. This work aims to instantiate the autonomous cycle in charge of identifying the type of input to be transformed in the production process, in the case of coffee production. This cycle analyzes the inputs of the production chain (quantity, quality, seasonality, durability, cost, etc.), based on information from the organization and the context, to establish the production process to be conducted. This autonomous cycle is instantiated in coffee production to identify the type of input to be transformed (bean quality) and determine the transformation process (level of decrease in the bean during the roasting process and coffee processing method). The entire article is in Appendix C.

3.2.2 Identification of the article

J Fuentes, J Aguilar, E Montoya, Autonomous cycle of data analysis tasks for the determination of the coffee productive process for MSMEs, Journal of Industrial Information Integration, Volume 44, 2025, 100788, ISSN 2452-414X, https://doi.org/10.1016/j.jii.2025.100788.

3.2.3 Abstract

Coffee production needs certain levels of efficiency to ensure that the quality of the bean, the roasting process, and in general, the coffee processing methods, achieve financial and environmental sustainability objectives. This requires tasks of monitoring and analyzing of features of the coffee bean, and the roasting process, among other aspects, so that stakeholders of the agro-industrial sector of MSMEs can know what happens in the coffee production and can make better decisions to improve it. In a previous article, three autonomous cycles of data analysis tasks are proposed for the automation of the production chains of the MSMEs. This work aims to instantiate the autonomous cycle responsible for identifying the type of input to transform in the production process, in the case of coffee production. This cycle analyzes the inputs of the production chain (quantity, quality, seasonality, durability, cost, etc.), based on information from the organization and the context, to establish the production process to be carried out. This autonomous cycle is instanced in coffee production to identify the type of input to transform (bean quality), and to determine the transformation process (level of decrease of the bean during the roasting process and coffee processing method). The quality model is defined by the K-means technique with a performance in the Silhouette Index of 0.85, the predictive model of the level of decrease of beans in the roasting process is defined by Random Forest with a performance in the accuracy of 0.81, and finally, the identification model

of the "production method" is carried out by the Logistic Regression technique with a quality performance in the accuracy of 0.72.

3.2.4 Link to the full article

https://www.sciencedirect.com/science/article/abs/pii/S2452414X25000123

Chapter 4. Using Fuzzy Cognitive Maps for the Assessment of the level of Digital Transformation in MSMEs

Digital transformation enables the use of digital technologies to innovate and generate value in organizations by digitizing data and automating processes. This chapter proposes the use of FCMs based on data and expert knowledge to develop models for the evaluation of the level of digital transformation in MSMEs. In section 4.1, FCMs based on experts are proposed. The key variables, organized into five groups according to the COBIT standard [3], are used to define the FCMs. Likewise, in section 4.2., an Explainability Analysis of the evaluation model based on FCMs is presented. Explainability methods help to analyze the current situation of the key factors related to digital transformation in these companies, supporting MSMEs in their digital adaptation process. Finally, in section 4.3, an evaluation model based on FCMs trained with data is proposed. With these articles, we meet subobjectives three and four.

4.1. Evaluation of the Level of Digital Transformation in MSMEs using Fuzzy Cognitive Maps based on Experts.

4.1.1 Motivation

The concept of digital transformation involves exploiting digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models, and services that produce value based on the digitization of data and processes. The application of digital technologies enables organizations to develop capabilities for innovation, automation, etc., utilizing both established and emerging technologies, including the widespread use of artificial intelligence. In this paper, the implementation of a fuzzy cognitive map based on experts for the evaluation of the level of digital transformation in MSMEs (Micro, Small & Medium Enterprises). The main variables of digital transformation used to define our fuzzy cognitive map were classified into five types: i) Organization and Culture variables related to strategies, way of working, and ecosystems, ii) Customer variables related to services and digital channels and products, iii) Operations and Internal Processes variables related to supply chain, suppliers, business model, iv) Information

Technologies variables related to innovation, digitization, data and analytic. Finally, the fifth type of variable is the target, which indicates the level of digital transformation of the organization. The entire article is in Appendix D.

4.1.2 Identification of the article

J. Fuentes, J. Aguilar, E. Montoya and W. Hoyos, "Evaluation of the Level of Digital Transformation in MSMEs Using Fuzzy Cognitive Maps Based on Experts," Proceeding *XLIX Latin American Computer Conference (CLEI*), La Paz, Bolivia, 2023, https://doi.org/10.1109/CLEI60451.2023.10346161

4.1.3 Abstract

The concept of digital transformation involves exploiting digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models, and services that produce value based on the digitization of data and processes. The application of digital technologies enables organizations to develop capabilities for innovation, automation, etc., utilizing both established and emerging technologies, including the widespread use of artificial intelligence. In this paper, the implementation of a fuzzy cognitive map based on experts for the evaluation of the level of digital transformation in MSMEs (Micro, Small & Medium Enterprises). The main variables of digital transformation used to define our fuzzy cognitive map were classified into five types: i) Organization and Culture variables related to strategies, way of working, and ecosystems, ii) Customer variables related to services and digital channels and products, iii) Operations and Internal Processes variables related to supply chain, suppliers, business model, iv) Information Technologies variables related to innovation, digitization, data and analytic. Finally, the fifth type of variable is the target, which indicates the level of digital transformation of the organization. Our model managed to specify with 99.4% the level of digital transformation of the organization. Thus, our fuzzy cognitive map can predict and analyze the factors associated with digital transformation in MSMEs.

4.1.4 Link to the full article

https://ieeexplore.ieee.org/document/10346161

4.2. Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps

4.2.1 Motivation

The concept of digital transformation involves the exploitation of digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models and services that produce value from the digitalization of data and processes. The application of digital technologies allows organizations to develop innovation, automation, etc. capabilities, using both established and emerging technologies, including the widespread use of artificial intelligence. This paper proposes the implementation of expert- and data-based Fuzzy Cognitive Maps (FCMs) for the assessment of the level of digital transformation in MSMEs (Micro, Small and Medium Enterprises). Additionally, this paper performs an explainability analysis of the FCM-based assessment models. The main digital transformation variables used to define our FCMs were classified into five groups, based on the COBIT standard. In addition, the explanatory capacity of the FCMs developed in this paper was explored using different explainability methods, some general and others specific to the FCMs. Overall, the results obtained in the work are very encouraging since the quality metrics obtained with the evaluation models are very good, almost always higher than 90%; and the explanations obtained with the explainability methods allow an in-depth analysis of the behavior of the variables in the results obtained, something very important to understand how to improve the levels of digital transformation in organizations. The entire article is in Appendix E.

4.2.2 Identification of the article

J. Aguilar, J. Fuentes, E. Montoya, W. Hoyos, D. Benito. "Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps." Vol. 27 No. 2 (2023): CLEI 2023 best papers Vol. I. https://doi.org/10.19153/cleiej.27.2.2

4.2.3 Abstract

The concept of digital transformation involves exploiting digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models, and services that produce value based on the digitization of data and processes. The application of digital technologies enables organizations to develop capabilities for innovation, automation, etc., utilizing both established and emerging technologies, including the widespread use of artificial intelligence. This article proposes the implementation of Fuzzy Cognitive Maps (FCMs) based on experts and data for the evaluation of the level of digital transformation in MSMEs (Micro, Small and Medium Enterprises). Additionally, this work carries out an explainability analysis of the evaluation models based on FCMs. The main digital transformation variables used to define our FCMs were classified into five groups, based on the COBIT standard: i) Organization and Culture variables related to strategies, way of working, and ecosystems, ii) Customer variables related to services and digital channels and products, iii) Operations and Internal Processes variables related to supply chain, suppliers, and business model, iv) Information Technologies variables related to innovation, digitization, data and analytic. Finally, the fifth type of variable is the target, which indicates the level of digital transformation of the organization. Our model managed to specify with 99.4% the level of digital transformation of the organization. Furthermore, the explanatory capacity of the FCMs developed in this work was explored using different explainability methods, some general and others specific to the FCMs. In general, the results obtained in the work are very encouraging since the quality metrics obtained with the evaluation models are very good, almost always higher than 90%; and the explanations obtained with the explainability methods allow for an in-depth analysis of the behavior of the variables in the results obtained, something very important to understand how to improve the levels of digital transformation in organizations.

4.2.4 Link to the full article

https://www.clei.org/cleiej/index.php/cleiej/article/view/648

4.3. Assessment of the Digital Transformation Level in MSMEs through Trained Fuzzy Cognitive Map

4.3.1 Motivation

The importance of digital transformation lies in leveraging digital technologies to foster novel approaches within organizations, leading to the creation of innovative processes, models, and services. This digitalization of data and operations generates significant value and enables organizations to embrace innovation, automation, and various capabilities, spanning both well-established technologies and cutting-edge advancements, such as artificial intelligence. In this research, two innovative fuzzy cognitive maps, one expert-led and one data-driven, are proposed to assess the degree of digital transformation in micro, small, and medium-sized enterprises. Accordingly, our fuzzy cognitive map is competent to forecast and study the factors linked to digital transformation in SMEs. Furthermore, the paper conducts a comprehensive explainability analysis to help SMEs understand their current situation regarding digital transformation based on the FCM results. The entire article is in Appendix F.

4.3.2 Identification of the article

J. Fuentes, J. Aguilar, W. Hoyos. "Assessment of the Digital Transformation Level in MSMEs through Trained Fuzzy Cognitive Maps", submitted to publication, *Applied Soft Computing.*

4.3.3 Abstract

The essence of digital transformation lies in harnessing digital technologies to foster novel approaches within organizations, leading to the creation of innovative processes, models, and services. This digitalization of data and operations generates significant value and empowers organizations to embrace innovation, automation, and diverse capabilities, encompassing both well-established technologies and cutting-edge advancements, like artificial intelligence. In this research, two innovative fuzzy cognitive maps are proposed, one led by experts and the other based on data to assess the degree of digital transformation in Micro, Small and Medium Enterprises. The essential digital transformation variables are classified into five distinct groups: i) Organization and Culture, which includes strategies, work methods, and ecosystems, ii) Customer, including services, digital channels, and products, iii) Operations and Internal Processes, which encompasses the supply chain, suppliers, and business model, and iv) Information Technology, comprising innovation, digitization, data, and analytics. The last category of variables is the target, which signifies the digital transformation level of the organization. Remarkably, our model demonstrated an impressive accuracy of 99.4% in determining an organization's digital transformation level. Consequently, our fuzzy cognitive map is proficient in forecasting and studying the factors linked to digital transformation in MSMEs. Additionally, the paper carries out a comprehensive explainability analysis to assist MSMEs in comprehending their present situation regarding digital transformation based on the FCMs results.

4.3.4 Link to the full article

See Appendix F.

Chapter 5. Advances in AI applied to MSMEs in the agroindustrial sector

This chapter presents other innovative AI techniques to improve production processes for the MSME agro-industrial sector. Section 5.1 presents meta-learning architectures for the MSME agro-industrial sector. Section 5.2 analyzes the use of QSVM in different types of classification problems, considering various kernels and comparing it with Random Forest. With this article, we finish covering objectives three and four.

5.1 Meta-learning Architecture for analysis in the agroindustrial sector.

In the agro-industrial sector of MSMEs, continuous monitoring and analysis of their products and processes is of vital importance to optimize production. In section 5.1.1., a meta-learning architecture is proposed to analyze knowledge models in a production chain. Depending on the similarity of the dataset with which the new knowledge model is to be built with the previous datasets, with respect to previous datasets existing in the architecture, the system can reuse an existing model, transfer parameters, transfer data or develop a model from scratch. In turn, in section 5.1.2, this meta-learning architecture is extended to manage ACODATs of production chains. In both works, a case study is carried out in a coffee company.

5.1.1 Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs

5.1.1.1 Motivation

In the coffee industry, efficiency is key to achieving quality goals and ensuring sustainable practices, both financially and environmentally. In the MSME agro-industrial sector, it is essential to continuously monitor and analyze coffee bean characteristics and the roasting process to improve production. This paper proposes a meta-learning (MTL) architecture that incorporates autonomous data analysis cycles in production chains. The MTL framework includes a meta-model table that stores information about pre-trained machine learning models, allowing the appropriate model to be selected when a new data set is introduced. Based on the similarity of the new data set to previous ones, the system can reuse an existing model, perform parameter transfer, build a new model, or transfer data and create a model from scratch. The paper explores case studies in coffee companies, showing that the system is effective in quality assessment, waste reduction, and process selection, achieving an average accuracy of 90.45%. The entire article is in Appendix G.

5.1.1.2 Identification of the article

J. Fuentes, J. Aguilar. “Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs”, submitted to publication, *International Journal of Data Science and Analytics.*

5.1.1.3 Abstract

In the coffee industry, efficiency is a determining factor in achieving product quality objectives, as well as ensuring sustainable practices, both from a financial and environmental point of view. On the other hand, in the MSME agroindustrial sector, it is crucial to carry out constant monitoring and detailed analysis of the characteristics of the coffee bean and the roasting process to identify areas for improvement in production. In a previous study, three autonomous cycles of data analysis tasks are described to automate operations in the production chains of MSMEs. The proposal of this article consists of a meta-learning (MTL) architecture for autonomous cycles. The proposed MTL framework has a meta-models table that stores relevant information about pre-trained machine learning models for autonomous cycles. This meta-model table is used to identify the most appropriate machine learning model each time a new coffee dataset is introduced into an organization. Specifically, the framework can do several things, use one of the models in the meta-models table on the new dataset if this new dataset is similar enough to one previously used to make that model, do a parameter transfer, and construct a new model if the similarity with another dataset is lower, or perform a data transfer and build a new model if that similarity is even lower. This article explores the use of this MTL framework in some case studies, using records of coffee roasting processes from different coffee companies. The results indicate the great effectiveness of our system in reusing/adjusting/creating machine learning models in various situations such as evaluating coffee quality, reducing waste, and selecting coffee processing processes, reaching an accuracy of 90.45% on average in the results.

5.1.1.4 Link to the full article

See Appendix G.

5.1.2 Meta-learning Architecture for ACODAT in the Context of Agro-Industrial Production Chains of MSMEs

5.1.2.1 Motivation

In the coffee industry, efficiency is key to achieving quality goals and ensuring sustainable practices, both financially and environmentally. In the MSME agro-industrial sector, it is essential to continuously monitor and analyze coffee bean characteristics and the roasting process to improve production. This paper proposes a meta-learning (MTL) architecture that incorporates autonomous data analysis cycles in production chains. The MTL framework includes a meta-model table that stores information about pre-trained machine learning models, allowing the appropriate model to be selected when a new data set is introduced. Based on the similarity of the new data set to previous ones, the system can reuse an existing model, perform parameter transfer, build a new model, or transfer data and create a model from scratch. The paper explores case studies in coffee companies, showing that the system is effective in quality assessment, waste reduction, and process selection, achieving an average accuracy of 90.45%. The entire article is in Appendix H.

5.1.2.2 Identification of the article

J. Fuentes, Dos Santos, J. Aguilar, D. Shi. “Meta-learning Architecture for ACODAT in the Context of Agro-Industrial Production Chains of MSMEs.” Submitted to publication, *Information Processing in Agriculture.*

5.1.2.3 Abstract

In the MSME agribusiness sector, constant monitoring and detailed analysis of product characteristics are essential to identify areas where production can be improved. The proposal of this work is a meta-learning (MTL) architecture that is integrated into autonomous cycles of data analysis tasks (ACODATs) that automate the production process of MSMEs to allow their self-adjustment. The proposed MTL framework includes a metamodel knowledge base that stores information about existing machine learning models, the machine learning techniques used to create them, and with which dataset they were created. Thus, the architecture has meta-data and meta-techniques that describe previous datasets and techniques used to build the models described in the metamodel. These are used to select the most appropriate machine learning model for each ACODAT task when it is instantiated on a new dataset. Particularly, from the information contained in the MTL framework, the system is capable of performing model transfer, data transfer or parameter

transfer processes, or invoking procedures for the generation of systematic data, in order to define the new machine learning models for the tasks of the new ACODAT. The framework has been tested in a MSME in the coffee area. In the context of the case study, an autonomous cycle for a coffee company composed of three tasks to control its production process is instantiated. The proposed architecture demonstrated its ability to adapt the ACODAT to a new scenario, defined by a new set of data, obtaining very satisfactory results for each of its tasks.

5.1.2.4 Link to the full article

See Appendix H.

5.2 Analysis of Quantum Support Vector Machines in Classification Problems

5.2.1 Motivation

A common problem solved by supervised learning techniques is classification, which includes variants such as binary, multiclass, and multilabel classification. Developing techniques that perform well on different types of classification problems and for datasets with diverse characteristics remains a challenge. Recent research suggests that machine learning can benefit from quantum computing to reduce computational complexity and improve performance. This research analyzes the use of Quantum Support Vector Machines (QSVM) for different types of classification problems, using various kernels (linear and Gaussian). In addition, the performance of QSVM is compared with the Random Forest algorithm. The entire article is in Appendix I.

5.2.2 Identification of the article

J. Fuentes, J. Aguilar, A. Pinto. “Analysis of Quantum Support Vector Machines in Classification Problems", Proceeding *L Latin American Computer Conference* (CLEI), Buenos Aires, Argentina, 2024, 10.1109/CLEI64178.2024.10700573

5.2.3 Abstract

A typical problem that is solved by supervised machine learning techniques is classification. There are many types of this problem, such as binary classification, multi-classification, and multi-labels. The development of techniques that perform well in different types of classification problems, and for datasets with many or few variables, or many or few data, remains a challenge. On the other hand, recent work indicates that machine learning can benefit from quantum computing in terms of reducing computational complexity and improving its performance. The purpose of this research is to analyze the use of Quantum Support Vector Machines (QSVM) for different types of classification problems. For this, several kernels (linear and Gaussian) are considered. In turn, a comparison is made with Random Forest. Also, a detailed analysis is carried out for the binary classification of coffee beans, obtaining a precision of 95% and an accuracy of 79%. With this article we cover objectives one, and three.

5.2.4 Link to the full article

https://ieeexplore.ieee.org/document/10700573

Chapter 6. Conclusions

6.1 Summary

This study aimed to apply AI to improve production processes in agro-industrial MSMEs. The most used machine learning techniques in MSME management were analyzed. The analysis determined that there are no studies where ACODAT is applied in MSMEs. ACODATs for the MSME agroindustrial management were specified. Three industrial automation processes within the production chain were identified, in which autonomous cycles can be applied. Thus, an architecture for industrial automation of MSME production chains based on the ACODAT concept was proposed. The architecture combines multiple variables (e.g. temperature, humidity, acidity, quantity, quality and time) and allows the integration and interoperability of actors in the production context. The proposed architecture proposes three autonomous cycles to give autonomy to the industrial production chains of MSMEs. In addition, this work shows the instantiation of one of the autonomous cycles for the characterization of the inputs to be transformed as a first step towards an efficient and effective industrial automation process of the production chain.

Novel AI techniques were used to improve the agro-industrial processes of MSMEs that allow the development of diagnostic and classification models, among others. For example, FCMs were used to build models to evaluate the level of digital transformation in MSMEs, or QSVM were used in different classification problems (binary and multiclass). In addition, the use of these models in the coffee industry was described in detail. Two meta-learning architectures were also proposed to manage ACODATs, one that manages its knowledge models, and another based on an ACODAT to automate the management of other ACODATs in a production chain. A notable advantage of this proposal is the ability to maintain a continuous learning process with the arrival of new data or information for a supervised ACODAT. In most cases, the Café Galavis company located in Cúcuta, Colombia, was used as case study. This company, founded in 1918, is dedicated to the roasting and distribution of Colombian coffee.

The relevance of the results obtained is multiple. The definition of ACODATs allowed the automation of production processes in MSMEs using their own data, from their contexts, or expert knowledge. The knowledge/machine learning models of ACODATs solve different tasks, such as diagnosis and prediction. On the other hand, the meta-learning architecture to manage ACODATs automates their construction in new contexts. Finally, the use of models to evaluate the level of digital transformation in MSMEs based on FCMs allows the analysis and study of the current state in an organization, but also, the identification of what actions to take to improve its level.

6.2 Limitations and future work

The results of this thesis achieved the proposed objectives. However, this thesis had some limitations, in particular, due to the difficulty in obtaining data from the agro-industrial sector. Regarding ACODAT, the main limitation is the complexity of its integration into the production processes of an organization. This would imply an instrumentation that would allow data to be captured, and then, to act on the production process based on the conclusions reached by ACODAT. Regarding the FCM to build the Digital Maturity Level Assessment Model in an organization, the model was based on the COBIT model. It is necessary to determine if there are more suitable standards for digital transformation in the current literature. Also, another important limitation has to do with the availability of data to use this assessment model, which must come from the organization itself (this may imply in some cases dataizing organizational processes). This work presents a first case study. Another limitation is linked to quantum computing, particularly as the sample characteristics in the SVMs increase, the representation of the qubit wave function would experience exponential growth in simulation requirements, making it unfeasible when studying large datasets of classifiers in current machines that simulate quantum processing. Finally, one limitation of the meta-learning architecture is the collection of large volumes of data required to train the initial models in the knowledge base. Another limitation is having a wide variety of models, techniques and data to make it functional and interesting; otherwise, it could face the cold start problem that affects many adaptation systems.

Regarding future works, some are as follows. A first future work is the integration of ACODAT with the Digital Maturity Model based on FCMs to automatically identify gaps in digital transformation in an organization, and how to respond using ACODATs. In this context, the role of the meta-learning architecture is vital, because it will allow monitoring and guiding the process. Another future work is to build a robust Digital Maturity Model based on FCMS, using multiple national and international data sources to build the model. For this, distributed strategies such as federated learning can be considered to conceive a global model, in this case of a country, a region, etc. that allows generalizing behaviors. Another future research consists of exploring the behavior of several synthetic data generation algorithms within our meta-learning architecture. In addition, eventually implement a Federated

Meta-Learning approach with ACODAT to take advantage of the Meta-Knowledge generated in the different MSMEs, which would contribute to the privacy and security of their data, and reduce the computational resources required. Other, future work should focus on implementing ACODAT for the Automation of the Productive Chain of MSMEs in the Agroindustrial Sector in a real context. A challenge in the implementation of this type of solutions is the cost associated with the sensors for data acquisition, their optimal distribution, the computational and storage requirements, and the Internet connectivity of the industrial plant, which should also be studied. Finally, future research will focus on evaluating how the sensitivity of our meta-learning architecture varies in different scenarios of ACODAT and process automation in MSMEs. In addition, its application in other autonomous cycles of agro-industrial chains will be considered.

References

[1] Sánchez M., Aguilar J., Cordero J., Valdiviezo-Díaz P., Barba-Guamán, L. Chamba-Eras L. Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service. Advances in Intelligent Systems and Computing, 444, 993–1002, 2016. https://doi.org/10.1007/978-3-319-31232-3_94

[2]. Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector. Information, 15(2), 2024. https://doi.org/10.3390/info15020086

[3] Fuentes, J.; Aguilar, J.; E Montoya, E.; Hoyos, W. Evaluation of the Level of Digital Transformation in MSMEs Using Fuzzy Cognitive Maps Based on Experts," XLIX Latin American Computer Conference (CLEI), La Paz, Bolivia, 2023 (best paper), https://doi: 10.1109/CLEI60451.2023.10346161

[4]. Hoyos W, Aguilar J, Toro M. A clinical decision-support system for dengue based on fuzzy cognitive maps. Health Care Manag Sci. 25(4):666-681, 2022 doi: 10.1007/s10729-022-09611-6. Epub 2022 Aug 16. PMID: 35971038.

[5]. Fuentes, J.: Aguilar, J.; Montoya, E.; Hoyos, W.: Benito, D. Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps: Explainability Analysis on Fuzzy Cognitive Maps, CLEI Electronic Journal, 27(2), 2024 https://doi.org/10.19153/cleiej.27.2.2

[6] Fuentes J, Aguilar L, Hoyos W, Montoya E. “Assessment of the level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps”, Progress in Artificial Intelligence.

[7] Fuentes, J. Aguilar. J; Montoya, E, “Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs.” Agricultural Systems, 2024.

[8] Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. Analysis of Quantum Support Vector Machines in Classification Problems, 50 Latin American Computer Conference (CLEI), Bahia Blanca, Argentina, 2024.

Appendix A. A SYSTEMATIC LITERATURE REVIEW ON THE MANAGEMENT OF SMALL AGROINDUSTRIAL PRODUCERS USING ARTIFICIAL INTELLIGENCE

A SYSTEMATIC LITERATURE REVIEW ON THE MANAGEMENT OF SMALL AGROINDUSTRIAL PRODUCERS USING ARTIFICIAL INTELLIGENCE

Jairo Fuentes [1,2] , José Aguilar [2,3,4,*]

[1] Universidad Francisco de Paula Santander, Av. Gran Colombia, Cúcuta, Colombia
[2] GIDITIC, Universidad EAFIT, Carrera 48 No. 7 Sur-50, Medellín, Colombia
[3] CEMISID, Facultad de Ingeniería, Universidad de Los Andes, Mérida, Venezuela
[4] IMDEA Networks Institute, Leganés, Madrid, Spain
[*] Correspondence author

Summary: Small agro-industrial producers make up a significant portion of companies and play a vital role in a country's economic and social advancement. However, it is crucial that they integrate technology and become agro-industrial producers. Artificial Intelligence (AI) can use process data to build knowledge models that allow agro-industrial producers to make decisions about production chains to achieve high levels of performance and competitiveness. The objective of this work is to perform a systematic literature review on the use of AI for the management of small agro-industrial productions. Seven scientific digital libraries (ACM, Web of Sciences, Scopus, Google Scholar, IEEE Xplore, Sciences Direct and EBSCO) were used to investigate the use of AI for the management of small agribusinesses. A search methodology based on research questions, inclusion, and exclusion criteria, among other aspects, was used to select the most relevant articles. An analysis of sixty-two articles was performed to assess the strengths and limitations of using AI in Small Agribusiness Production Management, as well as to identify the primary challenges and opportunities. In the systemic review of the literature, some studies based on AI for the management of small agro-industrial productions are evidenced. Among the limitations found are the low investment in new technologies, and the rejection of changes and innovative business models, very common aspects among Micro, Small & Medium Enterprises (MSMEs). The results of this systematic review offer significant insights for future studies on the application of AI in the production chain of small-scale agro-industrial producers.

Keywords: Artificial Intelligence, Agribusiness, Small Agribusiness Production Management, systematic literature review.

1. Introduction.

Agribusiness is the integration of agricultural production, processing, and marketing under a single integrated management. It is defined as the subset of the manufacturing sector that processes agricultural, forestry and fishery raw materials and intermediate products (Sokolov et al., 1998). The agro-industrial sector comprises a diverse range of industries, including food, beverage, and tobacco manufacturers, textile and clothing makers, wood and furniture producers, printing and paper product manufacturers, and rubber manufacturers, among others (da Silva et al., 2013). According to da (]. C. Silva, D. Baker, 2013), agribusiness encompasses suppliers of inputs for the forestry, fishing, and agricultural sectors, as well as providers of food and non-food agribusiness products.

For small agro-industrial producers to transition into technologically advanced agro-industrial producers and maintain their crucial role in a country's economic and social progress, it is essential that they adopt technology. Consequently, Information and Communication Technologies (ICTs) are an opportunity for agro-industrial producers to improve their production processes, promote their products and gain a competitive advantage that will benefit the entire market (Gutarra & Valente, 2018).

With Artificial Intelligence (AI), it is possible to exploit data with learning algorithms, reasoning mechanisms, etc. In today's business landscape, data has taken on a critical role and can provide a competitive advantage to those who leverage it effectively. Even if competitors use comparable methods, an organization can gain an advantage in a competitive industry by having the best data. Machine Learning (ML) is an analytical method that enables small agribusiness producers as players in production chains to achieve high levels of productivity and competitiveness with additional values, such as incorporating techniques to learn and discover patterns, trends, and relationships in data, and thanks to this knowledge, better insights are offered.

To understand the approaches and challenges of AI in small agro-industrial producers, we analyze two aspects in this article. The first step involves evaluating the existing level of AI utilization in the management of small agro-industrial producers. Second, a discussion of the current challenges in the intelligent management of small agro-industrial productions.

Therefore, this paper presents a systematic literature review (SLR) on this topic. Similar works such as (Sanchez-Calle & Castillo Armas, 2022) and (Navarro et al., 2020) analyzed the adoption of IoT in smart agriculture and outlined the key components, such as Smart Agriculture, Smart Farming, and Machine Learning, along with the potential use cases of IoT solutions. In their works, (Sanchez-Calle & Castillo Armas, 2022) and (Navarro et al., 2020) point out that management systems are increasingly incorporating cloud and Big Data technologies to process large data sets, often using other supporting technologies to complement these solutions. It is noted by the authors that the use of AI and image processing techniques has become more common for enhancing smart agriculture management. Among the various IoT applications for smart agriculture, crop monitoring was found to be the most prevalent. The review indicated that smart agriculture IoT solutions can incorporate multiple network protocols. Furthermore, (Mirani et al., 2019), (Nath et al., 2022) presented a review of IoT agribusiness: applications and challenges. Areas of AI such as IoT and Sensor Networks (WSN) enable efficient and reliable agricultural and livestock monitoring. (García et al., 2023) focused on the use of ML in precision livestock (PLF), directed at grazing and animal health. They venture into agricultural automation and find new opportunities in the field of agriculture and livestock. Finally, (Toscano-Miranda et al, 2022) presented the state of the art on the use of sensors and AI techniques for the management of diseases and insect pests in cotton crops.

The main contribution of this work is the presentation of the use of AI techniques in the management of small agro-industrial producers; a sector that performs most of its processes manually, which is due to the lack of technological investment and/or ignorance of the benefits of the use of these techniques. Their processes would improve with the adoption of AI techniques applied to their needs or requirements, contributing to a continuous improvement of their processes. The objective is to conduct a systematic literature review on the use of AI techniques for the intelligent management of small agro-industrial producers. To the best of our knowledge, no previous research has conducted this particular literature review. This article is structured into three sections. Section 2 outlines the systematic literature review methodology, Section 3 details the review findings, and Section 4 concludes the paper with an extensive discussion that highlights the research questions, limitations, and challenges.

2. SLR Methodology

The systematic literature review (SLR) is based on a modification of the PRISMA approach (Moher et al., 2009), which aims to identify relevant studies based on research questions. The research question is derived from the article's objective, which is to examine the use of AI in enhancing the management of small agribusiness productions. The PRISMA method guides the search process in identifying studies that align with the research questions, leading to a comprehensive and focused review of the relevant literature. In particular, a research question has been defined:

Q1. ¿How has AI been used for the intelligent management of small and medium agro-industrial productions?

2.1. Search strategy

To evaluate the number of studies that could be applicable and identify any existing systematic reviews, initial searches were conducted. These searches aimed to assess the relevance of available studies in the context of the research topic. Several combinations of keywords derived from the research question were used, which defined the search strings (see Table 1). The following digital libraries were used: ACM, Web of Sciences, Scopus, Google Scholar, IEEE Xplore, Sciences Direct and EBSCO.

Table 1. Search strings used by research question.

Question	String search
Q-1.	("Artificial Intelligence" OR "Machine Learning" OR "Fuzzy Logic" OR "Artificial Life" OR "Expert Systems" OR "Data Mining" OR "Data Analysis" OR "Ontologies" OR "Computer Vision" OR "Bayesian Networks" OR "Knowledge Engineering" OR "Artificial Neural Networks" OR "Reactive Systems OR "Reactive systems" OR "Evolutionary computing" OR "Rule-based systems" OR "Case-based reasoning" OR "Knowledge representation techniques" OR "Semantic networks" OR "Computational linguistics" OR "Natural language processing" OR "Agents" OR "Multiagent systems" OR "Intelligent systems"). AND ("Intelligent management model" OR "Intelligent management framework" OR "Intelligent management" OR "Intelligent enterprise" OR "Intelligent agribusiness" OR "Intelligent agriculture" OR "Management") AND ("microenterprise" OR "small enterprise" OR "small business" OR "medium enterprise" OR "SME" OR "MSME") AND ("agricultural industries" OR "agribusiness").

The following are the inclusion criteria used for selecting publications: CI-1 Publications in scientific journals or conference proceedings whose titles are related to the research question; CI-2, which included publications containing keywords that matched the search string; CI-3, which encompassed publications whose abstracts, introductions, or conclusions were related to the selected topic; and CI-4, which focused on publications published in the year 2013. CI-5 Works in two languages: English and Spanish. The criteria that allow us to discard publications that are not of interest for this work are: i) publications that do not meet the above inclusion criteria, ii) publications in sites/journals of low recognition, and iii) all duplicate publications.

2.2. Selection process

To ensure the selection of relevant articles, the authors conducted a selection procedure in three stages. Firstly, they eliminated duplicate articles and assessed the titles and keywords of the remaining articles using search strings. Secondly, they evaluated the abstracts of the candidate papers based on the inclusion and exclusion criteria. Lastly, they reviewed the full texts of the selected papers from stage 2. The selection process resulted in the retrieval of 129 articles, of which 94 were selected after eliminating duplicates and evaluating titles and keywords. From stage 2, 74 articles were selected based on their abstracts, and finally, 62 articles met all the eligibility criteria.

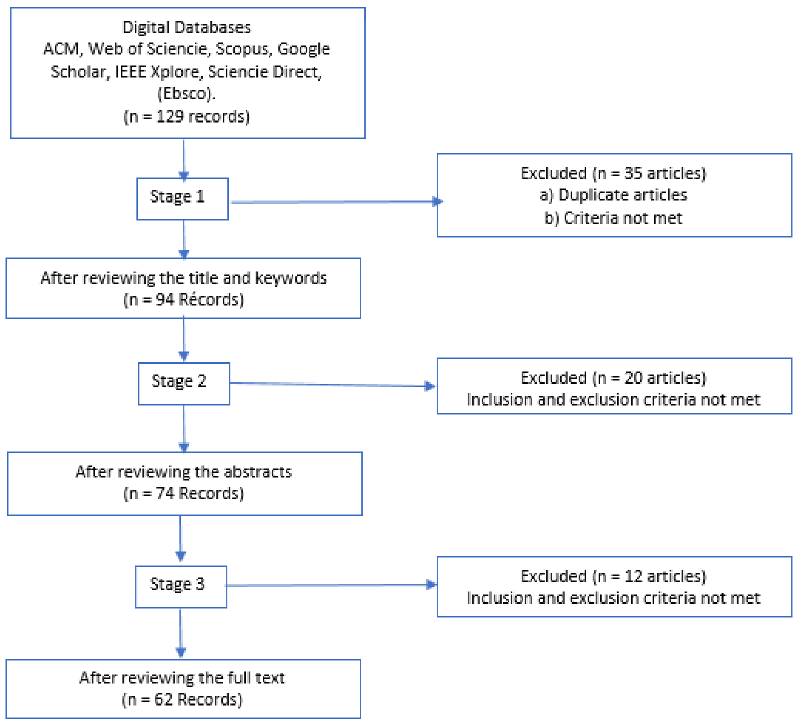

Figure 1. Flow chart of the selection process.

2.3. Preliminary results

In relation to the inquiry on the application of AI in managing small and medium agro-industrial productions, a total of 25 papers were selected, with 20 published in journals and 5 in conference proceedings. Figure 2 illustrates that Deep Learning (DL), Machine Learning, and Artificial Neural Networks are the focus of research in countries such as the USA, Brazil, and China (Fiehn et al., 2018), (Jagannath et al., 2019), (Davoudi Kakhki et al., 2019), (Schulte et al., 2019), (Pinheiro et al., 2020), (Garcia & Vieira, 2008), (Das et al., 2021), (S. Das et al., 2020), (Pereira et al., 2019). The utilization of CNNs for data analysis in IoT-based smart agriculture systems and the development of decision support systems for precision livestock farming were among the research findings.

It is followed by India and Israel, which have conducted IoT studies with autonomous robots and intelligent systems for management in agriculture MSMEs (Ram et al., 2020), (Joshi et al., 2017), (Heilbrunn et al., 2017), (Guo et al., 2020), (Li et al., 2018), (Stackpole, 2021). The integration of robotics with Green IoT gains further advantage with the collection of spatial information data using visualization techniques. Furthermore, due to the high energy consumption of IoT technologies, there have been efforts to develop energy-efficient IoT solutions known as Green IoT Technologies.

In addition, Canada, Switzerland, Ireland, South Korea, Morocco, Russia, and Belgium are working on Big Data, Deep Neural Networks, and ML ((Neethirajan, 2020), (Groher et al., 2020), (Stewart et al., 2017), (Kim et al., 2016), (Chiba et al., 2019), (van Witteloostuijn & Kolkman, 2019). In precision livestock farming, Big Data, sensors, and ML methods have been utilized alongside multimedia applications for precision agriculture (AP) and smart agriculture (SM), and a K-Means algorithm has been integrated to identify multimedia traffic.

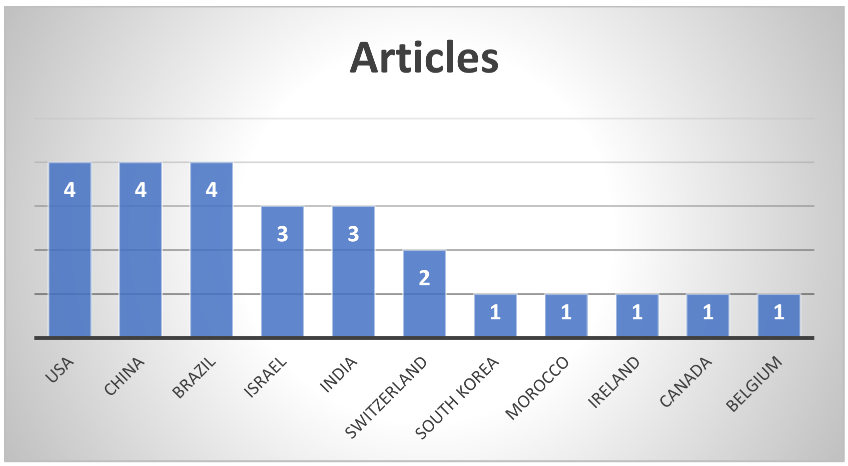

Figure 2. Countries that are working in the area of AI for small and medium agro-industrial production.

In addition, 37 articles, of which 29 are in journals and 8 in congresses, deal with advanced information technologies for the management of small agribusiness production. From Germany, Belgium, South Africa, and China there are papers on the implementation of intelligent sensing and monitoring cycles, analysis, and planning, as well as control of agricultural operations through a cloud computing (CC) based event management system (Carpio et al., 2017), (Roukh et al., 2020), (Wiseman et al., 2019), (Zhang et al., 2019). Works from Belgium, Germany, Korea, Ecuador, and Turkey have used Data Mining (DM) techniques (Park et al., 2017), (Benhamida et al., 2019), (C. P. Lopez et al., 2019), (Sen et al., 2016). Some of these works proposed using big data and Data Mining to foster partnerships in SMEs. In Korea, Spain, India, Bangladesh, and Italy there are works on IoT (Oliver et al., 2018), (Sanchez-Calle & Castillo Armas, 2022), (Kumar et al., 2021), (Talavera et al., 2017), (Pandey et al., 2019), (Happila et al., 2020), (Toyeer-E-Ferdoush et al., 2020),(Hajjaji et al., 2021). Some of these works suggested integrating IoT and Big Data technologies to develop real-world smart environmental applications that can monitor, safeguard, and improve natural resources. The examined areas include smart agriculture, smart environmental monitoring, metering, and smart disaster alerts. Furthermore, there are research works from India and Brazil that center on cyber-physical systems (Martini et al., 2019), (Neethirajan, 2020). Some of these works developed a computational model for ubiquitous smart services in indoor agriculture. Also, from Panama and Indonesia, there are works on blockchain technology to improve processes in the agricultural supply chain

(Pradana et al., 2020), (Das et al., 2021), (Caballero & Rivera, 2019). Some of these works use Blockchain modeling for traceability information systems in the coffee agribusiness supply chain, or in an automated tomato maturity grading system using transfer learning. Finally, in (Gezgin et al., 2017) analyze digital transformation (DT) in the context of supply chain transformations. They conclude that by combining digital applications with operational changes, companies can achieve substantial performance improvements that endure over time, while also boosting supply chain performance in MSMEs.

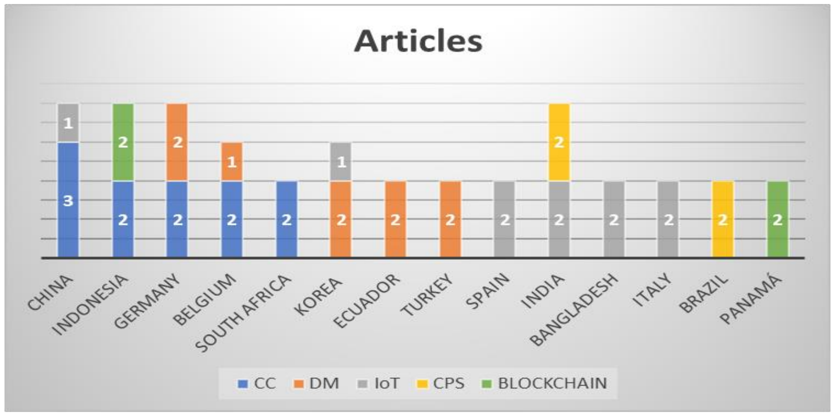

Figure 3. Countries with articles on advanced information technologies for small and medium agro-industrial production (CC - Cloud Computing, DM - Data Mining, IoT - Internet of Things, CPS- Cyber-Physical Systems, - Blockchain).

3. Analysis of the reviewed articles

The reviewed papers were classified into two categories: AI for data analysis of agro-industrial production processes, and AI for the management of agro-industrial production processes (supervision, monitoring).

3.1 Artificial Intelligence for data analysis of agro-industrial production processes.

(Cordeiro et al., 2022) recommended various deep neural network structures to develop soil moisture prediction models. They applied KNN techniques to fill in missing data with values that ensure a desired level of dependability. Moreover, they assessed the energy consumption of nodes in the cloud. Analyzing data is essential for improving the production chain, but the authors did not compare their results with other DL algorithms that could help achieve more accurate predictions. (Sen et al., 2016) utilized satellite image analysis, genetic algorithm, and convolutional neural network to develop a decision-making system for precision agriculture and agribusiness. Their objective was to make optimal decisions in agriculture using data analysis techniques. Nevertheless, they did not conduct case studies on MSMEs in sectors as energy, high-tech, media, advanced manufacturing, and nuclear power.

In their study, (Diaz-Gonzalez et al., 2022) recommend using ML-based models to estimate soil quality, in order to analyze data of different dimensions on soil management and crop conditions. However, they did not compare the use of different ML algorithms and autonomous cycles of data analyses to improve crop conditions. (Suarez et al., 2022) applied data mining techniques, specifically decision tree algorithms and random forests, to detect various climatic factors and types of agricultural production, which allowed the identification of smart strategies for agriculture. However, the research challenges do not include the exploration of alternative ML algorithms or other variables that could improve agro-industrial production. (Gupta et al., 2022) and (Elavarasan & Vincent, 2020) utilized ML approaches to forecast agricultural yields using an experimental dataset and utilizing classifiers such as particle swarm optimization, support vector machine, K-nearest neighbor, and random forest.

In the field of smart agriculture, authors like (Singh et al., 2022), (Emanuela et al., n.d.), and (Luiz Carlos & Ulson, 2021) suggest the application of artificial neural networks and wireless sensor networks to identify and manage pest infestation. By gathering data from sensors installed in the field, insect behavior can be analyzed, making it easier to detect and control pests. The next step in this research is to develop sensory circuits in agriculture that integrate GPS and GSM technologies to determine the precise location of the infestation, enabling prompt action to be taken to protect the crop and boost productivity. The study by (Rehman et al., 2022) involved utilizing ML approaches to collect and analyze written data from agricultural experts, which led to valuable information on present agricultural inputs for farmers. To improve crop productivity, it is recommended to capture farmers' opinions using voice recordings and various DL techniques. A DL classifier system was developed by (Kundu et al., 2021) to advance the accuracy of seed quality testing by classifying seed images based on texture, color, and shape. Healthy and diseased seeds from millet, pearl, and maize were included in the training dataset. However, a challenge that needs to be addressed is the improvement of DL's potential to efficiently identify seeds from the bulky pile gathered during harvesting.

Likewise, Data mining can be applied in organizations, using methodologies that define business problems (Vizcarrondo et al., 2012). There are related works that have used AI concepts, such as Intelligent Agents and ML, to improve the supply chain (H. A. G. Lopez & Cisneros, 2017), (Meyer & Dykes, 2020), (Nylander et al., 2017), (Ericson et al., 2020), (Ghobakhloo & Ching, 2019), or for image processing (Suarez et al., 2017), speech recognition (Huang et al., 2015), among other uses. However, there is no evidence of works in the area of Knowledge Engineering, Reactive Systems, Evolutionary Computing, Rule-Based Systems, Case-Based Reasoning, Knowledge Representation Techniques, Ontologies, Natural Language Processing, Multi-Agent Systems, among others, for the management of small agro-industrial processes.

Figure 4 shows the use of ML techniques in data analysis for precision agriculture, soil management and climatic factors, with 25% using convolutional neural network algorithms, and 17% using decision trees and deep learning.

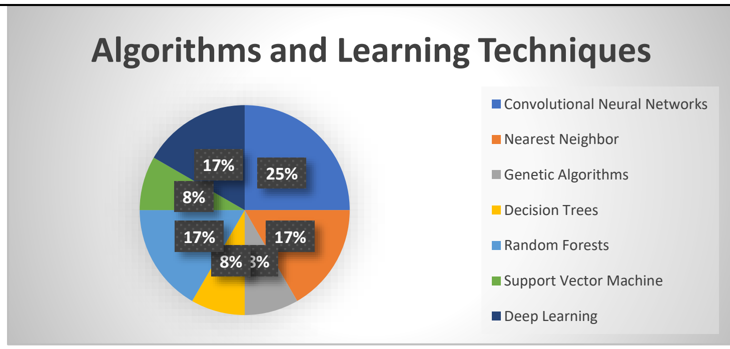

Figure 4. used ML Techniques.

In general, the most used techniques are: Linear/Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (K-NN) and boosting techniques (e.g., XGBoost). Table 2 shows the performance value reported in the literature in the field of agroproduction of each of the techniques for three classic metrics to evaluate ML models. We can see that ANN and XGBoos are the best techniques but Random Forests has very similar values. Now, if we value interpretability, the Linear/Logistic Regression technique gives good results with high explainability. If execution time were relevant, it would be Decision Trees.

Table 2. Comparison between ML approaches

Algorithm	Accuracy	F1-Score	AUC-ROC
Linear/Logistic Regression	0.85	0.84	0.88
Decision Trees	0.78	0.77	0.80
Random Forests	0.92	0.91	0.94
SVM	0.90	0.89	0.92
ANN	0.95	0.94	0.96
K-NN	0.86	0.85	0.87
Boosting (XGBoost)	0.94	0.93	0.95

3.2. Artificial Intelligence for the management of agro-industrial production processes.

(Luna et al., 2022) employed a genetic algorithm and a multi-criteria methodology to enhance the feeding sequence during the fattening period in the aquaculture industry. Despite this, the study did not consider the long-term planning needs of large aquaculture farms. (Kamyshova et al., 2022) have proposed a technique to improve the crop irrigation process, which consists of using a Phyto-indication system based on computer vision methods and a trained neural network to determine the percentage of plant irrigation in the irrigator's current sector. In (Nath et al., 2022), an IoT-enabled sensor clustering device was developed to determine soil quality indicators such as moisture, corrosivity and key minerals including nitrogen, phosphorus, and potassium.

Furthermore, (Shreyas & Narayana, 2022), (Cesar Flores, 2022), (Zhao et al., 2021) handled image processing techniques for early recognizable evidence of plant diseases. Several DL algorithms were applied. (Ghasemi-Mobtaker et al., 2022) use adaptive neuro-fuzzy inference system methods and artificial neural networks to predict the energy production, economic benefit, and global warming potential of wheat production. Also (Lima et al., 2021), (Schulte et al., 2019), (Jiménez Cercado et al., 2020), (V. H. M. García et al., 2014), worked around behavioral monitoring of dairy cows, using AI techniques and DL schemes for implementing computer vision techniques, and welfare monitoring, and behavioral tracking systems.

The dairy, bakery, beverage, fruit, and vegetable industries can benefit significantly from AI technology, according to (Addanki et al., 2022). The authors further suggest exploring 3D printing and robotics to improve the manufacturing and service processes of food companies. Another major challenge is the prediction of the expiration date of food present inside the package with the help of sensors would help to know the spoilage of food and prevent food-borne illnesses.

(Guisiano et al., 2020) and (Feijó et al., 2021) developed a learning model that significantly enhances the precision of weather forecasts over the short term to assist farmers in making decisions. An architecture designed to tackle the difficulties associated with acquiring, processing, and presenting large amounts of data, including real-time and batch data, and monitoring agricultural operations via a cloud is suggested by (Roukh et al., 2020). In (Oliver et al., 2018) applied ML techniques for land cover classification, which is important for analyzing climate change and monitoring ecosystems. (Sharaf et al., 2019) (Sharaf et al., 2019), used convolutional neural networks, to perform soil and climate change studies in precision agriculture and agribusiness.

(Ramirez-Asis et al., 2022), (Reddy & Rajanna, 2022), (Arkeman et al., 2022), (Lakshmi et al., 2022), and (Parasuraman et al., 2021) have developed Logistics Support Systems that utilize IoT and AI to provide farmers with fast transportation of crops from fields to consumers, which can help reduce cold storage fees and maintain crop quality. Future research may explore the integration of other AI and ML technologies to address problems such as storage and logistics management, supply chain management, and collaboration.

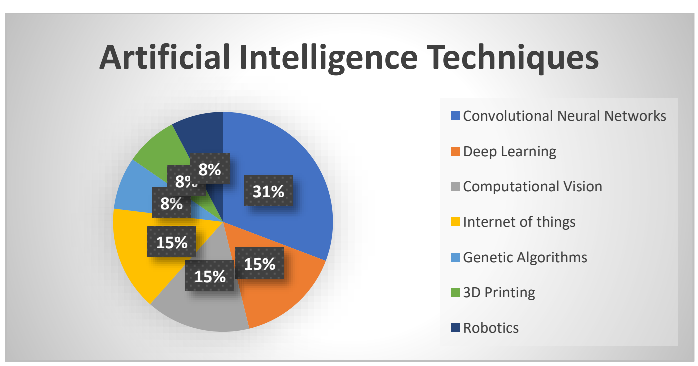

Figure 5. AI Techniques.

Figure 5 shows the AI techniques for the management of agro-industrial production processes, with 31% using convolutional neural network algorithms and 15% using computational vision, other DL algorithms, and IoT, among others.

3.3 Case studies

The application of AI in MSMEs in the agricultural field has been recognized for its ability to improve productivity, efficiency and sustainability. Below are some case studies that exemplify how AI has been integrated into agricultural MSMEs:

- Farmers' Cooperative Society in Kenya: A smallholder farmers' cooperative in Kenya was facing challenges managing crops and forecasting yields due to climate variability. To address this situation, the cooperative implemented the AI platform "Watson Decision Platform for Agriculture" developed by IBM (https://www.ibm.com/downloads/cas/0XZ0VRAW). This tool uses climate, soil and crop data to offer personalized recommendations on planting, irrigation and fertilization. Results included significant improvements in agricultural decision-making, a 15% increase in crop yields, and a 20% reduction in water and fertilizer use.
- Small Farm in India: A farmer in India growing rice and wheat was facing challenges related to pests and diseases that impacted the productivity of his crops (https://www.computerweekly.com/news/365534904/How-Indias-Cropin-is-tapping-the-groundswell-in-agritech). In collaboration with agritech startup "CropIn", he implemented an AI-based solution using satellite imagery and machine learning to monitor crop health and detect problems early (https://www.cropin.com/es/about). Results included early detection of diseases and pests, allowing for rapid response. There was also a 25% reduction in pesticide use and an 18% increase in crop yields.
- Vineyard in Chile: A small vineyard in Chile faced the challenge of optimizing irrigation to improve grape quality and reduce water use. To achieve this goal, the vineyard implemented an innovative AI solution developed by the Chilean company "Kilimo" (https://kilimo.com/home-eng/). This platform uses data from soil moisture sensors and weather forecasts to optimize the irrigation system. Results included a 30% reduction in water consumption, improvements in grape quality leading to the production of higher quality wines, and a reduction in operating costs associated with irrigation.
- Coffee Farm in Colombia: In Colombia, a small coffee farm had difficulties in properly managing fertilization and timely detection of diseases. To address these challenges, the farm implemented an AI-based solution provided by the company "Agrosmart" (https://www.agrosmart.app/). This solution leveraged real-time data from field sensors and image analysis to optimize fertilization and detect diseases early. The results obtained included a 20% increase in coffee yield, a 15% decrease in fertilizer use, and an improvement in coffee quality that facilitated entry into high-end markets.
- Vegetable Farm in Spain: a small farm dedicated to growing vegetables faced challenges related to climate variability and efficient management of water resources. To overcome these challenges, the farm implemented an AI platform developed by "Prospera" (https://prospera.ag/), which uses climate and soil data to offer precise recommendations on irrigation and crop protection. The results obtained included a 25% decrease in water consumption, a 15% increase in vegetable yield, and an improvement in resistance to adverse weather conditions.

These cases demonstrate that the integration of AI solutions can have a positive impact on the productivity and sustainability of MSMEs in the agricultural field. AI technologies facilitate more accurate and efficient resource management, improved decision-making, and greater ability to address climate and pest challenges. These improvements not only benefit farmers but also contribute to strengthening sustainability and food security globally.

4. Discussion

In the articles reviewed, there are studies in AI for data analysis of the agro-industrial production process and AI for the management of agro-industrial production processes in the agribusiness sector, but in the MSME sector, there are few works related to the use of AI techniques. Our challenge would be to use ML for the improvement of productive processes in the MSME agribusiness chain. This section presents a summary of the articles reviewed, according to their limitations and challenges.

4.1. Limitations of the work reviewed.

According to our review, platforms such as Customer Relationship Management (CRM), Business Intelligence (BI), Enterprise Resource Planning (ERP), Data Processing Systems, and Applications & Products in Data Processing (SAP), are the most used in agribusiness MSMEs. Few producers use intelligent data management techniques (Data Mining and Data Analytics) and ML techniques (CNN, Support Vector Machines (SMV), Randoms Forest (RF) and DL). In addition, there are other advanced information technologies, such as IoT, CC, smart farm architecture and cyber-physical systems (CPS), to consider (see Figure 6).

Among the limitations are a low investment in new technologies, rejection of changes and innovative business models, few adaptations, which are very common aspects among MSMEs. In addition, another limitation is the lack of real-time data acquisition and integration of heterogeneous data sources for control-level automation for MSMEs.

In general, in MSMEs in the manufacturing industry, data is not used enough to optimize processes. Some companies have implemented de facto "Industry 4.0 projects" but have encountered problems such as reducing the amount of data sent to the cloud from the local infrastructure. On the other hand, the introduction of IoT continuously generates large amounts of data. Therefore, it is necessary to create an architectural model that combines IoT, edge computing, semantic web, and CC. However, IoT has other limitations or problems that may hinder real-time monitoring, for example, security and connectivity. Finally, many technologies are not mature enough, such as CPS and 3D printing technologies, to enable the integration of communication networks, computational capability, intelligence, and autonomy with physical processes. Combining AI with 3D printing technologies could increase production throughput by reducing the risk of error and facilitating automated design (Terán et al., 2017).

The current situation is due to the fact that most small agro-industrial producers lack economic resources, or are insecure about innovating in new technologies, which leads them to use traditional software. Modern technological innovations are incorporated into the farming industry to improve decision-making and enhance resource management, known as smart agriculture. Incorporating advanced technologies is essential for boosting the efficiency of productive processes in MSMEs. These technologies include IoT, ML, CC, augmented reality, Big-data, autonomous robots, additive manufacturing, 3D printing, systems integration, and various other key technologies in the industry.

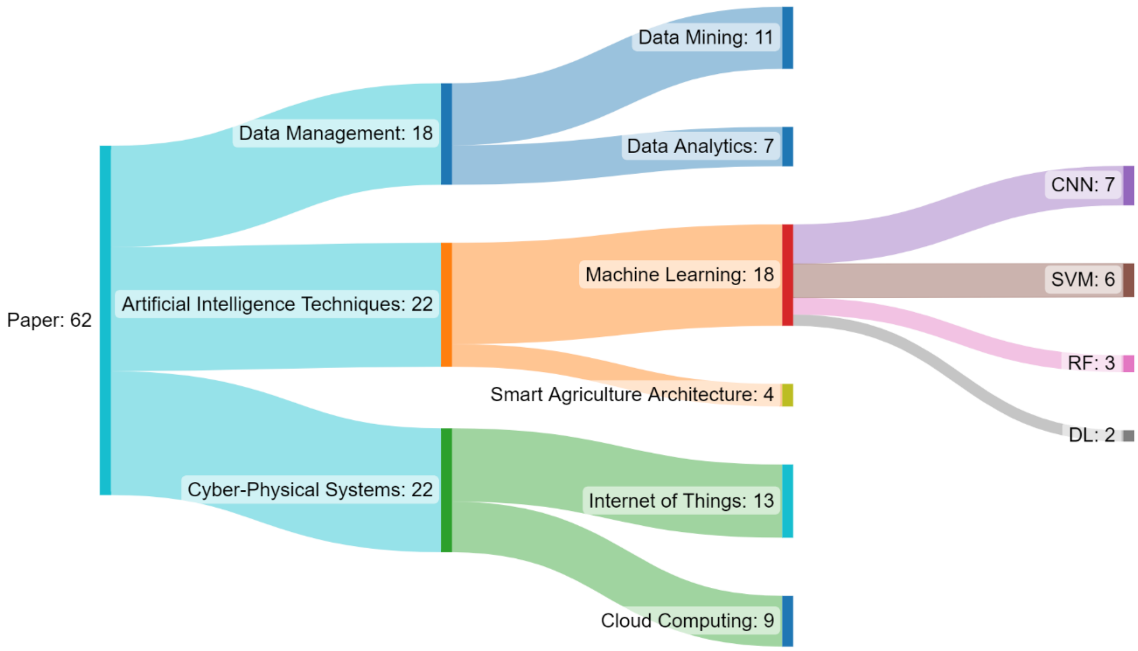

Figure. 6. Trends in reviewed works on technologies for agro-industrial process management.

In summary, predictive and diagnostic models for agro-industrial processes use various methods and techniques to identify problems, predict outcomes, and make informed decisions. These are some of the most common methods used in the literature. In the development of diagnostic models, the most common were Decision Trees, Bayesian Networks, Artificial Neural Networks and Rule-Based Systems. For the construction of predictive models, the most common techniques were Linear and Nonlinear Regression, Support Vector Machines, Random Forest, Deep Learning, and techniques for modeling time series such as ARIMA (Autoregressive Integrated Moving Average) and SARIMA (seasonal ARIMA). On the other hand, to build behavioral patterns, grouping techniques such as k-means or DBSCAN have been used. Finally, some classic data analysis tasks involve the exploration and cleaning of data as feature engineering processes. Also, Cross Validation and Hyperparameter Optimization are classic tasks. Thus, all these techniques have been adapted according to the specific needs, the nature of the available data and the objectives of the analysis.

4.2. Challenges

Two primary areas of concern for future work are data analysis and management of agro-industrial production processes. Each category of challenges is presented below.

4.2.1 Artificial Intelligence for data analysis of the agro-industrial production process.

With the increasing significance of extracting insights from data gathered from diverse sources, data analysis has become an essential field. As a result, it is crucial to create tools, approaches, and techniques to make their implementation simpler in different situations. Particularly for production processes of agro-industrial MSMEs, an "Autonomic Cycle of

Data Analysis Tasks" has been suggested for defining data analysis activities (Aguilar et al., 2022), (Sanchez et al., 2020), (Sánchez et al., 2016).

Data mining can also be used in an organization (Company), with a methodology that defines the company's problems that can be improved with Data Mining tasks. This methodology can be applied when there are different sources of data and knowledge, and their interactions are not well identified or defined (Sanchez et al., 2020). The implementation of these methodologies in managing small-scale agro-industrial productions would pose a challenge.

Processing data is a crucial stage in creating Data Mining tools within management models, as noted by (Pacheco et al., 2014). A significant challenge would be to implement automated data processing in the production processes of small agro-industrial MSMEs.

4.2.2. Artificial Intelligence for the management of agro-industrial production processes.

The SLR shows very few studies on the use of AI techniques in small agribusiness process management. However, they use information systems and business applications such as CRM, ERP and BI. In addition, they use SB and proprietary applications such as SAP (Sampaio & Bernardino, 2014).

Challenges include the development of Intelligent Systems for soil identification and analysis. Also, the implementation of feature engineering processes suitable for the automated Agri-Food sector, using advanced ML techniques to analyze local variables, with a focus on reinforcement learning and with the variation of hybrid algorithms, such as, for example, harvest time, cane age, precipitation (rainfall), climatic data, pest infestation, weeds. infestation, planting time, planting type, cane stage, soil type, and economic return of the site, among others (Elavarasan & Vincent, 2020).

Another task is to construct an ML-based model that can recognize plant diseases. An automated system can be built to detect, identify, and treat plant diseases using a time-series analysis of relevant variables. Another challenge is to record farmers' behavior to increase crop yield. Their feedback and gestures will be analyzed using diverse DL techniques. This would strengthen the analysis and yield more meaningful outcomes due to the large level of input from agricultural experts (Rehman et al., 2022).

AI and ML can play a crucial role in solving issues pertaining to supply chain management, collaboration, and warehouse and logistics management. Also, to enhance agricultural productivity, it may be beneficial to establish autonomous cycles consisting of identification, detection, and diagnosis, which can help determine the optimal time for harvesting crops, selecting the best storage and distribution center, and identifying the most effective shipping route, thus reducing post-harvest losses (Ramirez-Asis et al., 2022).

A technology of great interest and of little use even in the agro-industrial context is the wireless sensor network (WSN). In Zhou, Z. (2023), the adoption of WSN technology is suggested in smart agriculture. This technology allows the measurement of various environmental parameters, such as humidity, temperature, soil moisture and soil pH, with the purpose of improving the quantity and quality of crops. Likewise, Fang, J. (2022)

explores environmental factors for precision agriculture, such as irrigation, monitoring, soil qualities, and temperature. Particularly, he investigates crops using WSN technology and highlights communication technologies and sensors for precision agriculture. Research questions examine WSNs' impact on agriculture, seeking answers through the study. Also, the paper of Panda et al. (2021) proposes an agricultural monitoring system using a WSN to enhance farming production and quality without constant manual oversight. Key factors affecting plant productivity, such as temperature, humidity, and carbon dioxide levels, are regularly measured. This system allows farmers and agriculture specialists to view real-time readings online. If a critical change occurs, specialists notify farmers via mobile text and email. Thus, continuous monitoring of environmental data helps farmers study optimal conditions for maximum crop productivity, increased yields, and significant energy savings. Finally, Nourkhah et al. (2023) analyze the role of sensors in precision agriculture. By installing sensors and accurately mapping fields, farmers can closely study their crops on a microscopic scale, achieving more efficient resource management and significantly reducing their environmental impact.

On the other hand, an autonomous cycle for analyzing IoT variables in a production context is necessary to carry out a feature engineering process that involves extracting, merging, and selecting variables to create data sets for knowledge models. The autonomous cycle can also determine the ideal device placement for capturing the required process variables. Autonomous Cycles allow a natural introduction of AI technologies in crop management (Sánchez et al., 2016; Monsalve-Pulido et al. 2024). For example, an Autonomous Computing for Manufacturing Processes framework has been proposed in the context of Industry 4.0, which enables integration based on coordination, cooperation and collaboration needs (Sanchez et al., 2020). This framework could be implemented in the agro-industrial processes of small producers. Proper coordination among participants can enable the autonomous generation of new products, error detection, and recovery from failures, among other capabilities. Finally, the utilization of ML in agro-industrial production processes, and the definition of datasets for use in knowledge models (predict, identify, etc.), are other challenges.

Although significant progress has been made in the integration of AI in agribusiness, some AI techniques, such as knowledge engineering and reactive systems, have not been widely explored or adopted in this sector. *Knowledge engineering* focuses on the collection, representation, and application of specialized knowledge to address complex problems. In the agribusiness sector, this could encompass detailed knowledge of crops, agricultural techniques and pest control. However, its application has been restricted and faces the following challenges: i) Insufficient Knowledge Capture: Agribusiness often lacks systematic efforts to capture the expert knowledge of farmers and agronomists. This is partly due to the diversity and complexity of agricultural practices, which vary significantly by region, climate and crop types; ii) Challenges in Modeling: The variability and complexity of the agricultural environment make knowledge modeling a challenge. Systems based on rules and ontologies can quickly become obsolete or ineffective with changing conditions; iii) Interoperability: The integration of knowledge engineering systems with other AI systems and agricultural technologies is complex due to the lack of standards and the diversity of data and knowledge sources. Now, its incorporation will lead to impacts such as: i) Improved Decision Making: Better capture and use of expert knowledge could significantly improve real-time decision making in agriculture, such as for disease and pest management, and optimization of irrigation and fertilization, ii) Training and Knowledge Transfer: These systems can facilitate the transfer of knowledge to new farmers and workers, improving training and the dissemination of optimal agricultural practices.

In turn, *reactive systems,* which respond in real time to changes in the environment, have great potential in agribusiness, particularly in applications such as crop monitoring and irrigation management. However, its implementation also faces important challenges: i) Technological challenges: In many rural areas, the technological infrastructure may not be adequate to support reactive systems, which need a constant and solid connection; ii) High costs: Reactive systems can be expensive to develop and maintain, making them difficult to adopt by MSMEs, iii) Problems with real-time data: collection and processing real-time data processing can be complicated due to the variability of terrain conditions and the lack of suitable and inexpensive sensors. However, their use will impact the sector in the following ways: i) Operational efficiency: Reactive systems have the potential to improve operational efficiency by allowing immediate responses to changing field conditions, which will improve, for example, detection of diseases or the need irrigation; ii) Waste reduction: By acting quickly on issues they can help reduce waste and optimize resource use, which has a positive impact on sustainability and operating costs.

The lack of studies and applications of techniques such as knowledge engineering and reactive systems in agribusiness represents a lost opportunity to improve the efficiency, sustainability and resilience of the sector. To address this gap, more research and pilot projects are needed that explore how these techniques can be adapted and applied to the specific challenges of agriculture. This includes improving technological infrastructure in rural regions and training farmers in the use of advanced technologies.

4.2.3. Perspectives from agricultural economists and management experts.

It is possible to strengthen the analysis of the challenges and opportunities in the implementation of AI in agroindustrial production by considering both the perspectives of agricultural economists and management experts. This expanded approach allows technical aspects to be addressed along with economic, organizational and social implications. For example, from the perspective of Agricultural Economists is possible to consider:

- Efficiency and Productivity: The implementation of AI can optimize the utilization of essential resources such as water, fertilizers and pesticides, leading to increases in agricultural yields and, consequently, greater production. Agricultural economists can measure these impacts by calculating how production is increased and associated costs reduced;
- Impact on Markets and Prices: Improving predictability and efficiency can help stabilize agricultural prices, reducing market volatility. Economists can develop models to understand these impacts and anticipate how AI could affect prices both locally and globally. Improving the quality and traceability of agricultural products using AI could facilitate the opening of premium and export markets.
- Sustainability and Environment: AI can optimize the utilization of agricultural resources, which could result in reduced waste and lower environmental impact. Economists could calculate these effects in terms of economic savings and improved sustainability. Implementing AI can promote more responsible agricultural practices, such as precision agriculture and crop rotation, which contribute to improved soil health and biodiversity.

Economists can evaluate how these innovations impact the long-term viability and resilience of the agricultural sector.

On the other hand, the Management Expert Perspective can introduce:

- Change Management and Technology Adoption: To adopt AI, it is crucial to have a well-trained workforce. Managers must create training and education programs for farmers and agricultural workers, ensuring a smooth transition to the integration of new technologies. The implementation of AI may encounter opposition among workers due to fear of being replaced. Therefore, managers must devise strategies to counteract this resistance, involving employees in the adoption process and clearly communicating the benefits;
- Resource Management: To integrate AI, it is necessary to invest in technological infrastructure, including sensors, drones and software. Managers must plan and budget these investments, evaluating their profitability and return on investment. Technology requires regular maintenance and updates. Managers must establish preventive maintenance plans and ensure the availability of technical support,
- Innovation Strategies: Managers must promote an innovative culture within agroindustrial companies, encouraging experimentation and implementation of new technologies. And collaboration with universities, research centers and technology startups can accelerate the innovation and adaptation of AI in agribusiness. Managers must seek and establish these strategic alliances.

In general, these perspectives represent new challenges:

- Cost and Accessibility: For many agricultural MSMEs, the high initial cost of implementing AI technologies can be a considerable obstacle. Thus, it is necessary to implement subsidy programs and financing options that allow MSMEs to acquire advanced technologies more easily.
- Data and Privacy: The handling and use of sensitive data can present problems regarding the privacy and property rights of that data. Thus, it is necessary to establish clear rules on data ownership and implement measures to protect personal and business information.
- Interoperability and Standards: Differences in systems and technologies can complicate the incorporation of AI due to a lack of interoperability. It is necessary to encourage the adoption of open and compatible standards that facilitate integration and communication between various technological platforms.

Thus, the incorporation of AI in agroindustrial production presents not only a technical challenge, but also an economic, organizational and social one. Agricultural economists can analyze economic and market impacts, while management experts can create strategies for the effective and sustainable adoption of these technologies. This comprehensive approach ensures that AI implementation not only improves productivity and efficiency but is also viable and beneficial in the long term for all actors in the agribusiness sector.

5. Conclusions

This paper examines the use of AI techniques in aiding small agro-industrial producers, with an emphasis on recent developments. The study utilized an SLR approach, which provided a detailed description of the research outcomes. Scientific digital libraries were used to retrieve 129 articles, and after eliminating duplicates, 74 articles were selected based on their abstracts. Eventually, 62 articles met all the eligibility criteria and were included in the study. An important limitation of the study is that it only considered works published in high-impact journals referenced in the repositories used in the search. That implied that they were articles written in English. It does not include experiences that have not been recorded in digital media, nor experiences narrated or reported in other languages or in low-impact magazines or conferences. It is important to consider these biases when interpreting our assumptions and conclusions.

is that you recognize potential gaps or biases in the review methodology (e.g., languages included, publication bias). This helps reduce completeness assumptions.

During this study, authors from countries such as the United States, Brazil and China had a higher production of articles focused on DL, ML and ANN, followed by India and Israel, which have conducted studies on IoT with autonomous and intelligent robots. In addition, authors from Canada, Switzerland, Ireland, South Korea, Morocco, Russia and Belgium are working on Bigdata, Deep Neural Networks and ML.

The SLR found a limited number of studies exploring the use of AI techniques in managing small agribusiness firms. Nonetheless, these firms utilize information systems and business applications, including enterprise resource planning systems, customer relationship management systems, and business intelligence tools (Sampaio & Bernardino, 2014). Now, a limitation that we found is the non-use of WSN that allows capturing a lot of data from the production process, which would allow producing high-yield crops at a lower cost. On the other hand, it is necessary to educate and sensitize farmers, to facilitate a gradual change in their way of thinking and encourage the adoption of smart technologies.

Regarding the relationship between agroindustrial production and Industry 5.0, the latter is based on the integration of advanced technologies and human-centered approaches to optimize the efficiency, sustainability and resilience of agroindustrial processes. These technologies include Advanced Automation and Robotics, AI and Big Data, the Internet of Things (IoT), and paradigms such as Sustainability and Circular Economy, Personalization and Localized Production, and Worker Wellbeing. Industry 5.0 is applicable in sectors such as Precision Agriculture, Smart Greenhouses and Blockchain in the Supply Chain. Thus, this integration with agroindustrial production not only improves productivity and reduces costs, but also promotes more sustainable and responsible agricultural practices, benefiting producers and consumers. In that sense, there is still a lot of research work that delves into these topics.

The joint incorporation of robotics and AI in agroindustrial production provides a series of important benefits that can change the sector in terms of efficiency, sustainability and productivity. Some of the ways in which these technologies complement each other to be applied in agroindustrial practices are: Automation of Agricultural Tasks (Harvesting and Seeding Robots, and Automated Seeding), Crop Monitoring and Management (Drones for crop surveillance, Predictive Analysis, Optimization of Resource Use, Intelligent Irrigation, among others), Pest and Disease Management (Early Detection and Selective Application of Pesticides, Food Logistics and Processing (Sorting and Packaging, and Predictive Maintenance), Improving Animal Welfare (Livestock Monitoring and Cleaning and Automated Feeding), and Vertical Farming and Urban Farming (Autonomous Indoor Farming Systems). The combination of robotics and AI in agribusiness production not only improves efficiency and productivity, but also, drives sustainability by optimizing the use of resources and minimizing environmental impact. These advanced technologies offer solutions to meet the current and future challenges of agriculture, guaranteeing a safe and efficient food supply.

In general, the studies included in this article evidence a low use of AI techniques and ML algorithms to improve production in MSME agro-industrial production chains. However, it seems reasonable to assume that future solutions will need to implement ML techniques, which may reside in the cloud to obtain the benefits of a truly connected and intelligent AI ecosystem.

Conflict of interest

The authors declare that they have no competing financial interests or known personal relationships that could have influenced the work reported in this paper.

References

Addanki, M., Patra, P., & Kandra, P. (2022). Recent advances and applications of artificial intelligence and related technologies in the food industry. *Applied Food Research*, *2*(2), 100126. https://doi.org/10.1016/j.afres.2022.100126

Aguilar J., Buendia O., Pinto A. , Gutiérrez J. (2022) Social learning analytics for determining learning styles in a smart classroom, Interactive Learning Environments, 30(2), 245-261, DOI: 10.1080/10494820.2019.1651745

Arkeman, Y., Kailaku, S. I., Saefurahman, G., & Fatullah, R. (2022). *Special issue: Smart and Resilient Agri-Food Systems for Integrating Smallholder Farmers into Global Value Chains Applications of the Internet of Things (IoT) and Blockchain for Agriculture in Indonesia*. *3*, 57–68.

Benhamida, F. Z., Bennani, C., Casado-Mansilla, D., & López-De-Ipina, D. (2019). Toward a delay tolerant internet of things. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3365871.3365908

Caballero, R., & Rivera, B. (2019). Blockchain: An alternative to enable traceability in the agricultural supply chain in Panama. *Proceedings - 2019 7th International Engineering, Sciences and Technology Conference, IESTEC 2019*, 46–51. https://doi.org/10.1109/IESTEC46403.2019.00017

Carpio, F., Jukan, A., Sanchez, A. I. M., Amla, N., & Kemper, N. (2017). *Beyond Production Indicators*. https://doi.org/10.1145/3152130.3152140

Flores, F. (2022). Rot Corn Grain Classication by Color and Texture Analysis. *IEEE Latin America Transactions*, *20*(2), 208–214. https://doi.org/10.1109/TLA.2022.9661459

Chiba, Z., Abghour, N., Moussaid, K., el omri, A., & Rida, M. (2019). Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. *Computers and Security*, *86*, 291–317. https://doi.org/10.1016/j.cose.2019.06.013

Cordeiro, M., Markert, C., Araújo, S. S., Campos, N. G. S., Gondim, R. S., da Silva, T. L. C., & da Rocha, A. R. (2022). Towards Smart Farming: Fog-enabled intelligent irrigation system using deep neural networks. *Future Generation Computer Systems*, *129*, 115–124. https://doi.org/10.1016/j.future.2021.11.013

Das, P., Yadav, J. K. P. S., & Yadav, A. K. (2021). An automated tomato maturity grading system using transfer learning based AlexNet. *Ingenierie Des Systemes d'Information*, *26*(2), 191–200. https://doi.org/10.18280/isi.260206

Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, *117*, 257–262. https://doi.org/10.1016/j.ssci.2019.04.026

Diaz-Gonzalez, F. A., Vuelvas, J., Correa, C. A., Vallejo, V. E., & Patino, D. (2022). Machine learning and remote sensing techniques applied to estimate soil indicators – Review. *Ecological Indicators*, *135*, 108517. https://doi.org/10.1016/j.ecolind.2021.108517

Elavarasan, D., & Vincent, D. R. (2020). Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. *Journal of Intelligent and Fuzzy Systems*, *39*(5), 7605–7620. https://doi.org/10.3233/JIFS-200862

Emanuela, L., Cătălin, E., & Leonard, I. (n.d.). *Implementation of artificial intelligence systems in the agricultural field: agrodata project*. *XXIII*(1), 79–88.

Ericson, A., Lugnet, J., Solvang, W. D., Kaartinen, H., & Wenngren, J. (2020). Challenges of Industry 4.0 in SME businesses. *2020 3rd International Symposium on Small-Scale Intelligent Manufacturing Systems, SIMS 2020*, *825196*. https://doi.org/10.1109/SIMS49386.2020.9121542.

Fang, J. (2022). Smart phone based monitoring of agricultural activities. Computational algorithms and numerical dimensions, 1(4), 159-163.

Feijó, T., David, J. M., Braga, R., Otenio, M. H., Paula, V. R., Santos, G. M., Campos, F., & Stroele, V. (2021). @grogest Ambiental: A Web-based Decision Support System for agribusiness. *ACM International Conference Proceeding Series*, 81–84. https://doi.org/10.1145/3470482.3479612

Fiehn, H. B., Schiebel, L., Avila, A. F., Miller, B., & Mickelson, A. (2018). Smart agriculture system based on deep learning. *ACM International Conference Proceeding Series*, 158–165. https://doi.org/10.1145/3289100.3289126

Garcia, E., & Vieira, M. T. P. (2008). *Estudo de caso de mineração de dados multi-relacional: aplicação do algoritmo connetionblock em um problema da agroindústria*. 224–237. http://dl.acm.org/citation.cfm?id=1498932.1498954

García, R., Aguilar, J., Toro, M., Perez N., Pinto, A., & Rodríguez, P. (2023). Autonomic computing in a beef-production process for Precision Livestock Farming. Journal of Industrial Information Integration, *31*, 10.1016/j.jii.2022.100425.

García, V. H. M., López, J. F. Q., & Cely, S. M. B. (2014). Knowledge transfer to manage human talent processes at MSMEs through simulated enterprises. *MINDTREK 2014 - Proceedings of the 18th International Academic MindTrek Conference:* , 152–158. https://doi.org/10.1145/2676467.2676507

Gezgin, E., Huang, X., Samal, P., & Silva, I. (2017). Digital transformation : Raising supply-chain performance to new levels. *McKinsey & Company Operations*, *November*, 1–11.

Ghobakhloo, M., & Ching, N. T. (2019). Adoption of digital technologies of smart manufacturing in SMEs. *Journal of Industrial Information Integration*, *16*(January), 100107. https://doi.org/10.1016/j.jii.2019.100107

Groher, T., Heitkämper, K., Walter, A., Liebisch, F., & Umstätter, C. (2020). Status quo of adoption of precision agriculture enabling technologies in Swiss plant production. *Precision Agriculture*, *0123456789*. https://doi.org/10.1007/s11119-020-09723-5

Guisiano, J., Chiky, R., Orensanz, J., & Ahvar, S. (2020). Multilayered Ensemble Learning for short-term forecasting in agro-climatology. *ACM International Conference Proceeding Series*, 7–11. https://doi.org/10.1145/3397125.3397130

Guo, Y., Zhang, J., Yin, C., Hu, X., Zou, Y., Xue, Z., & Wang, W. (2020). Plant Disease Identification Based on Deep Learning Algorithm in Smart Farming. *Discrete Dynamics in Nature and Society*, *2020*. https://doi.org/10.1155/2020/2479172

Gupta, S., Geetha, A., Sankaran, K. S., Zamani, A. S., Ritonga, M., Raj, R., Ray, S., & Mohammed, H. S. (2022). Machine Learning-and Feature Selection-Enabled Framework for Accurate Crop Yield Prediction. *Journal of Food Quality*, *2022*. https://doi.org/10.1155/2022/6293985

Gutarra, R., & Valente, A. (2018). Las mipymes tecnológicas peruanas al 2030. Estrategias para su inserción a la industria 4.0. *Nova Scientia*, *10*(1), 754–778.

Hajjaji, Y., Boulila, W., Farah, I. R., Romdhani, I., & Hussain, A. (2021). Big data and IoT-based applications in smart environments: A systematic review. *Computer Science Review*, *39*, 100318. https://doi.org/10.1016/j.cosrev.2020.100318

Happila, T., Dhinesh, S., Nandhababu, N., Puhazhenthi, C., & Sriram, G. (2020). Smart agriculture monitoring system using IoT. *Test Engineering and Management*, *82*, 10973–10977. https://doi.org/10.14419/ijet.v7i2.7.10603

Heilbrunn, S., Rabin, N., & Rozenes, S. (2017). Detecting mutual configurations of applied planning strategies and performances in small and medium sized businesses with kernel based machine learning methods. *Applied Soft Computing Journal*, *61*, 1211–1225. https://doi.org/10.1016/j.asoc.2017.08.054

Huang, J. T., Li, J., & Gong, Y. (2015). An analysis of convolutional neural networks for speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2015-Augus*, 4989–4993. https://doi.org/10.1109/ICASSP.2015.7178920

]. C. Silva, D. Baker, S. Miranda. (2013). *Agroindustrias para el Desarrollo*.

Jagannath, J., Polosky, N., Jagannath, A., Restuccia, F., & Melodia, T. (2019). Machine learning for wireless communications in the Internet of Things: A comprehensive survey. *Ad Hoc Networks*, *93*, 101913. https://doi.org/10.1016/j.adhoc.2019.101913

Jiménez Cercado, M. E., Acosta Véliz, M. M., Muñoz Naranjo, D. I., & García Moreira, E. (2020). La relación entre el uso de las NTIC en la competitividad en las micro y pequeñas empresas comerciales de Guayaquil. *3C TIC: Cuadernos de Desarrollo Aplicados a Las TIC*, *9*(2), 119–137. https://doi.org/10.17993/3ctic.2020.92.119-137

Joshi, J., Polepally, S., Kumar, P., Samineni, R., Rahul, S. R., Sumedh, K., Tej, D. G. K., & Rajapriya, V. (2017). Machine learning based cloud integrated farming. *ACM International Conference Proceeding Series*, 1–6. https://doi.org/10.1145/3036290.3036297

Kamyshova, G., Osipov, A., Gataullin, S., Korchagin, S., Ignar, S., Gataullin, T., Terekhova, N., & Suvorov, S. (2022). Artificial Neural Networks and Computer Vision's-Based Phytoindication Systems for Variable Rate Irrigation Improving. *IEEE Access*, *10*, 8577–8589. https://doi.org/10.1109/ACCESS.2022.3143524

Kim, Y. H., Ha, J. H., Yoon, Y., Kim, N. Y., Im, H. H., Sim, S., & Choi, R. K. Y. (2016). Improved Correction of Atmospheric Pressure Data Obtained by Smartphones through Machine Learning. *Computational Intelligence and Neuroscience*, *2016*. https://doi.org/10.1155/2016/9467878

Kumar, I., Rawat, J., Mohd, N., & Husain, S. (2021). Opportunities of Artificial Intelligence and Machine Learning in the Food Industry. *Journal of Food Quality*, *2021*. https://doi.org/10.1155/2021/4535567

Kundu, N., Rani, G., & Dhaka, V. S. (2021). Seeds Classification and Quality Testing Using Deep Learning and YOLO v5. *ACM International Conference Proceeding Series*, 153–160. https://doi.org/10.1145/3484824.3484913

Lakshmi, G. J., Hilmi, S. A. N., & Obaid, A. J. (2022). Artificial intelligence agriculture recommendation model (AIARM). *International Journal of Health Sciences*, *April*, 1782–1808. https://doi.org/10.53730/ijhs.v6ns2.5370

Li, Y., Jiang, W., Yang, L., & Wu, T. (2018). On neural networks and learning systems for business computing. *Neurocomputing*, *275*, 1150–1159. https://doi.org/10.1016/j.neucom.2017.09.054

Lima, L. M., Cavalcante, V. C., de Sousa, M. G., Fleury, C. A., Oliveira, D., & de Andrade Freitas, E. N. (2021). Artificial Intelligence in Support of Welfare Monitoring of Dairy Cattle: A Systematic Literature Review. *Proceedings - 2021 International Conference on Computational Science and Computational Intelligence, CSCI 2021*, 1708–1715. https://doi.org/10.1109/CSCI54926.2021.00324

Lopez, C. P., Santórum, M., & Aguilar, J. (2019). Autonomous Cycles of Collaborative Processes for Integration Based on Industry 4.0. *Advances in Intelligent Systems and Computing*, *918*, 177–186. https://doi.org/10.1007/978-3-030-11890-7_18

Lopez, H. A. G., & Cisneros, M. A. P. (2017). Industry 4.0 & Internet of things in supply chain. *ACM International Conference Proceeding Series*, 1–4. https://doi.org/10.1145/3151470.3156646

Luiz Carlos, M., & Ulson, J. A. C. (2021). Real time weed detection using computer vision and deep learning. *2021 14th IEEE International Conference on Industry Applications, INDUSCON 2021 - Proceedings*, 1131–1137. https://doi.org/10.1109/INDUSCON51756.2021.9529761

Luna, M., Llorente, I., & Cobo, A. (2022). Determination of feeding strategies in aquaculture farms using a multiple-criteria approach and genetic algorithms. *Annals of Operations Research*, *314*(2), 551–576. https://doi.org/10.1007/s10479-019-03227-w

Martini, B. G., Helfer, G. A., Barbosa, J. L. V., da Silva, M. R., de Figueiredo, R. M., Modolo, R. C. E., & Yamin, A. C. (2019). A computational model for ubiquitous intelligent services in indoor agriculture. *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019*, 497–500. https://doi.org/10.1145/3323503.3360641

Meyer, M., & Dykes, J. (2020). Criteria for Rigor in Visualization Design Study. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 87–97. https://doi.org/10.1109/TVCG.2019.2934539

Mirani, A. A., Memon, M. S., Rahu, M. A., Bhatti, M. N., & Shaikh, U. R. (2019). A Review of Agro-Industry in IoT: Applications and Challenges. *Quaid-E-Awam University Research Journal of Engineering, Science & Technology, Nawabshah.*, *17*(1), 28–33.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7). https://doi.org/10.1371/journal.pmed.1000097

Monsalve-Pulido, J., Aguilar, J., Montoya, E. Salazar, C. (2024), Autonomous recommender system architecture for virtual learning environments", *Applied Computing and Informatics*, Vol. 20, pp. 69-88. 10.1016/j.aci.2020.03.001

Nath, S., Dey, A., Das, P., Mohapatra, D., Sing, J. K., & Sarkar, S. K. (2022). Application of Soil Sensors for Maximizing Productivity Using IoT Framework. *Proceedings of IEEE VLSI DCS 2022: 3rd IEEE Conference on VLSI Device, Circuit and System*, *2022*(February), 220–224. https://doi.org/10.1109/VLSIDCS53788.2022.9811456

Navarro, E., Costa, N., & Pereira, A. (2020). A systematic review of iot solutions for smart farming. *Sensors (Switzerland)*, *20*(15), 1–29. https://doi.org/10.3390/s20154231

Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, *29*(July), 100367. https://doi.org/10.1016/j.sbsr.2020.100367.

Nourkhah, S. A., Cirovic, G., & Edalatpanah, S. A. (2023). The role of sensors in smart agriculture. Computational algorithms and numerical dimensions, 2(4), 210-215.

Nylander, S., Wallberg, A., & Hansson, P. (2017). Challenges for SMEs entering the IoT world - Success is about so much more than technology. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3131542.3131547

Oliver, S. T., González-Pérez, A., & Guijarro, J. H. (2018). An IoT proposal for monitoring vineyards called senviro for agriculture. *ACM International Conference Proceeding Series*, *1*. https://doi.org/10.1145/3277593.3277625

Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. (2014). Methodological framework for data processing based on the Data Science paradigm," *Proceeding XL Latin American Computing Conference (CLEI)*, Montevideo, Uruguay, 2014, pp. 1-12, doi: 10.1109/CLEI.2014.6965184.

Panda A, Edalatpanah S, Godarzi R. "Improve Crop Production Through WSN: An Approach of Smart Agriculture". Big. Data. Comp. Vis. Vol. 1, No. 2 (2021) 71–82.

Pandey, A., Kumar, S., Tiwary, P., & Das, S. K. (2019). A hybrid classifier approach to multivariate sensor data for climate smart agriculture cyber-physical systems. *ACM International Conference Proceeding Series*, 337–341. https://doi.org/10.1145/3288599.3288621

Parasuraman, K., Anandan, U., & Anbarasan, A. (2021). IoT based smart agriculture automation in artificial intelligence. *Proceedings 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks, ICICV 2021*, *Icicv*, 420–427. https://doi.org/10.1109/ICICV50876.2021.9388578

Park, J., Choi, J. H., Lee, Y. J., & Min, O. (2017). A layered features analysis in smart farm environments. *ACM International Conference Proceeding Series*, 169–173. https://doi.org/10.1145/3175684.3175720

Pereira, D. F., Bugatti, P. H., Lopes, F. M., de Souza, A. L. S. M., & Saito, P. T. M. (2019). An intelligent system to enhance the productivity and sustainability in soybean crop enterprises. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. https://doi.org/10.1145/3330204.3330266

Pinheiro, T. C., Azevedo dos Santos, J. A., & Pasa, L. A. (2020). Gestão Da Produção De Frangos De Corte Por Meio De Redes Neurais Artificiais. *Holos*, *2*, 1–15. https://doi.org/10.15628/holos.2020.9043

Pradana, I. G. M. T., Djatna, T., & Hermadi, I. (2020). Blockchain modeling for traceability information system in supply chain of coffee agroindustry. *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, 217–224. https://doi.org/10.1109/ICACSIS51025.2020.9263214

Ram, C. R. S., Ravimaran, S., Krishnan, R. S., Julie, E. G., Robinson, Y. H., Kumar, R., Son, L. H., Thong, P. H., Thanh, N. Q., & Ismail, M. (2020). Internet of Green Things with autonomous wireless wheel robots against green houses and farms. *International Journal of Distributed Sensor Networks*, *16*(6). https://doi.org/10.1177/1550147720923477

Ramirez-Asis, E., Bhanot, A., Jagota, V., Chandra, B., Hossain, M. S., Pant, K., & Almashaqbeh, H. A. (2022). Smart Logistic System for Enhancing the Farmer-Customer Corridor in Smart Agriculture Sector Using Artificial Intelligence. *Journal of Food Quality*, *2022*. https://doi.org/10.1155/2022/7486974

Reddy, P. K., & Rajanna, G. S. (2022). *A Study on Various Applications of IoT Using Machine Learning*. Iconic Research and Engineering Journals, 5(8), 334–337.

Rehman, M., Razzaq, A., Baig, I. A., Jabeen, J., Tahir, M. H. N., Ahmed, U. I., Altaf, A., & Abbas, T. (2022). Semantics Analysis of Agricultural Experts' Opinions for Crop Productivity through Machine Learning. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2012055

Roukh, A., Fote, F. N., Mahmoudi, S. A., & Mahmoudi, S. (2020). WALLeSMART: Cloud Platform for Smart Farming. *ACM International Conference Proceeding Series*, 1–4. https://doi.org/10.1145/3400903.3401690

Sampaio, D., & Bernardino, J. (2014). Open source CRM systems for SMEs. *ACM International Conference Proceeding Series*, 154–157. https://doi.org/10.1145/2641483.2641526

Sanchez, M., Exposito, E., & Aguilar, J. (2020). Implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context. *Computers in Industry*, Volume 121, https://doi.org/10.1016/j.compind.2020.103247.

Sánchez, M., Aguilar, J., Cordero, J., Valdiviezo-Díaz, P., Barba-Guamán, L., Chamba-Eras, L. (2016). Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service. INew Advances in Information Systems and Technologies (Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M. (eds)). Advances in Intelligent Systems and Computing, Springer. 993–1002, vol 444.

Sanchez-Calle, J. E., & Castillo Armas, G. P. (2022). Algoritmos y su efecto en la agricultura. *Revista Científica de Sistemas e Informática*, *2*(2), e386. https://doi.org/10.51252/rcsi.v2i2.386

Schulte, L. G., Perez, N. B., de Pinho, L. B., & Trentin, G. (2019). Decision support system for precision livestock: Machine learning-based prediction module for stocking rate adjustment. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3330204.3330222

Sen, D., Ozturk, M., & Vayvay, O. (2016). An Overview of Big Data for Growth in SMEs. *Procedia - Social and Behavioral Sciences*, *235*(October), 159–167. https://doi.org/10.1016/j.sbspro.2016.11.011

Sharaf, M., Abusair, M., Eleiwi, R., Shana'a, Y., Saleh, I., & Muccini, H. (2019). Architecture description language for climate smart agriculture systems. *ACM International Conference Proceeding Series*, *2*, 152–155. https://doi.org/10.1145/3344948.3344992

Shreyas, K. P., & Narayana, G. V. S. (2022). *Bell Pepper Crop Disease Detection Using Convolution Neural Network*. 1–6. https://doi.org/10.1109/iccsea54677.2022.9936171

Singh, K. U., Kumar, A., Raja, L., Kumar, V., Singh Kushwaha, A. K., Vashney, N., & Chhetri, M. (2022). An Artificial Neural Network-Based Pest Identification and Control in Smart Agriculture Using Wireless Sensor Networks. *Journal of Food Quality*, *2022*. https://doi.org/10.1155/2022/5801206

Sokolov, I. A., Murin, I. v., Wiemhöfer, H. D., & Pronkin, A. A. (1998). Electric conductivity and the nature of electric current carriers in the PbF2-2PbO · SiO2 glasses. In *Glass Physics and Chemistry* (Vol. 24, Issue 2).

Stackpole, B. (2021). You're hired: Recruiting Mobile Robots. In *Automation World* (Issue January, pp. 44–48).

Stewart, J., Stewart, R., & Kennedy, S. (2017). Dynamic IoT management system using K-means machine learning for precision agriculture applications. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3018896.3036385

Suarez, A., Jimenez, A., Castro-franco, M., & Cruz-roa, A. (2017). *Clasificación y mapeo automático de coberturas del suelo en imágenes satelitales utilizando Redes Neuronales Convolucionales*. *21*(1). Orinoquia

Suarez, A. J. B., Singh, B., Almukhtar, F. H., Kler, R., Vyas, S., & Kaliyaperumal, K. (2022). Identifying Smart Strategies for Effective Agriculture Solution Using Data Mining Techniques. *Journal of Food Quality*, https://doi.org/10.1155/2022/6600049

Talavera, J. M., Tobón, L. E., Gómez, J. A., Culman, M. A., Aranda, J. M., Parra, D. T., Quiroz, L. A., Hoyos, A., & Garreta, L. E. (2017). Review of IoT applications in agro-industrial and environmental fields. *Computers and Electronics in Agriculture*, *142*(118), 283–297. https://doi.org/10.1016/j.compag.2017.09.015

Toyeer-E-Ferdoush, Tahsin, M., & Taher, K. A. (2020). Innovative smart farming system with WiMAX and solar energy. *ACM International Conference Proceeding Series*, 10–11. https://doi.org/10.1145/3377049.3377063

Toscano-Miranda, R., Toro, M., Aguilar, J., Caro, M., Marulanda, A., Trebilcok, A. (2022). Artificial-intelligence and sensing techniques for the management of insect pests and diseases in cotton: A systematic literature review. *The Journal of Agricultural Science*, 160(1-2), 16-31. doi:10.1017/S002185962200017X

van Witteloostuijn, A., & Kolkman, D. (2019). Is firm growth random? A machine learning perspective. *Journal of Business Venturing Insights*, *11*(December), e00107. https://doi.org/10.1016/j.jbvi.2018.e00107

Vizcarrondo J., Aguilar J. Exposito E., Subias A. (2012) "ARMISCOM: Autonomic reflective middleware for management service composition," *2012 Global Information Infrastructure and Networking Symposium (GIIS)*, doi: 10.1109/GIIS.2012.6466760.

Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS - Wageningen Journal of Life Sciences*, *90–91*(December 2018), 100301. https://doi.org/10.1016/j.njas.2019.04.007

Zhang, Y., Wei, S., Zhang, L., & Liu, C. (2019). Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features. *Journal of Medical and Biological Engineering*, *39*(3), 381–392. https://doi.org/10.1007/s40846-018-0411-0

Zhao, S., Peng, Y., Liu, J., & Wu, S. (2021). Tomato leaf disease diagnosis based on improved convolution neural network by attention module. *Agriculture (Switzerland)*, *11*(7). https://doi.org/10.3390/agriculture11070651

Zhou, Z. (2023).Soil quality based agricultural activity through iot and wireless sensor network. Big data and computing visions, 3(1), 26-31.

Appendix B. Autonomous cycles of data-analysis tasks for the automation of the production chain of MSMEs for the agroindustrial sector.

information

MDPI

Article

Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector

Jairo Fuentes [1,2], Jose Aguilar [2,3,4,*], Edwin Montoya [2] and Ángel Pinto [5]

1 GIDIS, Universidad Francisco de Paula Santander, Av. Gran Colombia, Cúcuta, 540001, Colombia; fuentesjairo@gmail.com
2 GIDITIC, Universidad EAFIT, Carrera 48 No. 7 Sur-50, Medellín, 050001, Colombia; e jafuentesc@eafit.edu.co, jlaguilarc@eafit.edu.co, montoya@eafit.edu.co
3 CEMISID, Facultad de Ingeniera, Universidad de Los Andes, Mérida, 5101, Venezuela
4 IMDEA Networks Institute, Leganés, 28910, Madrid, Spain
5 Universidad del Sinú, Montería, 230001, Colombia; angelpinto@unisinu.edu.co
* Correspondence: aguilarjos@gmail.com or aguilar@ula.ve

Abstract: In this paper, we propose autonomous cycles of data analysis tasks for the automation of the production chains aimed to improve the productivity of Micro, Small and Medium Enterprises (MSMEs) in the context of agroindustry. In the autonomous cycles of data analysis tasks, each task interacts with the others and has different functions, in order to reach the goal of the cycle. In this article, we identify three industrial-automation processes within the production chain, in which autonomous cycles can be applied. The first cycle is responsible to identify the type of input to be transformed—such as quantity, quality, time, and cost—based on information from the organization and its context. The second cycle selects the technological level used in the raw-material transformation, characterizing the platform of plant processing. The last cycle identifies the level of specialization of the generated product, such as the quality and value of the product. Finally, we apply the first autonomous cycle to define the type of input to be transformed in a coffee factory.

Keywords: production-chain; agroindustry; autonomous computing; artificial intelligence; data analysis; machine learning

1. Introduction

One of the great current challenges of Micro, Small and Medium Enterprises (MSMEs) is to dynamically innovate to improve their supply of goods, products, and services in order to respond to the changing needs of the market [1,2]. In particular, several studies have concluded that investment in innovation has a high impact on the

Citation: Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector. *Information* **2024**, *15*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Received: 11 December 2023
Revised: 3 January 2024
Accepted: 6 January 2024
Published: date

competitiveness of organizations, which can lead to the introduction of new products and processes. Thus, innovation is a means for companies to adapt to remain in the market, considering available resources.

Given the importance of innovation in MSMEs, and the opportunities that currently exist to exploit data from organizations and their contexts, data strategies can be defined to build models that guide the automation process in an organization. One of these strategies is the use of autonomous data analysis task cycles (ACODATS) defined in previous works [3] that allow the generation of knowledge models using different data sources for automation management. An ACODAT is composed of a set of data analysis tasks to achieve a given objective, such that each task has a specific role [4,5,6,7,8]: some observe the system, others analyze it, and finally, others make decisions to improve it. Thus, in ACODAT, there are interactions between the data analysis tasks in order to achieve the automation objective for which it was defined.

This paper proposes three ACODATs for the automation of the production chain of MSMEs for the agroindustrial sector to improve their competitiveness. These ACODATs allow automating the main subprocesses that have previously been determined as vital to enable the self-management of industrial automation of MSMEs in the agroindustrial sector to improve their competitiveness. The first ACODAT determines the type of input to be transformed in the agroindustrial productive chain using information from the organization and the context, the next ACODAT specifies the technological level required in the transformation process of the agroindustrial productive chain, and the last ACODAT establishes the level of production specialization. This paper also includes a detailed specification of the first autonomous cycle, which goal is the definition of the type of input to be transformed, in the context of the coffee factories. For the specification of the ACODATs, the MIDANO methodology [4] was used, which allows the development of data analysis applications and, in particular, the development of ACODATs. The main contributions of this work are the following:

The specification of three ACODATs to manage agroindustrial automation to improve the productive chains. These three ACODATs automate the most relevant subprocesses to enable self-management of industrial automation for MSMEs in the agroindustrial sector.

A multidimensional data model to manage industrial automation, which stores the necessary information of an organization and its context.

A detailed description of the ACODAT to define the type of input to be transformed in a coffee factory.

In the review of the existing literature (see section 2), it is confirmed that the great gap is that industrial automation processes have not been developed for the MSME sector based on integrated data analysis tasks, but rather isolated tasks have been developed, some using artificial intelligence techniques, to solve specific problems. The great novelty that this article proposes is that defines autonomous cycles of data analysis tasks, which integrate these tasks with the objective of allowing industrial automation of the production chain in MSMEs.

The organization of the work is the following: related work is introduced in Section 2. Afterward, in Section 3, the theoretical framework of this work is introduced. In Section 4, the identification of the agroindustrial processes of the production chains is described. After, in Section 5, the definition of the ACODATs is carried out. Section 6 presents the Multi-Dimensional Data Model for the autonomous cycles and in Section 7, the instantiation of the ACODAT to define the type of input to be transformed for the case of "Café Galavis" company is described. Section 8 presents a comparison with previous works. Finally, in Section 9, conclusions and future work directions are presented.

2. Related Work

In this section, we present the main recent works related to this paper's approach, which are the definition of schemes for the automation of agroindustrial production chains and the use of data analysis tasks in the automation of agroindustrial processes.

The concept of agroindustry refers to the establishment of links between companies and supply chains to develop, transform and distribute specific inputs and products in the agricultural sector. As an example, according to [9], in the raw sugar-cane cubes (in Spanish, panela) production chain, the following activities are intertwined: (i) agricultural production, (ii) transformation, (iii) intermediate and final commercialization; and (iv) consumption. These tasks are carried out by different production systems that are in different sectors of the economy: agriculture, manufacturing, and services. As another example, the structure of the coffee agroindustrial sector comprises five activities: (i) agricultural production, (ii) processing, (iii) roasting, (iv) marketing, and (v) export [9].

According to [10], a production chain is a relationship between companies to connect the stages of supplying inputs, manufacturing, distribution, and marketing of a specific good, where the different links make agreements that condition and subordinate their technical and productive processes. This relationship seeks to become competitive at the national and international level, by strengthening the value chain in organizations and increasing the added value of their products. Finally, in general, according to [11], a production scheme includes the following links: (i) producers of raw materials, (ii) transporters, (iii) collectors, (iv) industrial processors, (v) distributors, and (vi) final consumers.

On the other hand, in some agroindustrial sectors, smallholders have applied Machine Learning (ML) techniques for land classification to analyze climate change and monitor ecosystem service. Also, there is a work about the processing of satellite images using artificial intelligence methods (specifically, using a convolutional neural network and a genetic algorithm) to convert the images into useful data for decision making, precision agriculture and agribusiness [12]. Another example is the utilization of ML methods to estimate land-cover change [6]. Recently, computational-learning algorithms—such as Support Vector Machine (SVM) and Random Forest (RF)—have been used for automatic land classification with overall accuracy results of 86.5% [6] and 95.10%, respectively [13].

According to [14], technologies such as smart agents, Big Data, Internet of Things (IoT), Cloud Computing (CC), ML, and data mining aim to implement data and information exchange along a supply chain [15,16]. In the sugar sector, a case study using a connection-block algorithm and a multirelational data mining approach in a database of this agroindustrial sector was analyzed in [17].

In [18], the authors present an intelligent monitoring application for Industry 4.0, for a factory that operates in the agroindustrial sector. The authors present a consumption management system based on the widespread installation of sensors in production lines and the design of software to access the data [18]. On the other hand, the use of industrial robots in food manufacturing remains a challenge. Despite the significant reduction in production costs in the last decades, its adoption in food processing has been slow [19]. Finally, in [20], ML algorithms (SVM, Boosted Trees and Naïve Bayes) have been applied to predict the severity of worker injuries in industrial processes; in addition to the nature and cause thereof, with an accuracy of 92–98% [20].

Borghesan et al. [21] provide an analysis of process and energy industries, considering the different mechanical, sensing, situational awareness, and decision-making tasks involved in the operation of plants. Then, they map these tasks to possible autonomous systems, and as part of autonomous system capabilities, they make a connection to the adaptation of model-based solutions. Finally, they argue that reaching higher and wider levels of autonomy requires a rethink of the design processes for both the physical plants as well as the way automation, control, and safety solutions are conceptualized. The paper [22] analyzes the effect of smart manufacturing systems on procurement processes by distinguishing between procurement tasks that will likely stay in human hands and those that will be taken over by intelligent systems. In addition, the paper presents an operational, fully automated, and manufacturing-related procurement system. For this, they begin by looking at the manufacturing-related procurement types in the era of Industry 4.0 that may be performed by machines, followed by the discussion of innovation-based procurement practices that will require human input.

All previous works focused on the use of artificial intelligence techniques in specific tasks (e.g., soil, climate, environment, and humidity monitoring); however, this analysis is not applied to the agroindustrial MSME sector. Likewise, they did not integrate into autonomic cycles of data analysis tasks for the industrial automation of the production chain. To overcome the restrictions found in previous works, this article focuses on the definition of autonomic cycles for optimal decision making on industrial automation of production chains of agroindustrial MSMEs to improve their competitiveness.

3. Background

This section presents the theoretical basis related to the field of ACODAT, which consists of a set of data analysis tasks that act together to achieve an objective in the process they supervise. The tasks interact with each other and have different roles in the cycle. Also, this section presents the methodology for the specification of data analytics tasks (MIDANO), which is designed for the development of ACODAT for the processes of any institution/company.

3.1. ACODAT

The main goal of ACODAT is to view business problems from a data perspective [4]. The set of data analysis tasks must be performed simultaneously for the purpose of achieving a goal in the monitored process. This set of tasks is related to each other, and each task has a different role in the cycle (see Figure 1): (i) to observe the process, (ii) to analyze and interpret what happens in it, and (iii) to make decisions about the process, which allow achieving the objective for which the cycle was designed. This insertion of tasks in a closed cycle, allows complex problems to be solved. The functions of each task are described below [3,4].

Observation Tasks: They correspond to tasks that monitor the process under analysis. They capture data and information that describe the behavior of the environment, and eventually, they are also responsible for their generation.

Analysis Tasks: These are a set of tasks whose purpose is to interpret, understand and diagnose, from the data, what may be happening in the context monitored by the cycle. This means that these tasks build knowledge models about the dynamics observed by the cycle.

Decision-making Tasks: This set of tasks is in charge of implementing the tasks for decision making, leading to the improvement of the process where the cycle is applied. These tasks modify the dynamics of the process to improve it, and their effects are again evaluated in the observation and analysis stages of the cycle, restarting a new one.

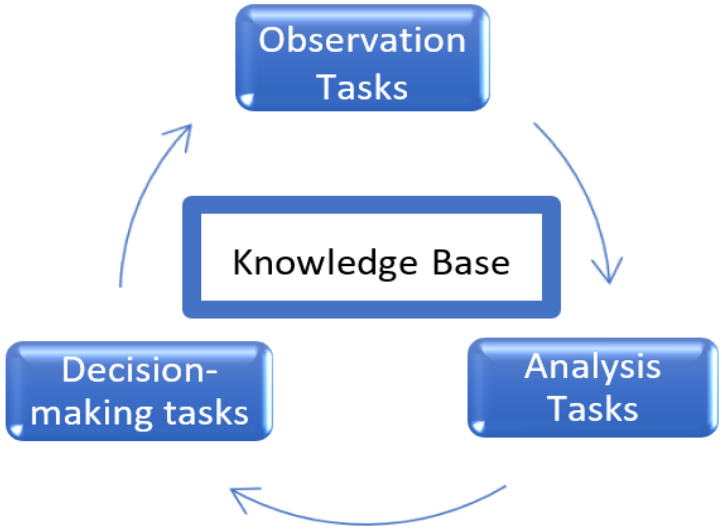

Figure 1. ACODAT task.

Observation tasks monitor and collect data and information about the monitored system or environment. This collected information is interpreted by the analysis tasks, to understand what may be happening in the system. Finally, according to the analysis carried out before, the decision-making tasks determine the activities to improve the system. Observation tasks monitor and collect data and information about the monitored system or environment. This collected information is interpreted by the analysis tasks to understand what may be happening in the system. Finally, according to the analysis carried out before, the decision-making tasks determine the activities to improve the system. It is a closed loop, where the supervision process is permanently carried out on the system under study.

ACODAT was suggested in [3], is based on concepts put forward by IBM in 2001 [23,24] and has been implemented in areas such as smart classrooms [25], telecommunications [26], Industry 4.0 [7], and smart cities [4]. ACODAT is based on the paradigm of autonomic computing [8, 27, 28], with the purpose of providing autonomic features to systems using smart control cycles. In general, an ACODAT requires:

A multidimensional data model to store the data collected to characterize the behavior of the context, which will be used by the different data analysis tasks.

A platform to integrate the technological tools required by the data analysis tasks.

The ACODAT concept has been successfully used in different fields, but it has not been applied to the processes of the agroindustrial MSME production chain. This will enable MSMEs to have autonomous management of their production processes.

3.2. MIDANO

MIDANO is a methodology for the development of applications based on data analysis for any organization [4], which is composed of three phases:

Phase 1. Identification of data sources for the extraction of knowledge about the organization: This phase performs knowledge engineering about the organizations, particularly its processes and its experts. This information is useful to define the objective of the application of data analysis in the organization. Likewise, the autonomous cycles with their data analysis tasks are designed.

Phase 2. Data preparation and processing: This phase prepares the data to be able to apply data analysis tasks to the problem under study. To do this, operations are performed on the data, such as extraction and transformation of each of the variables associated with the problem. In particular, a feature engineering process is carried out that analyzes the variables to define the variables of interest and prepare them (for example, cleaning, transforming, and reducing them). Finally, in this phase, the multidimensional data model of the autonomous cycles is designed, which is the structure of the data warehouse.

Phase 3. Development of data analytics tasks: In this phase, the data analysis tasks of the autonomous cycle are implemented. Each of them generates a knowledge model (for example, predictive, descriptive, or diagnostic models) required by the cycle. Subsequently, these tasks are integrated to form the autonomous cycle. This phase can use existing data mining methodologies for the development and validation of data analysis tasks.

4. Definition of Autonomous Cycles for the Agroindustrial Sector

This section analyzes the agroindustrial production chain of MSMEs, using the MIDANO methodology, to define the processes in which the autonomous cycles will operate to improve their competitiveness. The tables used in this section are defined by MIDANO.

4.1. Application of MIDANO to the Agroindustrial Production Chain of MSMEs

Table 1 summarizes the use of the MIDANO to analyze the agroindustrial production chain of MSMEs.

Table 1. Use of MIDANO phases for Industrial Automation Processes**.**

Phase	Use
Phase 1	Analysis of the production chain in the agroindustrial to improve their competitiveness. To this end, this research proposes ACODAT to improve industrial production.
Phase 2	Identification of data sources (e.g., quantity, quality, time, and cost).
Phase 3	Implementation of autonomic cycles for the automation of the production chains in agroindustry.

4.2. Agroindustrial Production Chain of MSMEs

Based on an analysis of agribusiness, it is possible to establish the macro conditions under which the production chains operate. In general, there is a wide range of activities and, therefore, actors that are part of a chain. It is, therefore, necessary to clearly establish the limits of the production chain and define its operating structure by establishing a model. The work [29] proposes a production chain model (see Figure 2), the basis of our proposal.

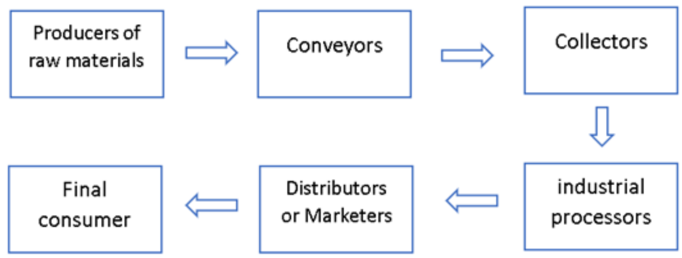

Figure 2. Production chain links (source: United Nations Industrial Development Organization).

In this context, Nonaka [30] has defined a scheme of the production chain and its main actors (see Figure 3).

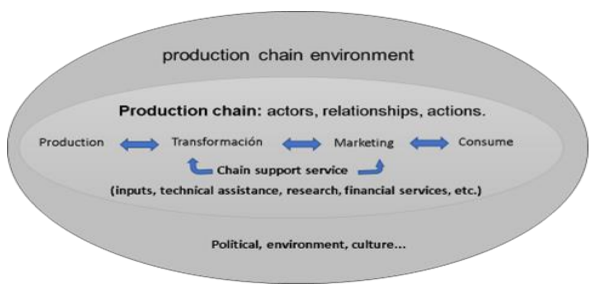

Figure 3. The production chain and its actors (source [30]).

As a starting point from such a model, it is necessary to understand the framework of production chains and establish how they are conceived, what their constituent elements are, and what characteristics they should have. The steps to follow to properly model a production chain are summarized below [31,32].

Establish the links in the production chain (defined by the blue arrows in Figure 3).

Determine the segments that make up each of the links in the production chain by using segmentation instruments and their corresponding variables.

Represent the material and capital flows that take place in the chain.

Establish the institutional and organizational environment of the chain.

The above activities must be accompanied by a validation by the chain actors. For the construction of the proposal, 15 experts from the MSME agribusiness sector validated each of the links in the production chain. Figure 4 shows our proposal of an Agroindustrial Production Chain for MSMEs.

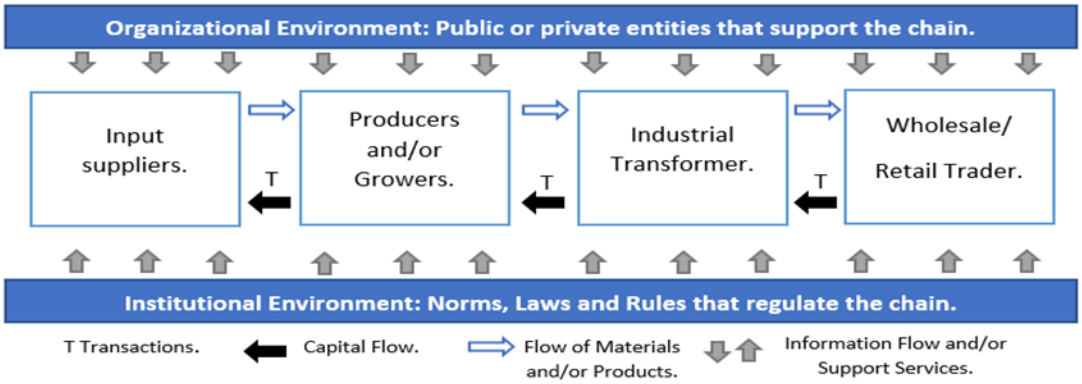

Figure 4. Our Proposal of an Agroindustrial Production Chain for MSMEs.

Our proposal has four processes (see Table 2): Input Suppliers, Producers and/or Growers, Industrial Transformer and Marketing.

Table 2. Processes in our agroindustrial production chain.

Process	Examples
Suppliers of Input Materials	Irrigation systems, organic and inorganic fertilizers, certified seeds.
Producers and/or Growers	Land tenure, area, labor force, technological level, degree of specialization, market share, working capital, forestry support services, agricultural support activities, post-harvest activities, seed processing.
Industrial Transformer	Size of the property, labor force, type of input to be processed, quality parameters, technological level, the added value of the product, market scope and coverage, level of specialization of the business.
Wholesaler/Retailer	Stockpiling and Distribution. Storage, Classification, Standardization, Packaging and Transportation.

In what follows, each process is described [91,7,29,33,34]:

Input Suppliers: the entities that provide or supply certain products or services to companies for their use, for example, agrochemicals and packaging.

Producers and/or Growers: this process is in charge of the harvest and post-harvest preparation. It must consider, among other things, the workmanship, the forestry support services and post-harvest activities.

Industrial Transformer: this process transforms or adapts the inputs for the materialization of the intended products or services. It must consider, among other things, The type of input to be transformed, the technology required for the transformation, etc.

Wholesaler/Retailer Commercialization: this process is the distribution of products or services to the market. It must consider, among other things, stockpiling and marketing.

Each process of the Agroindustrial Production Chain of MSMEs has different subprocesses, which should be prioritized according to whether data analysis tasks can be used and what its relevance is to improve the competitiveness of an MSME in the agroindustrial sector. We identify 15 subprocesses, listed in Table 3.

Table 3. Subprocesses in the Agroindustrial Production Chain of MSMEs.

Process	Subprocess	ACRONYM
Input suppliers	Certified seeds	SCS
	Organic and inorganic fertilizers	AOEI
	Primary, secondary and tertiary packaging	EPST
	Applications of agrochemicals in particular and fertilizers.	AAPF
Producers and/or growers	Workmanship	MOEP
	Forestry support services	SAAF
	Post-harvest activities	APAC
	Technological level	NTAP
Industrial transformer	Type of input to be processed	TIAT
	Transformer technology level	NTAT
	Market reach and coverage.	ACDM
	Level of business specialization.	NEDN
Wholesaler/retailer commercialization	Stockpiling	ACOP
	Leveling	NIVE
	Distribution	DIST

4.3. Prioritization of Subprocesses of the Agroindustrial Production Chain of MSMEs

A prioritization table has been used to select the subprocesses. The criteria to evaluate the relevance of the subprocesses were defined according to the importance of each subprocess in the MSME Agroindustrial Production Chain, and the possibility of performing data analysis tasks. Thus, these values determine the level of importance of each subprocess (see Table 4).

Table 4. Evaluation metrics.

Meaning	Weight
Subprocess is not important	1
Subprocess is slightly important	2
Subprocess is important	3
Subprocess is very important	4

For the construction of the prioritization table, 15 experts in the MSME Agroindustrial sector, and 10 research professors rated each of the criteria. In the result, each of the answers provided by them was averaged. The results are shown in Table 5, where the columns with the numbers in red represent the subprocesses with the highest priority (they represent the highest scores), for which the ACODATs are proposed in this work.

Table 5. Prioritized subprocesses by experts.

Weight	Evaluation Criteria	Processes.														
		Input suppliers.				Producers and/or Growers.				Industrial Transformer.				Wholesaler / Retailer.		
		SCS	AOEI	EPST	AAPF	MOEP	SAAF	APAC	NTAP	TIAT	NTAT	ACDM	NEDN	ACOP	NIVE	DIST
Relevance to Production Management.																
4	the factors involved in the process are characterised.	4	3	2	4	3	3	4	4	4	4	4	4	3	3	4
4	the uses and functions of the materials and tools used are distinguished	3	4	3	4	2	2	4	4	4	4	4	3	2	3	3
4	information and knowledge management is identified	2	3	3	4	2	3	3	4	4	4	4	4	2	3	4
4	Production, service and support processes are identified.	2	4	3	4	3	4	4	4	4	4	4	3	2	3	4
4	Environmental responsibility, good use and conservation of biodiversity.	3	4	4	4	2	4	4	3	4	4	4	4	2	3	3
4	Machinery capacity.	2	3	4	4	2	3	3	3	4	4	3	4	3	3	2
4	Accessibility to technology.	3	3	4	2	2	3	3	4	4	4	4	4	3	3	4
4	Skilled Labour (Requirement and Availability).	4	3	3	2	4	2	3	4	4	4	4	4	3	3	4
4	Identification of suppliers of raw materials and inputs (national, international origin).	4	4	4	3	2	3	3	2	4	4	4	4	4	3	4
Relevance for performing data analysis tasks.																
4	How many internal or external sources of information exist: databases, Excel sheets, reports.	3	3	4	3	3	3	3	4	3	4	4	4	4	3	4
4	What level of access you have to information.	4	3	3	4	3	2	4	4	3	4	4	4	4	3	4
4	Level of use of computer tools (Words, excel, power point, etc.).	2	3	3	3	3	2	3	4	3	3	4	3	4	3	4
4	Frequency of information collection at this stage of the process.	3	3	3	3	3	3	3	4	3	3	4	4	4	3	4
	Total unweighted.	39	43	43	44	34	37	44	48	48	50	51	49	40	39	48
	Weighted total.	36	36	37	39	37	33	38	39	42	41	38	41	37	36	38

Weight	Evaluation Criteria	Processes														
		Input Suppliers				Producers and/or Growers				Industrial Transformer				Wholesaler / Retailer		
		SCS	AOEI	EPST	AAPF	MOEP	SAAF	APAC	NTAP	TIAT	NTAT	ACDM	NEDN	ACOP	NIVE	DIST
Relevance to Production Management																
4	the factors that intervene in the process are characterized.	4	3	2	4	3	3	4	4	4	4	4	4	3	3	4
4	the uses and functions of the materials and tools used are distinguished	3	4	3	4	2	2	4	4	4	4	4	3	2	3	3
4	information and knowledge management is identified	2	3	3	4	2	3	3	4	4	4	4	4	2	3	4
4	Production, service and support processes are identified.	2	4	3	4	3	4	4	4	4	4	4	3	2	3	4
4	Environmental responsibility, good use and conservation of biodiversity.	3	4	4	4	2	4	4	3	4	4	4	4	2	3	3
4	Machinery capacity	2	3	4	4	2	3	3	3	4	4	3	4	3	3	2
4	Accessibility to technology.	3	3	4	2	2	3	3	4	4	4	4	4	3	3	4
4	Skilled Labor (Requirement and Availability)	4	3	3	2	4	2	3	4	4	4	4	4	3	3	4
4	Identification of suppliers of raw materials and inputs (domestic, international origin)	4	4	4	3	2	3	3	2	4	4	4	4	4	3	4
Relevance for performing data analysis tasks																
4	How many internal or external sources of information exist: databases, Excel sheets, reports, etc.	3	3	4	3	3	3	3	4	3	4	4	4	4	3	4
4	What level of access do you have to the information	4	3	3	4	3	2	4	4	3	4	4	4	4	3	4
4	Level of use of computer tools (Words, excel, power point, etc.).	2	3	3	3	3	2	3	4	3	3	4	3	4	3	4
4	Frequency of information gathering at this stage of the process	3	3	3	3	3	3	3	4	3	3	4	4	4	3	4
	Total unweighted	39	43	43	44	34	37	44	48	48	50	51	49	40	39	48

	Weighted total	3 6	36	37	39	37	33	38	39	42	41	38	41	37	36	38

From the above table, the subprocesses "Type of Input to Transform", "Transforming Technological Level" and "Business Specialization Level" were the highest prioritized.

5. Definition of Autonomous Cycles of Data Analysis

Figure 5 presents the ACODATs of the prioritized subprocesses, to allow an autonomous coordination of the industrial-automation process of MSMEs (ACPCPA-000)

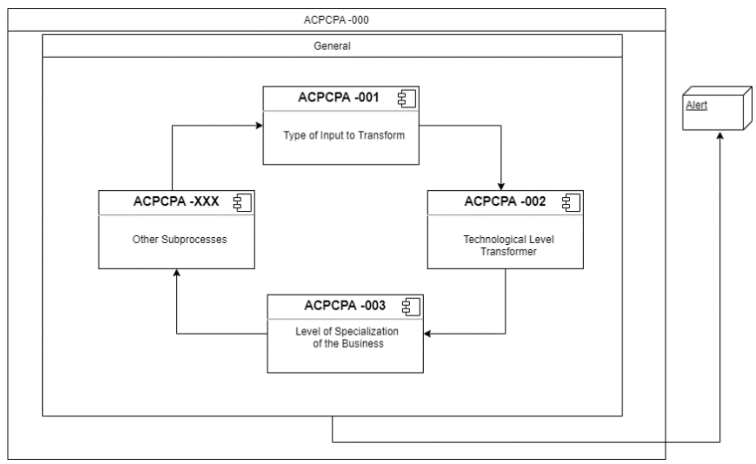

Figure 5. ACPCPA-000: Autonomous cycles of the prioritized processes of MSMEs.

The objective of ACPCPA-000 is the self-management of the industrial-automation of MSMEs within the agroindustrial sector to improve their competitiveness. To achieve this objective, we propose three ACODATs:

ACPCPA-001 (Type of Input to Transform): This cycle is responsible for obtaining useful information to identify the type of input to transform in the agroindustrial production chain, such as quantity, quality, time (seasonality and durability), cost, etc., based on information from the organization and the context.

ACPCPA-002 (Transforming Technological Level): This cycle is responsible for obtaining useful information to identify the technological level required in the transformation process of the agroindustrial production chain, such as identifying the raw material transformation processes and selecting the technology for processing.

ACPCPA-003 (Business-Specialization Level): This cycle is responsible for obtaining useful information to identify the Business-Specialization Level of the production, such as the specific customer sector to cover (target market), the quality of the product, etc.

5.1. Specification of the Autonomous Cycles for the Type of Input to Transform

In this section, we detail ACPCPA-001, which obtained the highest score in the prioritization. This autonomous cycle determines the type of input to transform. Mainly, this cycle is composed of four tasks (see Figure 6). Table 6 shows the general description of each task of this autonomous cycle.

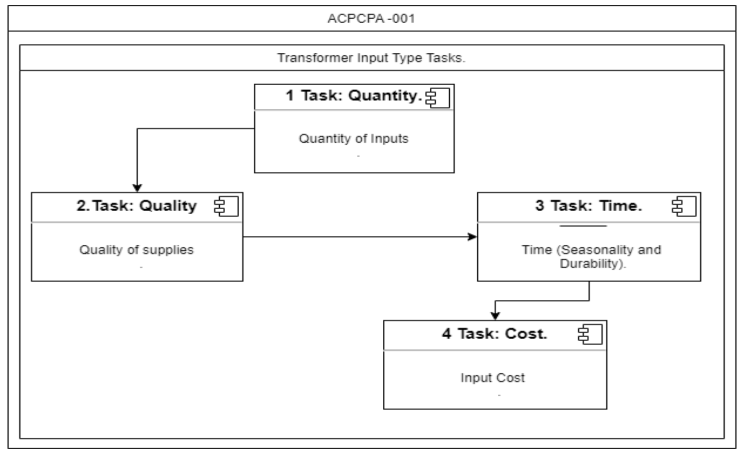

Figure 6. Tasks of ACPCPA-001.

Table 6. Description of ACPCPA-001 tasks.

Task Name	Knowledge Models	Data Sources
1. Determine the amount of input to transform.	Predictive Model	Production demand, customers.
2. Determine the quality of the input to be processed.	Diagnostic Model	Inputs used on the farm, cultural practices, storage, and transport services.
3. Determine the time (Seasonality and Durability) of the input to be transformed.	Predictive Model	Annual demand.
4. Determine the cost of input to be transformed.	Predictive Model	Operating Costs.

Task 1. Quantity of Input to be transformed: This task identifies the raw materials required to produce the product to satisfy customer demands. Its objective is to identify the historical evolution, yields and alternative uses of the product, etc. This task uses a predictive model to determine the amount of input to be transformed.

Task 2. Quality of Input to be processed: This task identifies the quality of the inputs. Its quality affects the quality of the products and depends on the cultural practices and storage and transport services used, among other things. This task uses a diagnostic model.

Task 3. Timing (Seasonality and Durability) of the Input to be processed: This task identifies the factors related to timing. For example, seasonality because many raw materials or inputs are only produced during certain times of the year. Also, the durability of the raw materials and inputs used because some are perishable. For this task, a predictive model can be used.

Task 4. Cost of the Input to transform: This task determines the cost of raw materials or inputs, to determine if it is low enough to contribute to the production profitability.

The first task predicts the quantity of raw material to be transformed, which is the input of the second task to diagnose its quality. The next task uses the result of the previous task to predict its durability, as it can be perishable or seasonal. Finally, the last task predicts the cost of that raw material.

5.2. Specification of the Autonomous Cycles for the Transformer Technology Level

The Autonomous Cycle for the Transforming Technology Level (ACPCPA-002) has as its main objective the characterization of the technology to be used to transform the raw material or input. In general, this autonomous cycle is composed of three tasks (see Figure 7). Table 7 shows the description of each task.

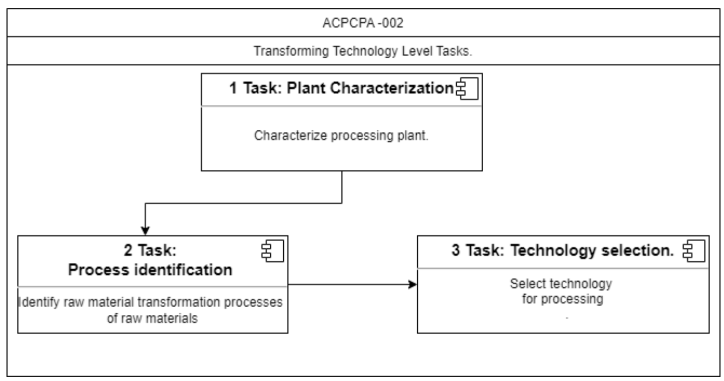

Figure 7. Structure of the ACPCPA -002- Transforming Technology Level.

Table 7. Description of ACPCPA-002 tasks.

Name Task	Knowledge Models	Data Sources
1. Characterize the processing plant.	Classification Model	Location, the exclusivity of the premises, access roads, structure and finishes, lighting, ventilation.
2. Identify raw material transformation processes.	Prescription Model	Phases (collection and production), transportation and storage, sources of supplies, availability of labor, availability of infrastructure, cost of land and raw material, quality standards.
3. Select technology for processing.	Classification or identification Model	Type of production process (made-to-order or job, batch, mass, and continuous flow).

Task 1. Characterize Processing Plant: This task identifies and classifies how the MSMEs work through the identification of essential

elements that allow the management and control of the production, such as location, edification, access roads, structure and finishes, lighting, and ventilation. This task uses classification models.

Task 2. Identify raw-material transformation process: This task identifies the raw material or input transformation process (collection and production). This process is related to distances to input, location of the different processing plants, ease of transportation, availability of labor, availability of infrastructure, and cost of land, among other things. It also manages inventories, due to the seasonal and perishable nature of certain raw materials, storage capacity, inventory financing, for which it can use a prescriptive model.

Task 3. Select technology for processing: This task identifies the technology for processing raw materials or inputs to satisfy the demand. Its objective is to identify the technology dedicated to production. For that, it must select the type of production process, such as on-demand, batch, mass, and continuous flow. This task uses a classification or identification model to select the type of technology for the production processes to be transformed.

The first task identifies the production process to be carried out in the MSME using a classification model; from there, the next task prescribes the transformation process of the raw material. Finally, the last task selects the processing technology for that raw material using a classification or identification model.

5.3. Specification of the Autonomous Cycles for the "Business-Specialization Level"

The Autonomous Cycle for the Business-Specialization Level (ACPCPA-003) has as its main objective the characterization of the final product (target market, quality, and value of the product). This cycle is composed of two tasks (see Figure 8). Table 8 shows the description of each task.

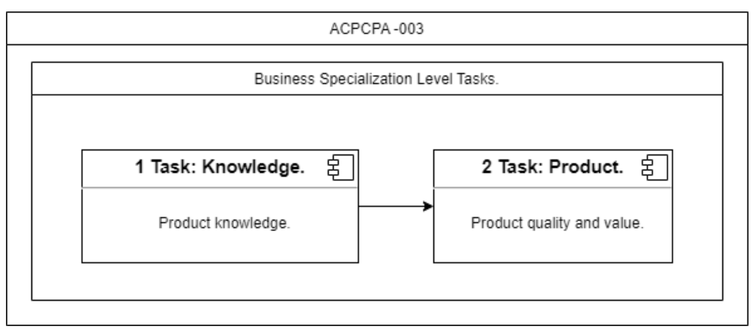

Figure 8. Structure of the ACPCPA-003 Business Specialization Level.

Table 8. Description of ACPCPA-003 tasks.

Name Task	Knowledge Models	Data Sources
1. Identify product knowledge.	Classification or identification Model	Demand needs, access to markets.

2.	Determine the quality and value of the product.	Diagnostic Model	Hygienic-sanitary quality and bromatological quality management in production.

Task 1. Identify product knowledge: This task identifies knowledge of the target market of the product. It defines the needs of demand and is the means to access sophisticated international markets with greater value added. Its objective is to acquire knowledge of the relationship of the product with a target market. This task uses a classification model to identify knowledge of the product.

Task 2. Determine the quality and value of the product: This task identifies the quality and value of the product in terms of hygienic-sanitary quality management in production and quality management in production. This task uses a diagnostic model.

In this ACODAT, the first task identifies the target market of the product using a classification model; and from that identification, the next task determines the quality and value of the product using a diagnostic model.

6. Multidimensional Data Model for the Autonomous Cycles

This section defines the multidimensional data model for the above autonomous cycles. The multidimensional data model is defined in Figure 9.

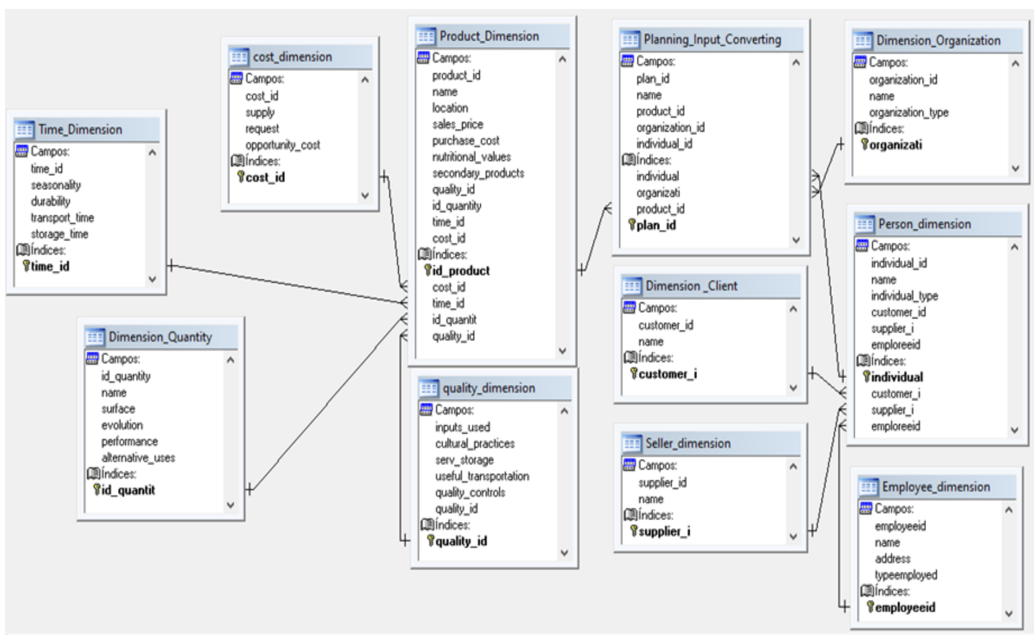

Figure 9. Multidimensional Model.

The model in Figure 9 includes different data sources (e.g., about the raw materials, the production process, the context (market, etc.)). Data are grouped in different dimensions of the data model, depending on their characteristics. The multidimensional data model is a star model that has a fact table (Planning–Input–Converting) and three main dimensions (product, organization, and person). In turn, the product dimension is composed of the dimensions of quantity, quality, time and cost. Finally, the person dimension is made up of the client, seller and employee dimensions. Each dimension is described below. The dimensions are as follows:

Product Dimension: Stores product data (e.g., location, selling price, production cost).

Quantity Dimension: Stores data on the quantity of raw materials required to satisfy demand, for example, historical evolution, yield, and alternative uses.

Quality Dimension: Stores data on the quality of raw materials or products, and the information related; for example, inputs used on the farm, cultural practices, storage and transportation services, and quality controls established.

Time Dimension: Stores time data of raw materials or inputs identifying the various factors related to time, for example, seasonality, durability, and storage time.

Cost Dimension: Stores data on the cost of raw materials or inputs; for example, supply and demands, opportunity cost, logistic services, government interventions, alliances with producers and contracting standards in the area.

Organization Dimension: Stores company or organization data; for example, name, address, and type of organization such as producers, suppliers, processors, transporters, warehousing, financial, marketing and distributors.

Person Dimension: Stores individual data; address, phone, and email.

Client Dimension: Stores client data (e.g., type and frequency of demand).

Seller Dimension: Stores seller data (e.g., type of product to sell).

Employee Dimension: Stores Employee data; for example, the type of employees such as producers, coordinators, or operators.

Thus, the Planning–Input–Converting is the fact table of the multidimensional model, which stores the information generated by the ACODATs (the measures that are being analyzed) and the keys to join each dimension table.

7. Case Study of Café Galavis

For this case study, this section presents the experimental context and the instantiation of ACPCPA-001 (Type of input to be processed).

7.1. Experimental Context

The raw material to be processed is a key element in the agroindustrial sector. The processing must deliver to the market a product that has an acceptable quality, in an adequate and sufficient quantity for the chosen market, in a time that allows covering the needs of the demand, and at a reasonable and competitive price.

In this case study, we use data from the company "Café Galavis", located in the industrial zone of the city of Cúcuta, Colombia. This operation is dedicated to the roasting and distribution of coffee.

To illustrate the functionality of ACPCPA-000, this case study analyzes ACPCPA-001 according to the following scenario. The coffee beans enter the factory in a process called "reception of the beans" with 70 kg bags and are stored in optimal conditions complying with the quality standards of the market to which the product is directed (ambient temperature between 20 °C and 25 °C and humidity between 10% and 12%). The coffee is stored in the warehouse, in green beans, and classified according to various criteria, mainly size and density. The beans are measured by passing them through a sieve to classify the size, a process called "storage and weighing", which is a large and adequate place to protect the raw material from the sun, wind, water, animals, odors, dust and dirt. Once the coffee has been processed and classified into qualities, it is packaged into 60 kg sacks, which is an international standard.

Depending on the orders or demand, the quantity for the production order of the day is then weighed. Once the above is fulfilled, the beans are dehydrated until they reach the required point, through the application of heat, which causes several physical changes and chemical reactions that develop the aroma and flavor. During the roasting process, the temperatures that the coffee beans reach are around 193 °C for a light roast, close to 200 °C for a medium roast, and close to 218 °C for a dark roast. This allows processing 280 kg in 20 min, that is to say, a production capacity of 60,000 pounds per month.

The coffee then goes through the cooling process that consists of activating the system for a period of approximately 5 min; through an air-suction system, coffee is elevated to three meters high, producing the first cleaning of the pure grain. Next is the milling process, which is carried out through two types of mills, in which one part is processed in 75% in the granulating mills and the other in the fractional mills in 25%. The milled grain is then mechanically conveyed through special ducts. A second cleaning of the coffee is carried out by means of a vibration system, separating it from the remaining particles.

The milled grain enters the hoppers of the packing machines dosing the product according to its presentation, according to its weight previously selected, and programmed by portions of 2500, 500, 250, 125, 50 and 25 g. Thus the process is finished in perfect conditions.

The finished product is placed on plastic pallets where the process of control, coding and sealing of the packaging is carried out. Once the coffee is in inventory, it is taken by carts to the finished product warehouse. It is then delivered to the team of distributors who are in charge of delivering it to the final consumer.

7.2. Instantiation of ACPCPA-001 (Type of Input to Transform)

The instantiation of ACPCPA-001 should consider, for example, the quantity, quality, time, and cost of coffee beans used in production. The following steps describe how ACPCPA-001 has been instantiated for this case study.

Task 1. Input quantity task: The first task consists of automatically determining the quantity of coffee beans, for which a prediction model is used. The prediction model is built with historical data found in the company's software. The prediction model uses variables such as surface area, evolution, yield, ambient temperature, and humidity. An example of the results of the prediction model is shown in Table 9. Two cases of this task are described below.

Case 1: For storage of 110 bags of green coffee beans, the model estimates that it should have a temperature of 25 °C and humidity of 14%. The recommendation is that green beans be stored between 20 °C and 25 °C. In this case, Task 2 can be carried out because storage is maintained at the desired levels.

Case 2: For 150 bags stored, the model estimates the ambient temperature of 18 °C and humidity of 15%. As the recommendation is a temperature between 20 °C and 25 °C and humidity between 12 and 14%. Given that the temperature and humidity conditions for storage are not favorable, in this case, a warning is generated to the warehouseman to take the appropriate decisions.

Table 9. Predictions generated by the first task.

Winery	Week	Quantity (Bags)	Ambient Temperature °C	Humidity %
Bod_01	Week 1	100	20	12
Bod_01	Week 2	120	22	13
Bod_01	Week 3	150	18	15
Bod_01	Week 4	110	25	14

The first task predicts the amount of coffee to be transformed according to the contextual conditions. For the first case, it is expected that 110 bags of green coffee beans can be processed/transformed, and for the second case, 150 bags can be processed.

Task 2. Input-quality task: The second task uses a diagnostic model based on historical data from raw-material sale websites. The diagnostic model uses variables such as inputs used, acidity, aroma, cultural practices, storage, and transportation services used and quality controls. An example of the results of the diagnostic model is shown in Table 10. Two cases of this task are described below.

Case 1: It is recommended that green beans have a pH of acidity between 4.9 and 5.2. This is an important factor in determining the quality of coffee in terms of flavor. When the coffee has a pH lower than 4.9, it acquires a flavor that is too acidic and above 5.2 it is bitterer. Another factor is the size and density of the bean. For a coffee to be a specialty coffee, it must have zero defects (category 1) and a maximum of five defects is category 2. Table 9 shows that the model diagnoses a storage of 120 bags of green coffee beans, with an acidity of 5.0 and a category of 2. In this case, Task 3 would be carried out since the quality of the beans is maintained at the desired levels.

Case 2: For 150 bags, the model diagnoses an acidity of 5.5 with a category of 6, as presented in Table 9 (see Week 3). Given that the conditions for storage (the inputs used: coffee body and aroma) are not favorable, in this case, a warning is generated to the warehouse personnel so they can make the appropriate decisions.

Table 10. Diagnostics generated by the second task.

Winery	Week	Quantity (Bags)	Acidity (4.9–5.2)	Category (0–5)
Bod_01	Week 1	100	4.9	0
Bod_01	Week 2	120	5.0	2
Bod_01	Week 3	150	5.5	6
Bod_01	Week 4	110	5.2	4

Thus, this task diagnoses the quality characteristics of the coffee determined to be processed in the first task. For the first case (120 bags of green coffee beans), it diagnoses an acidity of 5.0 and a category 2, and for the second case (150 bags), the model diagnoses an acidity of 5.5 and a category 6.

Task 3. Input-Time Task: In the third task, the system will use a predictive model to identify the various time-related factors such as seasonality and durability of the input. The predictive model uses variables such as surface area, evolution, yield, roasting temperature, and time. An example of the results of the prediction model is shown in Table 11. Two cases of this task are described below.

Case 1: It is recommended that the roasting temperature in the industrial machines start with the oven preheated to 200 °C. After loading the coffee, the temperature drops to half, and then, gradually rises again at a rate of 10 °C per minute. If it rises faster, the coffee beans are roasted externally but remain raw and hollow on the inside. The model estimates that the roasting of 120 bags of green coffee beans has a temperature of 200 °C and a time of 13 min. In this case, Task 4 can be carried out, as the roasting and cooling process is maintained at the desired levels.

Case 2: For 110 bags, the model estimates a roasting temperature of 300 °C and a time of 20 min, as presented in Table 11 (see Week 4). Since the temperature conditions in the roasting machine are not favorable, in this case, a warning is generated to the roasting machine operator to make the appropriate decisions.

Table 11. Predictions generated by the third task.

Toaster	Week	Quantity (Bags)	Temperature (°C)	Time (Minutes)
Tost_01	Week 1	100	193	12
Tost_01	Week 2	120	200	13
Tost_01	Week 3	150	218	14
Tost_01	Week 4	110	300	20

Specifically, this task estimates the durability characteristics of the production process of the coffee. For the first case (120 bags of green coffee beans), the model estimates that roasting must have a temperature of 200 °C and a time of 13 min, and for the second case, it estimates a roasting temperature of 300 °C and a time of 20 min.

Task 4. Input Cost Task: In the fourth task, the system will use a predictive model to estimate the cost of raw materials or inputs. These costs should be low enough to contribute to the processing plant's profitability. This model uses variables such as supplies, demand, and quality and time of inputs. Two cases of this task are presented below.

Case 1: If the tasks of quantity, quality, and time of inputs to be transformed are kept at the desired levels, it could be predicted that costs will remain stable in production.

Case 2: If one or all of the tasks of quantity, quality, and time of inputs to be transformed are not maintained at the desired levels, it can be predicted that costs will increase in production. In this case, a warning is generated to the operators or managers to make the appropriate decisions.

Finally, the last task predicts the cost of the coffee. For the first case, due to the quantity and quality of the coffee, it predicts a stable cost in production. In the second case, because the quantity and quality are not maintained at the desired levels, it predicts that costs will increase in production.

8. General Discussion

8.1. Comparison with Previous Works

In this section, we propose several criteria to analyze the automation of the production chain of MSMEs for the agroindustrial sector to improve their competitiveness. This is followed by a qualitative comparison of this work with related works, based on the above criteria (see Table 12) [35].

Criterion 1: Automation of the entire industrial production chain of MSMEs.

Criterion 2: Use of data mining techniques in the industrial automation of the production chain.

Criterion 3: Quantity, quality, time, and cost are jointly analyzed in the industrial-automation process.

Criterion 4: Consider efficient and environmentally friendly production.

Table 12. Comparison with previous work.

	Criterion 1	Criterion 2	Criterion 3	Criterion 4
[12]	X	√	√	X
[17]	X	√	√	X
[18]	X	X	X	X
[19]	X	X	X	X
[32]	X	X	X	X
[36]	X	√	√	X
This work	√	√	√	√

Table shows whether the analyzed works meet the criteria indicated above, indicating √ that it does and X that it does not. As shown in Table 12, the related articles did not meet all the criteria. Specifically, in criterion 1, our research enables, through autonomic cycles, the automation of the entire production process. For this automation, it is necessary to use paradigms such as multi-agent systems together with ACODAT architecture to model the whole production process [7,30].

For criterion 2, García et al. [17] analyzed a case study of multirelational data mining, using the Connection-Block algorithm, applied to the database in this agroindustrial sugar sector. The basis of our proposal is data-driven autonomous decision making, with knowledge extracted from the industrial transformation of the production chain.

For criterion 3, Roukh et al. [36] focused on the acquisition, processing, and visualization of massive amounts of data, both batch and real time. Sen et al. [12] and Garcia et al. [17] focused on converting data useful for decision making, especially for precision agriculture and agribusiness, to estimate land-cover change. Meyer et al. [18] showed how consumption management problems can be solved by the widespread installation of sensors on production lines. Bader et al. [19] stated that the adoption of industrial robots, in food processing, has been slow. This proposal considers the quantity, quality, time, and cost of the inputs to be transformed, in the industrial automation of the production chain of MSMEs.

Finally, regarding criterion 4, this proposal meets the criterion of efficient and environmentally friendly production because the operator (warehouseman, processor) and manager can know the quantity, quality, time, and cost of the raw material. This can be used to reduce greenhouse gases, manage the hygienic-sanitary quality in production, make agroindustrial production more sustainable, etc.

8.2. Quality of the Knowledge Models

In this section, we carry out a quantitative analysis of the behavior of the knowledge models used in each task (see Table 13). To do this, we have used R^2 and MAPE (Mean Absolute Percentage Error) as metrics for the predictive models, and, in the case of the diagnostic model, the silhouette index. R^2, MAPE, and the silhouette index are self-contained metrics, so a value close to 1 indicates a very good quality of the models. In addition, we have used machine learning techniques to build the models to RF in the case of the prediction models, and K-means in the case of the diagnostic model. We see that the results obtained are very good in general. For example, in the first task, it gives us a fairly good R^2 and MAPE (95 and 89%, respectively). Likewise, the quality of the diagnostic model is very good, with a Silhouette index value of 0.87.

Table 13. Quality of the Knowledge Models for each task.

Task Number	Knowledge Models	Quality Metrics
1	Predictive Model	$R^2 = 0.95$ MAPE = 89%
2	Diagnostic Model	Silhouette index = 0.87
3	Predictive Model	$R^2 = 0.92$ MAPE = 88%
4	Predictive Model	$R^2 = 0.97$ MAPE = 95%

9. Conclusions and Direction of Future Work

This paper presents an architecture for the industrial automation of the production chains of MSMEs. The architecture combines multiple variables (e.g., temperature, humidity, acidity, quantity, quality, and time) and allows the integration and interoperability of actors in the context of production. The proposed architecture is based on the ACODAT concept and proposes several autonomous cycles to give autonomy to the industrial production chains of MSMEs. Furthermore, this article shows the instantiation of the autonomous cycle for the characterization of the inputs to be transformed as the first step toward an efficient and effective industrial automation process of the production chain.

At the level of industrial automation, our main contribution is the specification of the three main ACODATs to manage the most relevant subprocesses for the agroindustrial automation of the production chains of MSMEs. The main benefit of our approach is that it allows an industrial automation process to be carried out based on the data of the organization and its environment, which allows MSMEs to exploit their data without increasing their operational costs. Observing the case study in a coffee factory, our first ACODAT determines the characteristics of the coffee to be processed (both in terms of quantity, quality, and the characteristics of its production process), to later estimate the cost of production, according to the target market aspired to reach. Thus, we see, in this specific case, the comprehensive behavior of ACODAT to carry out an exhaustive analysis of coffee production using its organizational and context data.

Other results of this research are (i) the use of environmental variables to make decisions on optimal industrial automation, (ii) the ability to use mining techniques to improve system knowledge and decision-making processes (e.g., data mining, text mining and web mining techniques), and (iii) real-time analysis for production in the industrial automation process. In particular, the case study shows that extraction techniques of knowledge are necessary to address self-management in industrial automation. This case study serves as a guide for incorporating self-management in industrial automation of the production chain of the coffee industry, based on the paradigm of autonomic computing and data mining techniques.

The main limitations of this study are the following. The first is that only one of the autonomous cycles was instantiated in a very specific case study, so its scalability in other contexts must be evaluated. The second major limitation is that the integration of the different autonomous cycles that automate the different processes present in the automation of the production chain has not been evaluated. A third limitation is the lack of evaluation of the operational costs involved in maintaining the data repository and models based on data in real-time.

Future work is aimed at implementing this framework in a real-time context to verify the functionalities of this solution. In this sense, we plan to use historical data of the company "Café Galavis" to develop the different data analysis tasks (e.g., to determine the quality of the coffee bean). Finally, one of the biggest challenges to implement this type of system is the cost associated with the sensors for data acquisition. In addition, other challenges are the optimal distribution of these sensors, and the Internet connectivity in the industrial installation. Thus, future studies should analyze these aspects and how to consider them in the proposed architecture.

Author Contributions: Conceptualization, J.F. and J.A.; methodology, J.F. and J.A.; formal analysis, J.F., J.A. and E.M.; resources, Á.P.; data curation, J.F.; writing—original draft preparation J.F. and J.A.; writing—review and editing, E.M. and Á.P.; supervision, J.A.; funding acquisition, Á.P. All authors have read and agreed to the published version of the manuscript.

Funding: Jairo Fuentes holds a Bicentenary PhD scholarship funded by the Ministry of Science, Technology, and Innovation (Minciencias). All authors would like to thank the Vice Rector's Office for Discovery and Creation, Universidad EAFIT, for their support in this research. The authors would like to thank the HAMADI 4.0 project, code 22-STIC-06, of the STIC-AmSud regional program, for supporting this work.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data will be available upon request.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Solleiro, J.; Del Valle, M. *El cambio Tecnológico en la Agricultura y las Agroindustrias en México*, Siglo, Ed.; México, 1996; p. xxi.
2. Solleiro-Rebolledo, J.L.; García-Martínez, M.B.; Castañón-Ibarra, R.; Martínez-Salvador, L.E. Smart specialization for building up a regional innovation agenda: the case of San Luis Potosí, Mexico. *J. Evol. Stud. Business-JESB* **2020**, *5*, 81–115. https://doi.org/10.1344/jesb2020.1.j069.
3. Sánchez, M.; Aguilar, J.; Cordero, J.; Valdiviezo-Díaz, P.; Barba-Guamán, L.; Chamba-Eras, L. Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service. In *New Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; Volume 444, pp. 993–1002.
4. Aguilar, J.; Garces-Jimenez, A.; Gallego-Salvador, N.; De Mesa, J.A.G.; Gomez-Pulido, J.M.; Garcia-Tejedor, A.J. Autonomic Management Architecture for Multi-HVAC Systems in Smart Buildings. *IEEE Access* **2019**, *7*, 123402–123415.

5. Candia, G. Industry 4.0 and its aberrations. *İnformasiya Cəmiyyəti Probl.* **2022**, *1*, 48–57.
6. Eisavi, V.; Homayouni, S.; Yazdi, A.M.; Alimohammadi, A. Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environ. Monit. Assess.* **2015**, *187*, 291.
7. Sanchez, M.; Exposito, E.; Aguilar, J. Implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context. *Comput. Ind.* **2020**, *121*, 103247.
8. Valencia-Cárdenas, M.; Restrepo-Morales, J.A.; Día-Serna, F.J. Big Data Analytics in the Agribusiness Supply Chain Management. *AiBi Revista de Investigación, Administración e Ingeniería*, **2021**, *9*, 32–42.
9. Flórez, D. Prospective research guidelines for the production chain of sugarcane—(focus on panela, not centrifuged sugar). *Tecnura* **2013**, *17*, 72-86.
10. Chaves, J.; Díaz, R.; Hernández, A.; Hidalgo, O. Cadenas productivas agroindustriales y competitividad: definición de políticas y estrategias en el meso nivel. *Econ. Y Soc.* **2000**, *13*, 5–18.
11. Isaza, J. Cadenas productivas. Enfoques y precisiones conceptuales. *Sotavento* **2008**, *11*, 8–25.
12. Sen, D.; Ozturk, M.; Vayvay, O. An Overview of Big Data for Growth in SMEs. *Procedia-Soc. Behav. Sci.* **2016**, *235*, 159–167. https://doi.org/10.1016/j.sbspro.2016.11.011.
13. Li, Y.; Jiang, W.; Yang, L.; Wu, T. On neural networks and learning systems for business computing. *Neurocomputing* **2018**, *275*, 1150–1159
14. Marinagi, C.; Skourlas, C.; Galiotou, E. Advanced information technology solutions for implementing information sharing across supply chains. In *ACM International Conference Proceeding Series, Proceedings of the PCI '18: Proceedings of the 22nd Pan-Hellenic Conference on Informatics, Athens, Greece, 29 November–1 December 2018*; ACM: New York, NY, USA, 2018; pp. 99–102. https://doi.org/10.1145/3291533.3291575
15. Lopez, H.A.G.; Cisneros, M.A.P. Industry 4.0 & Internet of Things in Supply Chain. In Proceedings of the CLIHC '17: 8th Latin American Conference on Human-Computer Interaction, Antigua Guatemala, Guatemala, 8–10 November 2017; pp. 1–4. https://doi.org/10.1145/3151470.3156646.
16. Luque, A.; Peralta, M.E.; Heras, A.d.L.; Córdoba, A. State of the Industry 4.0 in the Andalusian food sector. *Procedia Manuf.* **2017**, *13*, 1199–1205. https://doi.org/10.1016/j.promfg.2017.09.195.
17. García, E.; Vieira, M. (2008). Estudo de caso de mineração de dados multirelacional: aplicação do algoritmo connetionblock em um problema da agroindústria. In Proceedings of the Simpósio Brasileiro de Bancos de Dados, Campinas, Brazil, 13–15 October 2008; pp. 224–237.
18. Meyer, M.; Dykes, J. Criteria for Rigor in Visualization Design Study. *IEEE Trans. Vis. Comput. Graph.* **2019**, *26*, 87–97. https://doi.org/10.1109/TVCG.2019.2934539.
19. Bader, F.; Rahimifard, S. Challenges for Industrial Robot Applications in Food Manufacturing. In Proceedings of the ISCSIC '18: The 2nd International Symposium on Computer Science and Intelligent Control, Stockholm, Sweden, 21–23 September 2018.
20. Kakhki, F.D.; Freeman, S.A.; Mosher, G. Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Saf. Sci.* **2019**, *117*, 257–262. https://doi.org/10.1016/j.ssci.2019.04.026.
21. Borghesan, F.; Zagorowska, M.; Mercangöz, M. Unmanned and Autonomous Systems: Future of Automation in Process and Energy Industries. *IFAC-Pap.* **2022**, *55*, 875–882.
22. Uygun, Y. Autonomous Manufacturing-Related Procurement in the Era of Industry 4.0. In *Digitalisierung im Einkauf*, Schupp, F., Wöhner, H., Eds.; Springer: Gabler, Wiesbaden, 2023.
23. Kephart, J.; Chess, D. The vision of autonomic computing. *Computer* **2003**, *36*, 41–52
24. Papetti, A.; Gregori, F.; Pandolfi, M.; Peruzzini, M.; Germani, M. Iot to enable social sustainability in manufacturing systems. *Adv. Transdiscipl. Eng.* **2018**, *7*, 53–62.
25. Aguilar, J.; Jerez, M.; Exposito, E.; Villemur, T. CARMiCLOC: Context Awareness Middleware in Cloud Computing. In Proceedings of the 2015 XLI Latin American Computing Conference (CLEI), Arequipa, Peru, 19–23 October 2015.
26. Morales, L.; Ouedraogo, C.A.; Aguilar, J.; Chassot, C.; Medjiah, S.; Drira, K. Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform. *Serv. Oriented Comput. Appl.* **2019**, *13*, 199–219.
27. Verdouw, C.; Sundmaeker, H.; Tekinerdogan, B.; Conzon, D.; Montanaro, T. Architecture framework of IoT-based food and farm systems: A multiple case study. *Comput. Electron. Agric.* **2019**, *165*, 104939.
28. Yadav, S.; Luthra, S.; Garg, D. Modelling Internet of things (IoT)-driven global sustainability in multi-tier agri-food supply chain under natural epidemic outbreaks. *Environ. Sci. Pollut. Res.* **2021**, *28*, 16633–16654.
29. Ramírez-Valverde, B. "Gerardo Torres Salcido y Rosa María Larroa Torres (coord): Sistemas agroalimentarios localizados: desarrollo conceptual y diversidad de situaciones" (Reseña), *agric. soc. desarro.* **2013**, 10, 133-137.
30. Nonaka, I. The knowledge creating company. *Harv. Bus. Rev.* **1991**, *85*, 162–171.

31. Castellanos, O.; Rojas, J. Conceptualización y papel de la cadena productiva en un entorno de competitividad. *Innovar* **2001**, *18;* 87-98.
32. Fletes, H.; Ocampo, G.; Valdiviezo, G. Agroindustry dynamism in the Corredor Costero, Chiapas, Mexico. Coordination and territorial competitivity. *Mundo Agrar.* **2016**, *17*, e038.
33. Organización de Cooperación y Desarrollo Económicos, ocde. *Manual de Oslo: Guía Para la Recogida e Interpretación de Datos Sobre Innovación*, 3rd ed.; Traducción española Grupo Tragsa: Madrid, España, 2005; p. 188.
34. Salimbeni, S.; Redchuk, A.; Rousserie, H. Quality 4.0: technologies and readiness factors in the entire value flow life cycle. *Prod. Manuf. Res.* **2023**, *11*, 2238797.
35. Bell, M.; Pavitt, K. The development of technological capabilities. In *Trade, Technology, and International Competitiveness*; Haque, I., Ed.; Economic Development Institute, The World Bank: Washington, DC, USA, 1995; pp. 69-101.
36. Roukh, A.; Fote, F., Mahmoudi, S. Mahmoudi, S. WALLeSMART: Cloud Platform for Smart Farming. In *ACM International Conference Proceeding Series, SSDBM '20: Proceedings of the 32nd International Conference on Scientific and Statistical Database Management*, Vienna, Austria, 7–9 July 2020; pp. 1–4. https://doi.org/10.1145/3400903.3401690.

Appendix C. Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs

Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs

Jairo Fuentes[1,2], Jose Aguilar[2,3,4, *], Edwin Montoya[2]

[1] Universidad Francisco de Paula Santander, Av. Gran Colombia, Cúcuta, Colombia
[2] GIDITIC, Universidad EAFIT, Carrera 48 No. 7 Sur-50, Medellín, Colombia
[3] CEMISID, Facultad de Ingeniera, Universidad de Los Andes, Mérida, Venezuela
[4] IMDEA Networks Institute, Leganés, Madrid, Spain
*Corresponding Author (aguilarjos@gmail.com; aguilar@ula.ve)

Abstract

Coffee production needs certain levels of efficiency to ensure that the quality of the bean, the roasting process, and in general, the coffee processing methods, achieve financial and environmental sustainability objectives. This requires tasks of monitoring and analyzing of features of the coffee bean, and the roasting process, among other aspects, so that stakeholders of the agro-industrial sector of MSMEs can know what happens in the coffee production and can make better decisions to improve it. In a previous article, three autonomous cycles of data analysis tasks are proposed for the automation of the production chains of the MSMEs. This work aims to instantiate the autonomous cycle responsible for identifying the type of input to transform in the production process, in the case of coffee production. This cycle analyzes the inputs of the production chain (quantity, quality, seasonality, durability, cost, etc.), based on information from the organization and the context, to establish the production process to be carried out. This autonomous cycle is instanced in coffee production to identify the type of input to transform (bean quality), and to determine the transformation process (level of decrease of the bean during the roasting process and coffee processing method). The quality model is defined by the K-means technique with a performance in the Silhouette Index of 0.85, the predictive model of the level of decrease of beans in the roasting process is defined by Random Forest with a performance in the accuracy of 0.81, and finally, the identification model of the "production method" is carried out by the Logistic Regression technique with a quality performance in the accuracy of 0.72.

Keywords: Autonomous Computing, Artificial Intelligence, Data Analysis, Machine Learning, Production-chain, coffee agro-industry.

1. Introduction

This work is related to improving production processes in the MYPIMES of the coffee sector to obtain better participation in the market. Particularly, the automation of the production processes of MYPIMES is very important to make them financially and environmentally sustainable (Rocha et al., 2016). The issue is important since the future of these companies can depend on it. Especially, a greater understanding of automation in production processes is required in the context of MYPIMES.

Specifically, coffee processing methods are important to achieve quality organoleptic characteristics. Some preliminary works have tried to focus on aspects related to this. The objective of the work (Rocha et al., 2016) was to evaluate the effect of drying using an electromechanical infrared-based system on the cup sensory quality of specialty coffees to assurance the coffee quality. The authors of (Sánchez-Aguiar et al., 2019) propose a method to automatically discriminate good beans from immature ones based on digital image processing techniques. However, in general, coffee producers have numerous difficulties in detecting faults in their production systems, which directly affects the quality and production methods of coffee. The problem lies in the inability to make real-time, data-driven decisions about what may be happening in the process of grading coffee quality and production methods.

In this context, we propose an approach for the automation of the production processes of MYPIMES in the coffee sector. In this sense, this work proposes an approach that allows automating the process of analyzing the quality of the coffee bean, to determine the processing method, based on the concept of autonomous cycles of data analysis tasks. Specifically, an autonomic cycle is implemented to ensure that coffee bean quality and processing methods meet financial and environmental sustainability targets, with the aim of helping coffee producers make decisions to increase production efficiency. Autonomic computing is a paradigm in which the computer system has the self-managing capacity to adapt to its environment (Horn,2001), (Vizcarrondo et al., 2017). The concept of autonomous cycles of data analysis tasks (ACODAT) has been defined by (Chamba-Era et al. 2017), (Sánchez et al. 2016) as a set of data analysis tasks that act autonomously to monitor, analyze, and control a process, which has been applied in different contexts (Aguilar et al., 2022), (Aguilar et al. 2019), (Morales et al., 2019). These data analysis tasks are based on knowledge models (prescriptive, predictive, diagnosis, etc.), and interact with each other according to the cycle objectives. Each data analysis task has a different function: to observe the process, to analyze and interpret what happens in the process, and to make decisions to improve the process. The main contributions of this paper are:

- It integrates various data analysis tasks, among which is the analysis of the coffee beans' quality, the estimation of the decrease of the coffee beans during the roasting process, and the determination of the coffee processing method to be used.
- It uses an ACODAT to automate the production process, which mixes knowledge models based on supervised and unsupervised algorithms to make decisions to improve the production process.

This article is organized as follows. Section 2 presents related work. Then, section 3 presents the ACODAT used in this paper. Section 4 shows the instantiation of this ACODAT in coffee production. Then, section 5 describes and analyzes the results obtained. Finally, section 6 shows a comparison with previous work, and the paper ends with the conclusions and future works.

2. Related work

The goal of the work (Eshetu et al., 2022), (Corrales et al., 2014) was to evaluate the effect of drying on the quality of specialty coffees. For this purpose, the authors designed an infrared-based electromechanical system that combines electromagnetic radiation with conventional convection heating. They concluded that the electromechanical system with infrared drying increased the quality of the coffee compared to traditional drying. The objective of the study (Guevara-Sánchez et al., 2019) was to evaluate the effect of two types of drying (mechanical and traditional), at five altitudes, on coffee quality. They determine that there are significant differences such that the mechanical drying was superior. Thus, in the case of the organoleptic quality of the coffee, mechanical drying was found to be the better option.

In addition, (Puerta-Quintero, 2016) showed that the cup quality of coffee is influenced by moisture and the type and number of physical defects in the bean. Coffee quality deteriorates even if defective beans are removed before roasting. Therefore, it is necessary to perform good sorting practices from the farm to the processing stages to improve the coffee quality. On the other hand, in (Firdissa et al., 2022) has been estimated that coffee at a temperature of less than 20°C and relative humidity of less than 70% is preserved for longer and does not show significant wetting. The study carried out allows establishing the acidity and aroma, registering a significant statistical variation over time, although the quality characteristics are preserved satisfactorily during 6 months of storage of the coffee in refrigeration.

In (Campo et al., 2020), the authors defined a hardware for the detection of fragrance and aroma thresholds at three levels of light, medium and dark roast, suitable for coffee quality evaluation. The hardware is based on two chambers, one for the MOx gas sensor array, and one for the concentrated coffee sample. The authors of (Sánchez-Aguiar et al., 2019) proposed a method to automatically discriminate good beans from immature ones based on digital image processing techniques. They developed a prototype to capture images under controlled lighting conditions.

In (Tamayo-Monsalve et al., 2022), the authors presented a framework for coffee maturity classification from multispectral image data based on Convolutional Neural Networks (CNNs). The system uses multispectral image acquisition systems that generate large amounts of data and takes advantage of the ability of CNNs to extract meaningful patterns from very high-dimensional data to provide a feasible method for grading coffee fruits. According to the authors (Anita & Albarda, 2020), the quality of coffee is determined by 60% when it is planted, 30% when it is roasted, and 10% when it is brewed. They proposed a coffee bean selection process based on the next classification: ripe, undercooked, raw, and damaged, using the GLCM (Gray-Level Co-Occurrence Matrix) method for feature extraction, and KNN (k-Nearest Neighbor) and CNN for the classification model. They obtained an accuracy of 72.12%.

In (Montalbo & Hernandez, 2020), authors proposed a deep learning method to identify diseases such as rust, spots and insect infestation that cause many losses to farmers due to inadequate diagnosis and treatment. They use a CNN to perform classifying diseases. The work (García et al., 2019) presented an image processing and machine-learning technique integrated with an Arduino Mega board to evaluate four important factors in selecting the best quality green coffee beans. They considered the next factors: Colour, morphology, shape, and size, to identify the best quality beans.

In general, there is great interest in proposing mechanisms to improve the coffee production process to guarantee its quality (Horta De Oliveira et al., 2018),(Lince Salazar & Sadeghian Khalajabadi, 2015). Some of the papers have proposed hardware, and other papers have proposed data-driven models to analyze coffee bean quality. Our work has a more holistic vision because it proposes to integrate several data analysis tasks, among which is analyzing the quality of the coffee beans, estimating the decrease of the coffee bean in the roasting process, and determining the coffee processing method to be used.

3. ACODAT for the determination of the Production Process.

This autonomous cycle determines the type of input to be transformed, and according to it, the production process to be deployed. Mainly, this cycle is composed of three tasks (see Figure. 1). Table 1 shows the general description of each task of this autonomous cycle.

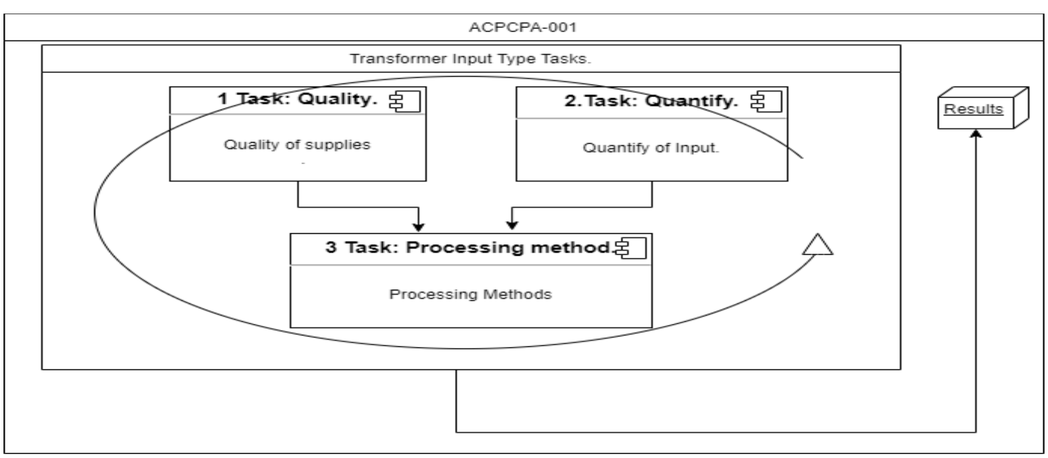

Figure 1. ACPCPA - 001 tasks.

Table 1. Description of ACPCPA - 001 tasks.

Name of the task	Knowledge models	Data sources
1. Determine the quantity of input to be transformed.	Predictive model.	Production demand, customers, temperature, and humidity of the environment.
2. Determine the quality of the input to be processed.	Diagnostic model.	Inputs used on farm, cultural practices, storage and transport services, demand, size, density, acidity, Defect category.
3. Determine the method of processing.	Predictive model.	aroma, flavour, acidity, body, uniformity, sweetness, and moisture.

Task 1. Quantity of input to transform: This task identifies the raw materials required to meet customer demands. It analyzes different aspects as the historical evolution, uses of the product, etc. This task uses a predictive model to identify the product and the quantity to transform.

Task 2. Quality of inputs to process: This task determines the quality of the inputs, which depends on cultural practices, storage and transport services used, among other things. This task uses a diagnostic model to determine this quality.

Task 3. Method of processing to use: This task identifies factors related to the processing method. For example, natural washed, natural, or mechanical drying, automatic or manual selection, among others. A predictive model is used for this task.

4. Case Study in the Coffee Agro-industrial Sector of MSMEs

This cycle is responsible for obtaining useful information to determine the features of the agro-industrial production chain, such as the roasting process and the method of coffee processing. This will allow making decisions to reduce production costs and improve the company's profitability.

The cycle was instanced in the Café Galavis Company located at Cucuta, Colombia. This company, founded in 1918, is dedicated to the roasting and distribution of Colombian Coffee. They have sophisticated equipments for the process that takes place there, from the time the raw material is received at the company's facilities to its storage in optimal temperature and humidity conditions. In addition, they include the most rigorous selection of the bean, its purification prior to processing, and the most hygienic conditions for its internal transport to the packaging machines of the company.

In this case of study, the autonomous cycle proposed in the previous section has some adjustments, since task 1 is not necessary, and in task 3 two subtasks are carried out, to analyze the roasting process and to determine the coffee processing method to use. Task 1 is not necessary because it is a company that receives the raw material (coffee beans) from a predefined network of producers (it is not necessary to estimate how much it will receive), and task 3 analyzes those two specific aspects (roasting process and coffee processing method) because, as identified in the literature review, they are the two elements that affect the quality of coffee [7].

4.1. Task of determination of the coffee quality.

4.1.1. Feature engineering process

The dataset used is from experts' scoring of different coffee preparations. Thus, a score value is assigned to each cup of coffee tasted, where each sample has specific information on the coffee, its origin, the recipe used, characteristics of the raw material and the final product. The dataset has a total of 1,347 records and 44 variables, from which the quality of the coffee will be estimated such that the categorization will be defined using unsupervised learning algorithms (Kim, 2022).

Initially, it starts with a feature engineering process to eliminate the features/variables with many missing or outlier values, and reduce variables because they are highly correlated (they provide the same information), among other things. Thus, a dataset of 845 records and 34 columns is obtained. Finally, an analysis of the most significant variables to determine the quality of the coffee bean was made, for which a Pearson and Spearman correlation analysis was carried out, which determined that they are flavour, aroma, consistency, and acidity. These variables were used to build the unsupervised model.

4.2.1. Coffee Bean Quality Model

For the definition of the quality model was used three unsupervised learning techniques: K-means, Agglomerative Clustering and DBScan. Thus, two distance-based clustering techniques and one density-based clustering technique are used. In the case of K-means, to determine the number K, the score for different values of K is plotted (see Figure 2). The optimal K for this problem is three.

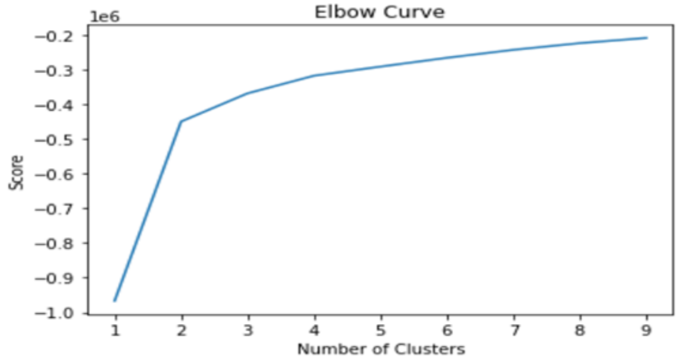

Figure 2. Determination of K.

Similarly, in the Agglomerative Clustering technique is necessary to define the number of groups. Figure 3 shows the dendrogram with the results of this technique to identify the number of categories. According to Figure 3, different clusters can be defined. For example, two is a good number, but so is 3 or 4. Like the previous model, we have selected 3 categories of coffee quality.

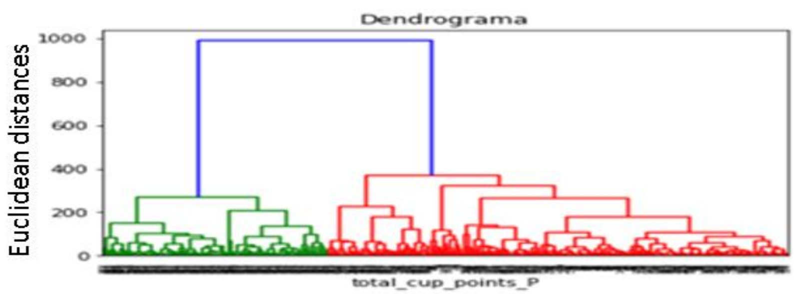

Figure 3. Dendrogram for the Agglomerative Clustering techniques in our problem

Finally, a density-based model is analyzed. In this case, we have used DBSCan with the next values for the parameters: for the radio (eps) 9.92 and for the number of points necessary to form a dense region (minPts) 4. Figure 4 shows the results of this technique. Particularly, the distribution of the categories is not clearly identified by density groupings, and therefore most of the data go to the same cluster

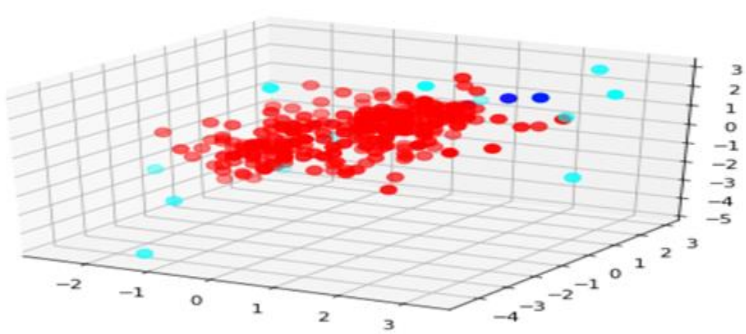

Figure 4. Distribution of Categories by DBSCan.

For the selection of the best model, we have used the Silhouette Index and the Density-Based Clustering Validation (Moulavi et al., 2014). Both metrics range from -1 to +1, where in the case of the silhouette index a high value indicates that the groups are well-constituted and far from their neighbors. While in the other metric, greater values of the measure indicate better density-based clustering solutions

Table 2. Comparison of models.

	Silhouette Index	Density-Based Clustering Validation
K-Means	0,85	-0,41
Clustering	0,64	-0,45
DBScan	0.41	-0.122

Clearly, the K-means model performs better because its silhouette index is the highest, close to 1. With respect to density-Based Clustering Validation all have very bad values. Therefore, this model is considered for the determination of coffee quality. Once the groups are obtained with K-means with three categories, we analyze the centroid of each group (flavour, aroma, consistency, and acidity) to the quality of the coffee (see Figure 5). For group 1, the values of flavor, aroma, consistency, and acidity represent good quality of the coffee.

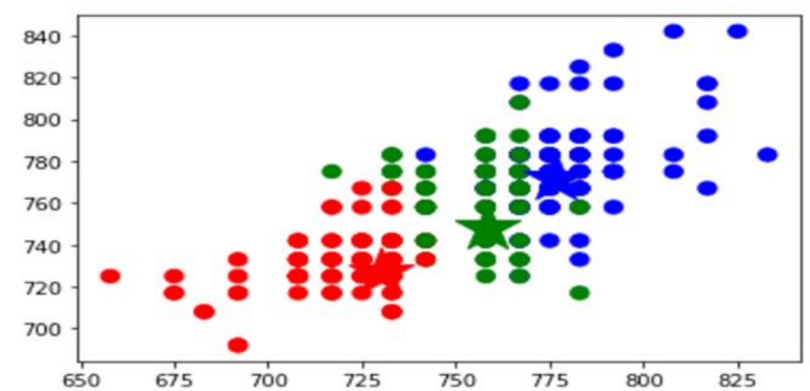

Figure 5. Results with K-means.

Thus, this model with three categories represents the quality of coffee as low, medium, and high quality, where the distribution of the data for the original dataset in each of the categories is shown in Table 3.

Table 3. Quality Vs Amount.

Quality	Amount
High	264
Medium	311

Low	270

Thus, the model built with K-means with 3 categories presents the best behaviour for the identification of the quality of coffee using the flavour, aroma, consistency and acidity variables.

4.2. Task for the prediction of the decrease of the bean in the roasting process

During the coffee roasting process, it is common for coffee beans to increase in size and lose moisture, decreasing the final weight of the batch (Porras-Zúñiga et al., 2019), (Abdul Majid et al., 2015). In the process, there is a set of product and process variables, some of which can be controlled to reduce the level of coffee bean loss. This task will focus on the design of a supervised model for predicting the level of bean decrease in the roasting process, considering the product input characteristics (One of them is the quality of the grain previously determined by the previous task) and the controllable variables of the process.

4.2.1. Feature engineering process

As before, a feature engineering process is carried out. The Dataset has 2,632 records and 23 variables, from which the bean decrease will be estimated. The date is transformed into datatime variable to obtain the day, month, year, and week values. It is noted in this part of the process that all the data were taken in a period between January and September 2022. Next, the bean decrease value is calculated for each record as:

$$Merma\ (wastage) = \frac{(Green\ weight - Toasted\ weight)}{Green\ weight}$$

It corresponds to the percentage of weight lost in the roasting process. In the case of the sample (see Figure 6), the loss ranges between 7.6% and 19.87% with 75% of the data above 14.5% and 50% of the data between 14.5% and 15.8%. There are very high and low values that can be considered outliers.

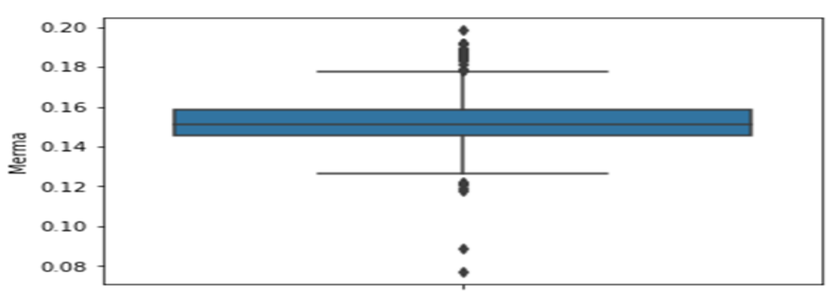

Figure 6. Box and Whisker Plot of Wastage (merma) Values in the Original Dataset

The decision is taken to use only the wastage values that are within the central behaviour, i.e., values between 12.5% and 18%. After cleaning the outliers, the dataset was reduced to 2103 records.

Now, the categories of wastage are defined. Theoretically, there are no values to define high, medium, or low wastage, so it was decided to use the 33rd and 66th percentiles as the limits between the three categories. Thus, low wastage is anything below 14.76%, high wastage is anything above 15.48%, and medium wastage is anything between these two values. Figure 7 shows the three categories, which are balanced, thus reducing the possibility of bias towards a specific class during the prediction process.

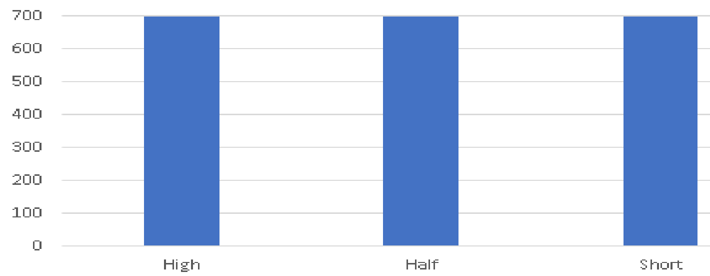

Figure 7. Number of observations per category

Finally, we analyze all the variables in the Dataset. It is observed in Figure 8 that some variables have very few values, which do not provide any relevant information to the analysis, and it is decided to exclude them.

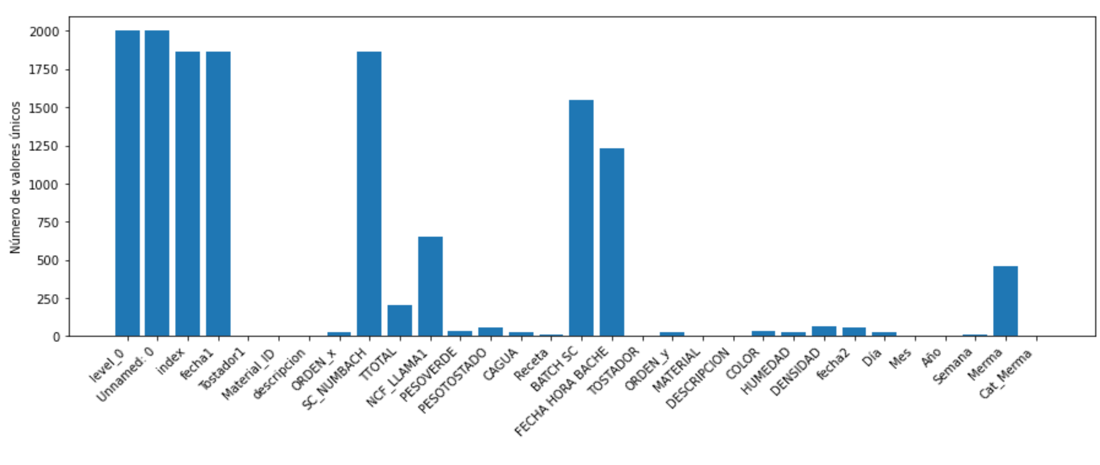

Figure 8. Number of unique values for each variable

Additionally, it is observed that the variable DENSITY has many null elements, so it is decided to exclude it from the analysis. Finally, only variables that contain sufficient relevant information and are considered controllable within the roasting process are selected. Such variables are "TTOTAL" (total roasting time), "NCF_LAMMA1" (flame power), "GREEN WEIGHT" (initial weight of the load), "CAGUA" (amount of water added in the process), "WEIGHT TOASTED" (total weight of the load after the roasting process), "CAGUA" (amount of water added in the process), "WEIGHT TOASTED" (total weight of the load after the roasting process), "WEIGHT TOASTED" (Total weight of the load after the roasting process), "RECIPE" (Type of recipe used), "COLOUR" (Colour of the final roasted load), "MOISTURE" (Final moisture content of the load) and finally "Cat_Merma", which is the category of the Merma (target).

Finally, to make better use of all the information offered by the selected variables, new variables are created that relate to the original numerical variables in pairs. Some of them are GREEN WEIGHT/HUMIDITY or TTOTAL/COLOUR, which have a clear physical explanation. In other cases, such as WEIGHT GREEN/NCF_FLAME, they may not be explainable a priori. The idea here is to provide the model with as much input information as possible, to achieve a prediction with a good degree of accuracy. Once the values of the numerical variables have been standardized, the correlation matrix between all the variables is calculated (see Figure 9).

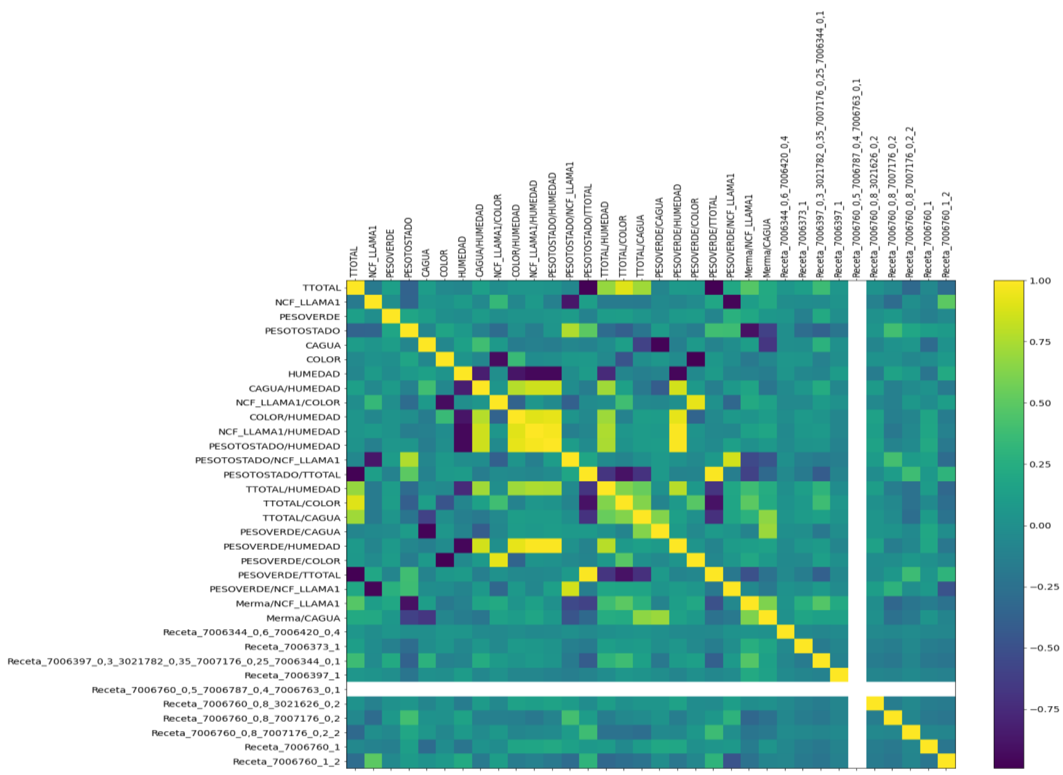

Figure 9. Heatmap of the correlation matrix between all input variables.

Finally, it is intended that only variables that can provide new information are used to build the model, which other variables do not offer. Then, it is decided to select those variables that are highly correlated with the target variable (merma), and when there are two highly correlated descriptor variables, exclude the one with less correlation with the target variable. Thus, it is decided as high Pearson correlation values those greater than 0.75 (which shows a color between light green and yellow) and less than -0.75 (which is shown in dark blue) in the heat map of Figure 9. The new dataset has been considerably reduced, eliminating many artificial variables. Finally, only three of these have been retained: "TTOTAL/WETTING", "TOTAL/CAGA" and "MERMA/CAGUA”. The highest correlation in the new data set is between the variables "TOTAL/MOISTURE" and "MOISTURE" with -0.75, which is at the lower limit of the selection parameter.

A final review of the variables in the dataset shows that there are two that do not correspond to an original characteristic of the product and that cannot be controlled by those responsible for the process: "WEIGHT TOASTING" and "Merma/CAGUA", as both are the result of the combination of other variables. Since the aim of this exercise is also to define how to vary the controllable values of the process from some initial product properties, it is decided to exclude these two non-controllable variables. The new Dataset consists of 17 input variables, one of which is the quality of the coffee bean, and nine of which are categorical. The output variable is "Cat_Merma". On the other hand, remains considerably well balanced, as shown in Figure 10. The 17 input variables are:

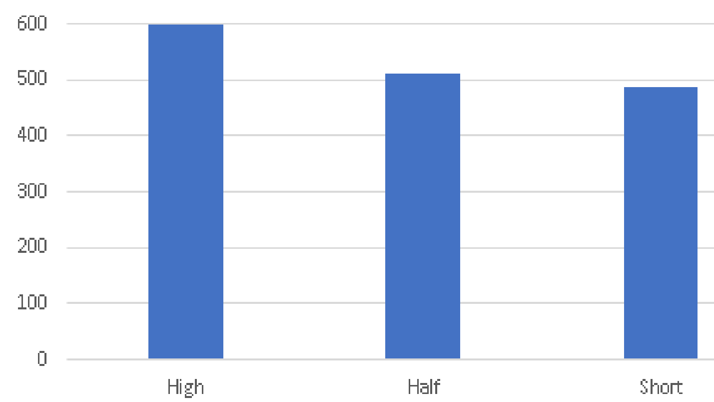

Figure 10. Distribution of observations by category of the “merma” variable in the final dataset

4.2.4. Wastage Prediction Model

To build the predictive model, three traditional supervised learning techniques were used: Random Forest, Logistic Regression and Support Vector Machine (Zhang et al., 2019). For the evaluation, the final dataset was separated into a training set and a test set, which correspond to 80% and 20% of the dataset, respectively. Then, the accuracy and confusion matrices were generated to analyze the results.

In the case of Random Forest, it was used with 50 estimators or trees trained, with results in the test phase of accuracy of 0.81. Taking advantage of the fact that Random Forest allows visualizing the impact of all input variables on the classifier variable, we review which input variables have a higher importance in the model (see Figure 11).

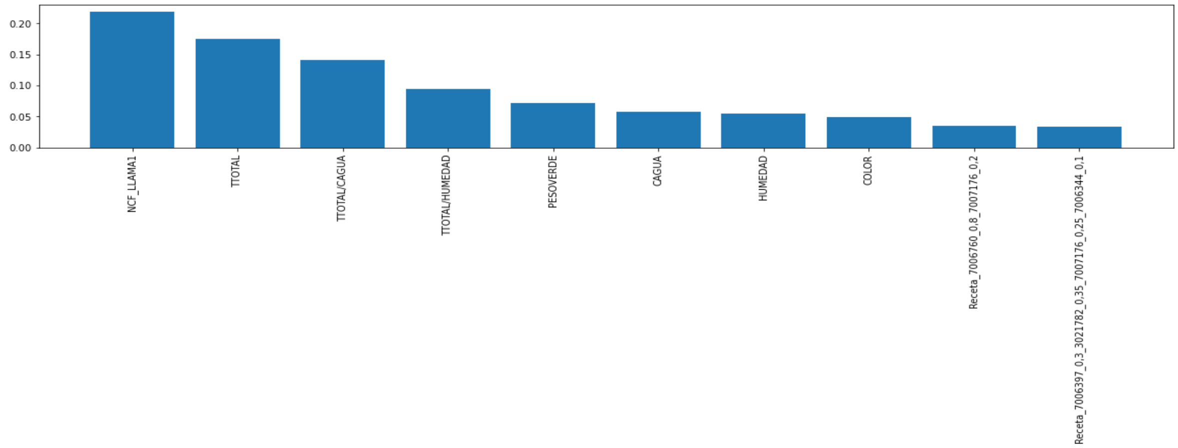

Figure 11. Contribution of the 10 most important variables in Random Forest

Of the 17 original variables, the 10 most important ones explain more than 90% of the behaviour. The final intensity of the flame, together with the total roasting time and the ratio between the total roasting time and the amount of water added, are the variables that have the greatest influence. This makes sense, considering that the longer the roasting time and the more powerful the flame, compared to the amount of water added, the lower the final weight of the load is expected to be and, therefore, the higher the decrease. The model is run again, including only the 10 variables that help the most in classification, and the results are found to be almost unchanged with a test accuracy of 78.8%.

When applying a logistic regression technique to the reduced dataset, it has a much lower performance than with Random Forest, with a test accuracy of 71%. There is, however, a logistic regression adjustment called LASSO regularization, which penalizes the effect of redundant variables that have high covariance with other variables and makes the coefficients associated with those variables exactly zero, which improves the regression performance in models with high dimensionality and covariance. Using LASSO regularization, the test accuracy is 73%. Thus, accuracy seems to have increased somewhat with regularization, leading to the conclusion that logistic regression with LASSO regularization is better.

A final technique is the Support Vector Machine. With this technique, an n-dimensional hyperplane is constructed that has the maximum possible distance to nearest-neighbours, which clearly separates the space into two sub-spaces, each corresponding to a class. Any new observation will be classified according to the subspace to which it belongs. One of the main parameters is C, which corresponds to the regularization of the algorithm's Loss function. We seek to select a setting for which the training error is minimal, to avoid overfitting. This is inversely proportional to the strength of the regularization, and its value is strictly positive. It has been decided to test different values of C, to use the one that generates the lowest test error (see Figure 12).

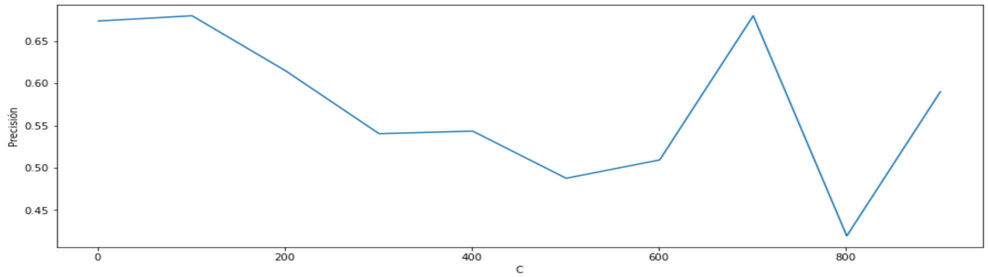

Figure 12. Variation of test accuracy for different values of C.

The lowest test error is achieved with C=101. However, the test accuracy is very low (0.68), compared to previously fitted models.

Of all the models tested, the best result was obtained with Random Forest (see Table 4), which had a training accuracy of 100% and testing of 80%. There is some concern that there is a problem of overfitting for future observations, however, the relatively high testing accuracy suggests that it could perform very well on unobserved data.

Table 4. Training and testing accuracy for all fitted models.

Model	**Train**	**Test**
Random Forest	100%	81%
Logistic Regression	67%	71%
Logistic Regression with LASSO	70%	73%
SVM con C=101	66%	68%

The confusion matrix of Random Forest is shown in Figure 13. In that case, the highest probability of false positives is for the middle class.

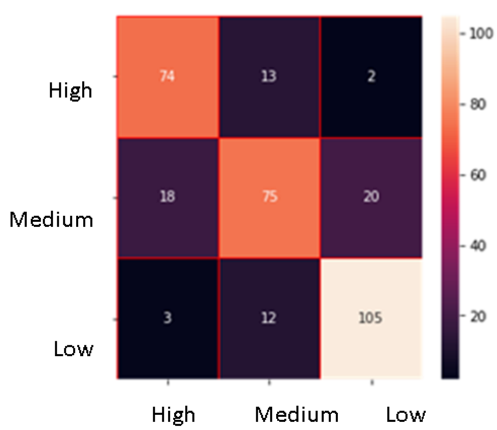

Figure 13. Confusion matrix test for Random Forest.

In terms of the variables selected for modelling, flame intensity, total roasting time and the amount of water added during the process are the most critical variables in defining the final level of decrease. During the roasting process, the coffee loses much of its moisture, and because of this, the weight of the beans will decrease. The model allows determining the level of reduction of the coffee bean or decrease in the weight of the beans. Although the scope of this project does not allow establishing the exact relationship in which variations in the values of these input variables affect the wastage values, it is possible to state that the three variables mentioned above, all controllable, should be closely monitored, as they are critical to achieve a reduction in decrease.

4.3. Task for the determination of the "Coffee processing method".

Apart from the quality of the coffee, it is also possible to establish the "Processing method". For this task, a dataset with 1347 records and 44 variables is studied, from which the "Coffee processing method" will be determined. Within these variables are also included the variables of bean quality and bean decrease in the roasting process previously estimated

4.3.1. Feature engineering process

The variables of the dataset are reviewed, identifying those that are relevant. The processing methods are washed/wet, Natural/Dry, Pulped Natural/Honey, semi-washed/semi-pulped, and others. The variables smell, flavor, aftertaste, acidity, body, balance, sweetness, moisture are analyzed with respect to their quartiles to determine the behavior of the data. They will allow classifying the most appropriate processing method. For example, the flavor variable is correlated with the processing method according to Figure 14.

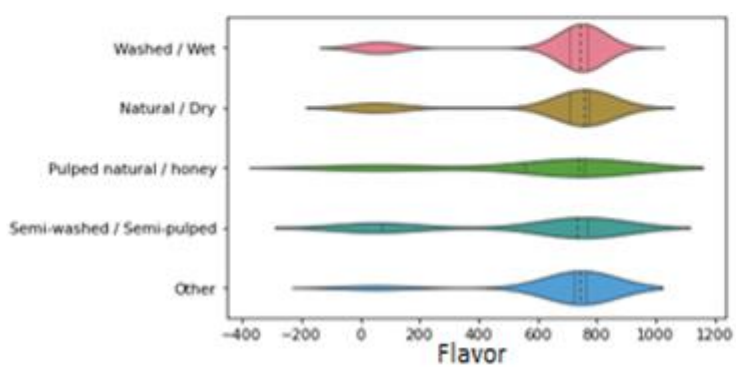

Figure 14. Flavor quartile and processing method

4.3.2. Identification Model of the "Method of coffee processing".

To build the identification model of the "Production method", we will use two machine-learning techniques: Logistic Regression and K nearest neighbours. The techniques were trained with 883 registers and tested with 464. In the case of K nearest neighbours, a test was carried out to determine the optimal value of k (see Figure 15).

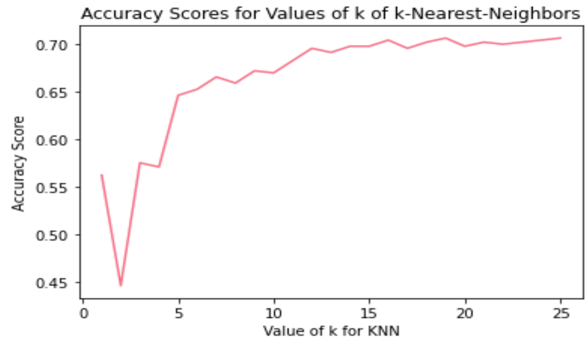

Figure 15. Test to determine the value of k.

The best result was obtained with logistic regression (0.72), with respect to K nearest neighbours (0.64) See Table 5.

Table 5. Training and testing accuracy for the models.

Model	Train	Test
Logistic Regression	0.69	0.72
K-nearest neighbors	0.63	0.64

5. Results Analysis

This research presents the analysis of three aspects of the coffee agro-industrial production chain such as quality, level of wastage and optimal coffee processing methods. We propose an autonomic cycle to analyze these aspects based on information from the organization and the context. The analysis tasks use supervised and unsupervised models.

The first task defines a diagnosis model based on unsupervised techniques to determine coffee quality. In this case, the best technique was K-Means with a silhouette index of 0,85. To determine the coffee production process, a main point is the coffee roasting process. The second task estimates the bean wastage due to this process based on a set of product and process variables that can be controlled. In this case, supervised classification techniques were used, and the best result was obtained with Random Forest, with a test accuracy of 81%. Regarding the variables, the intensity of the flame, the total roasting time and the amount of water added during the process are the most critical variables.

Another aspect that can also be established is the "Processing Method" in the productive chain. The cycle has a task to determine the coffee processing method, which analyzes the data to predict the coffee processing method. For this goal, the best result was obtained with logistic regression, whose value was 0.72. It can be concluded that with the input data such as aroma, flavor, aftertaste, acidity, body, balance, sweetness, moisture, the most appropriate processing method can be determined. The results of these tasks (e.g., quality of the inputs to be processed, the production method, etc.) allow making decisions to improve the production process.

6. Comparison with other works.

To compare this work with other similar works, we proceeded to define four criteria:

- **Criterion 1:** for the improvement of coffee production are analyzed contextual data.
- **Criterion 2:** the coffee production process is monitored using autonomic approaches.
- **Criterion 3:** machine learning techniques are used to analyze the contextual data.

Table 11. Comparison with previous work.

	Criterion1	Criterion 2	Criterion 3
(Eshetu et al., 2022)	x	x	x
(Guevara-Sánchez et al., 2019)	√	x	x
(Puerta-Quintero, 2016),	x	x	x
(Firdissa et al., 2022)	x	x	x
(Campo, et al., 2020)	√	x	x
(Sánchez-Aguiar et al., 2019)	x	x	x
(Anita & Albarda, 2020).	x	x	√
(García et al., 2019)	√	x	√
This work	√	√	√

In (Guevara-Sánchez et al., 2019) used a prototype to determine the organoleptic quality. Also, mechanical drying was presented as the best option, indicating that its use can optimize the organoleptic quality of coffee. Thus, this work jointly analyzed the quality and processing methods of coffee. However, in the coffee production process, they do not use machine-learning techniques. In (Campo et. al, 2020), the study of the fragrance and aroma of roasted coffee with the Coffee-NOSE prototype is presented. The work analyzes contextual data in the processing of coffee for evaluating coffee quality.

(Anita & Albarda, 2020) delve into the coffee bean selection process using the dry method. Coffee beans are classified into ripe, undercooked, raw, and damaged using the GLCM (Gray-Level Co-Occurrence Matrix) method for feature extraction, and the KNN (k-Nearest Neighbor) and CNN (Artificial Neural Network) classification algorithms. Machine learning techniques are used to analyze contextual data.

In (García et al., 2019) is presented image processing and machine learning techniques integrated with an Arduino Mega board to evaluate four important factors when selecting the best quality green coffee beans. For this, the k-nearest neighbor algorithm is used to determine the quality of the coffee beans and their corresponding types of defects. Thus, they use machine-learning techniques to analyze data from the context.

As shown in Table 11, previous studies did not meet all criteria. We propose an autonomous cycle of data analysis tasks that are based on machine-learning techniques to study contextual data to automate coffee production.

7. Conclusions

The autonomous cycle of data analysis tasks to automate the production process of the Coffee Agro-industrial Sector of MSMEs obtains useful information to identify the quality of beans to transform, the level of decrease/waste of coffee beans during the roasting process, and the coffee processing method based on organizational and contextual data. For the construction of the cycle, three knowledge models (predictive, diagnostic, and identification) were defined using supervised and unsupervised machine learning algorithms.

One of the models is for the estimation of the quality of coffee beans considering specific information about the coffee bean such as its origin, recipe used, characteristics of the raw material, etc. With this information, the quality of the coffee was classified as High, Medium, and Low, which was the optimal number of categories for an adequate classification of the quality of the coffee bean. Another model is for the prediction of the level of decrease of the bean in the roasting process. In the process, there is a set of product and process variables, some of which can be controlled to reduce the level of final loss in the process. Regarding the variables selected for modelling, the intensity of the flame, the total roasting time, and the amount of water added during the process are the most critical variables. Furthermore, it is possible to identify the "Coffee processing method" using a supervised model allow determining the type of processing (e.g., washed/wet, natural/honey polished, semi-washed/semi-polished, among others) using, among other variables, those previously calculated. With these models, the cycle can make decisions to improve the production, but also, the company can identify the farmers who grow the best quality beans, among other things.

One of the future works is to integrate this autonomous cycle of identification of the production method with another to evaluate the level of digital transformation in the company, in order to determine the improvements in the production chain of the company. This last cycle would to identify the technological level to be improved, and would be responsible for identifying the technological level to introduce in the agro-industrial production chain according to the identified production method, such as the characterization of the processing plant, the identification of the raw material transformation processes, and the selection of the technology for its processing.

Acknowledgments

Jairo Fuentes holds a Bicentenary PhD scholarship funded by the Ministry of Science, Technology, and Innovation (Minciencias). All authors would like to thank the Vice-Rector's Office for Discovery and Creation, Universidad EAFIT, for their support in this research. The authors would like to thank the HAMADI 4.0 project, code 22-STIC-06, of the STIC-AmSud regional program, for supporting this work.

References

Abdul Majid, N. A., Edzuan Abdullah, M. F., & Diana, A. (2015). A Review of Quality Coffee Roasting Degree Evaluation. *Journal of Applied Science and Agriculture*, *10*(7), 18–23.

Aguilar J., Buendia O., Pinto A. Gutiérrez J. (2022) Social learning analytics for determining learning styles in a smart classroom, *Interactive Learning Environments*, 30(2) 245-261, DOI: 10.1080/10494820.2019.1651745

Aguilar J. Garcès-Jimènez A., Gallego-Salvador N., Gutierrez J. Gomez-Pulido J. Garcìa-Tejedor A. (2019) Autonomic Management Architecture for Multi-HVAC Systems in Smart Buildings, *IEEE Access*, vol. 7, 123402-123415.

Anita, S., & Albarda. (2020). Classification Cherry's Coffee using k-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). *2020 International Conference on Information Technology Systems and Innovation, ICITSI*, 117–122. https://doi.org/10.1109/ICITSI50517.2020.9264927

Chamba-Eras L., Aguilar J (2017) Augmented Reality in a Smart Classroom—Case Study: SaCI, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 12(4) 165-172, doi: 10.1109/RITA.2017.2776419

Corrales, D., Ledezma, A., Peña Q., Hoyos, J., Figueroa, A., & Corrales, J. (2014). A new dataset for coffee rust detection in Colombian crops base on classifiers. *Sistemas y Telemática*, *12*(29). https://doi.org/10.18046/syt.v12i29.1802

Eshetu, E. F., Tolassa, K., Mohammed, A., Berecha, G., & Garedew, W. (2022). Effect of processing and drying methods on biochemical composition of coffee (Coffea arabica L.) varieties in Jimma Zone, Southwestern Ethiopia. *Cogent Food and Agriculture*, *8*(1). https://doi.org/10.1080/23311932.2022.2121203

Firdissa E, MohammedA,Gezahegn A., Garedew W. (2022) Coffee Drying and Processing Method Influence Quality of Arabica Coffee Varieties (Coffee arabica L.) at Gomma I and Limmu Kossa, Southwest Ethiopia, Journal of Food Quality. 2022. https://doi.org/10.1155/2022/9184374

García, M., Candelo-Becerra, J. E., & Hoyos, F. (2019). Quality and defect inspection of green coffee beans using a computer vision system. *Applied Sciences (Switzerland)*, *9*(19). https://doi.org/10.3390/app9194195

Guevara-Sánchez, M., Bernales del Águila, C. I., Saavedra-Ramírez, J., & Owaki-López, J. (2019). Effect of altitude on coffee (Coffea arabica L.) quality: Comparison between mechanical and traditional drying. *Scientia Agropecuaria*, *10*(4), 505–510. https://doi.org/10.17268/sci.agropecu.2019.04.07

Horta De Oliveira, G., de Oliveira, A., Botelho, F., Treto, P., & Botelho, S. (2018). Coffee Quality: Cultivars, Blends, Processing, and Storage Impact. *Journal of Food Quality*, *2018*. https://doi.org/10.1155/2018/9805635

Horn, P. J. (2001) Autonomic computing: Ibm's perspective on the state of information 130 technologies, 2001. http://www.research.ibm.com/autonomic/manifesto

Kim, J. (2022). Coffee Beans Quality Prediction Using Machine Learning. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.4024785

Lince Salazar, L., & Sadeghian Khalajabadi, S. (2015). Producción de café (Coffea arabica L.) en respuesta al manejo específico de la fertilidad del suelo. *Revista de Investigación Agraria y Ambiental*, *6*(2), https://doi.org/10.22490/21456453.1401

Montalbo, F., & Hernández, A. A. (2020). An Optimized Classification Model for Coffea Liberica Disease using Deep Convolutional Neural Networks. *Proceedings16th IEEE International Colloquium on Signal Processing and Its Applications,* 213–218. https://doi.org/10.1109/CSPA48992.2020.9068683

Morales, L., Ouedraogo, C., Aguilar, J. Chassot C., Medjiah S., Drira K. (2019) Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform. *Service-Oriented Computing & Applications,* 13, 199–219. https://doi.org/10.1007/s11761-019-00266-w\

Moulavi D., Jaskowiak P., Campello R. Zimek A. Sander J. (2014) Density-Based Clustering Validation Proceedings 14th SIAM International Conference on Data Mining.

Porras-Zúñiga, M. C., Vargas-Elías, G., Araúz- Madrid, L., & Abarca-Alpízar, Y. N. (2019). Efecto de la temperatura en la rapidez del tostado de café. *Revista Tecnología En Marcha*, *32*, 20–27. https://doi.org/10.18845/tm.v32i7.4255

Puerta G. (2016), "Physical quality of coffee from various regions of Colombia according to altitude, soils and good processing practices", journal Cenicafé 67 (1): 7-40.

Puerta-Quintero, G. (2016). *Calidad Física Del Café De Varias Regiones De Colombia Según Altitud, Suelos Y Buenas Prácticas De Beneficio Physical Quality of Coffee From Various Regions of Colombia According To Altitude, Soils and Good Processing Practices*. *67*(1), 7–40.

Rocha, Á., Correia, A. M., Adeli, H., Reis, L. P., & Teixeira, M. M. (2016). Preface. *Advances in Intelligent Systems and Computing*, *445*(V–VI). https://doi.org/10.1007/978-3-319-31232-3

Sánchez-Aguiar, A. F., Mario Ceballos-Arroyo, A., & Espinosa-Bedoya, A. (2019). Toward the recognition of non-defective coffee beans by means of digital image processing. *Proceedings 22nd Symposium on Image, Signal Processing and Artificial Vision*. https://doi.org/10.1109/STSIVA.2019.8730267

Sánchez M., Aguilar J., Cordero J., Valdiviezo-Díaz P., Barba-Guamán, L. Chamba-Eras L. (2016). Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service. Advances in Intelligent Systems and Computing, 444, 993–1002. https://doi.org/10.1007/978-3-319-31232-3_94

Tamayo-Monsalve, M. A., Ruiz, E. M., Pulgarin, J. P. V., Ortiz, M. A. B., Arteaga, H. B. A., Rubio, A. M., Alzate-Grisales, J. A., Garzon, D. A., Cano, V. R., Arias, S. O., Osorio, G., & Soto, R. T. (2022). Coffee Maturity Classification Using Convolutional Neural Networks and Transfer Learning. *IEEE Access*, *10*, 42971–42982. https://doi.org/10.1109/ACCESS.2022.3166515

Vizcarrondo J. Aguilar J., Exposito E. Subias A (2017) MAPE-K as a service-oriented architecture," in *IEEE Latin America Transactions*, 15(6) 1163-1175, doi: 10.1109/TLA.2017.7932705

Zhang, Y., Wei, S., Zhang, L., & Liu, C. (2019). Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features. *Journal of Medical and Biological Engineering*, *39*(3), 381–392. https://doi.org/10.1007/s40846-018-0411-0

Appendix D. Evaluation of the Level of Digital Transformation in MSMEs using Fuzzy Cognitive Maps based on Experts.

Evaluation of the Level of Digital Transformation in MSMEs using Fuzzy Cognitive Maps Based on Experts

Jairo Fuentes
Universidad Francisco de Paula *Santander, Cúcuta, Colombia*
GIDITIC, Universidad EAFIT, Medellín, Colombia
fuentesjairo@gmail.com

Jose Aguilar
GIDITIC, Universidad EAFIT, Medellín, Colombia
CEMISID, Universidad de Los Andes, Mérida, Venezuela
IMDEA Networks Institute, Madrid, Spain
aguilar@ula.ve

Edwin Montoya
GIDITIC, Universidad EAFIT, Medellín, Colombia
emontoya@eafit.edu.co

William Hoyos
Grupo de Investigación ISI Universidad Cooperativa de Colombia
Montería, Colombia
GIDITIC, Universidad EAFIT, Medellín, Colombia
william.hoyos@campussucc.edu.co

***Abstract*— the concept of digital transformation involves exploiting digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models, and services that produce value based on the digitization of data and processes. The application of digital technologies enables organizations to develop capabilities for innovation, automation, etc., utilizing both established and emerging technologies, including the widespread use of artificial intelligence. In this paper is proposed the implementation of a fuzzy cognitive map based on experts for the evaluation of the level of digital transformation in MSMEs (Micro, Small & Medium Enterprises). The main variables of digital transformation used to define our fuzzy cognitive map were classified into five types: i) Organization and Culture variables related to strategies, way of working, and ecosystems, ii) Customer variables related to services and digital channels and products, iii) Operations and Internal Processes variables related to supply chain, suppliers, business model, iv) Information Technologies variables related to innovation, digitization, data and analytic. Finally, the fifth type of variable is the target, which indicates the level of digital transformation of the organization. Our model managed to specify with 99.4% the level of digital transformation of the organization. Thus, our fuzzy cognitive map can predict and analyze the factors associated with digital transformation in MSMEs.**

Index Terms—**Digital Transformation, Maturity models, Fuzzy cognitive maps, Artificial Intelligence**

I. INTRODUCTION

Digital transformation is the process of integrating digital technologies in the business model, processes, and culture in an organization to improve its strategies, and performance, among other things [1]. Digital transformation is a long-term endeavor that entails determining how technology is employed within an organization to affect its behavior. Thus, digital transformation involves envisioning how digital technologies can improve service levels, reduce costs, and optimize inventory management, among other things, through the implementation of organizational changes using these technologies to achieve operational excellence.

Advanced technologies, such as artificial intelligence, data analytics, and the Internet of Things, are integrated into the digital transformation process to enable organizations to collect and analyze real-time data, automate processes, and improve the supply chain.

On the other hand, fuzzy cognitive maps (FCMs) are graph-based structures that support causal reasoning and enable the propagation of causality in both forward and backward directions [2]. They are particularly useful in domains where the concepts and relationships between them are inherently fuzzy, such as politics, military affairs, and historical events. They have been used in different areas where the modeling of the problem is very complex, such as health [3], analysis of social networks [4], or emerging processes [5], [21], among others.

A. Related Works

At the level of FCM in classification and prediction tasks, the approach described in [3] combines different indicators used in the standard diagnosis of dengue, such as clinical symptoms, laboratory tests, and physical findings to determine the severity of dengue. The model achieved an accuracy of 89.4% in diagnosing dengue, indicating a slightly lower performance than other well-established machine learning models like artificial neural networks (97.9%) and support vector machine (98.1%). Also, to study the vulnerability factors to climate variability in Andean rural micro-watersheds, an FCM was employed in [4], the results of the FCM application did not present an important difference between the two climatic periods to which it was applied. Another work based on an FCM, and a fuzzy linguistic methodology, called VIKOR, determines the vendor selection in the Industry 4.0 [5]. This research aims to emphasize the supplier selection process through using FCM. VIKOR enables the ranking and selection of suitable suppliers [5].

The investigation carried out in [6] analyzes the effect of digital transformation on the performance of MSMEs in terms of competitive advantage. The study revealed that competitive advantage acted as a mediator in the association between digital transformation and the performance of MSMEs, and all proposed hypotheses were confirmed. The results of this research highlight the positive influence of digital transformation on competitive advantage and the overall performance of MSMEs. In [7], the goal is to develop a holistic maturity model for digital transformation using a theory-based approach, known as DX-CMM. This model aims to help organizations assess their current level of capability and maturity in digital transformation, identify gaps, and create a standardized roadmap for enhancing their digital transformation efforts.

According to [6], technological advancements and digital information are expected to be utilized by traditional entrepreneurs. However, this study concluded that all proposed hypotheses were acceptable, and that competitive advantage partially mediated the relationship between digital transformation and MSMEs' performance. In [8], the authors propose a staged digital transformation capability maturity model framework that enables organizations to assess their present digital capability and establish a plan of improvements to move to a higher level. In [7] conducted a systematic literature review and found that none of the 18 existing maturity models in the field of digital transformation met all the criteria of appropriateness, completeness, clarity, and objectivity.

According to the literature review, there is no digital transformation approach that enables firms to evaluate their existing digital competencies and create a roadmap for improvement to reach a higher level. Thus, there is a lack of research and artifacts available for maturity models related to digital transformation.

B. Contributions

The main goal of this research is to utilize FCMs to evaluate the digital transformation levels in MSMEs. The FCM is responsible for analyzing the factors associated with digital transformation data and providing a diagnosis. The causal relationships between the digital transformation concepts are optimized to match the experts' opinions.

The study is based on five groups of concepts issued by the Governance and Management Objectives in COBIT [7]. The concepts are classified into five types and reflect the behavior of the company at all levels such as: i) Organization and Culture, ii) Customer, iii) Operations and Internal Processes, iv) Information Technology, and v) Objective, which indicates the digital transformation level of the organization. The main contributions of this paper are:

- Develop an FCM to determine the level of digital transformation in MSMEs.
- Analyze the behavior of factors associated with digital transformation in MSMEs.
- Carry out an explainability analysis for MSMEs about their current situation in digital transformation.

This paper is organized as follows: Section 2 discusses the theoretical framework employed, Section 3 defines our FCM for the evaluation of the digital transformation levels in MSMEs, Section 4 details the computational experiments conducted, Section 5 compares this research with previous studies, and Section 6 presents the conclusions and future works.

II. THEORETICAL BACKGROUND

This section presents a background of Digital Transformation in MSMEs. In addition, it provides the basic concepts of FCMs.

A. Digital Transformation

The application of digital technologies to generate significant changes in an organization with the goal of improving its operations is what defines digital transformation. [9]. Digital transformation is based on the adoption of digital technologies in an organization to transform its business processes. The interpretation of "digital technology" in the context of digital transformation differs across the literature. Some authors encompass new Information Technologies solutions [10], and others emphasize in cutting-edge technologies like big data, cloud computing, artificial intelligence, blockchain, and the Internet of Things [11]. Also, some authors include social media analytics, and embedded devices [12].

As described in [13], there are two possible strategies for companies to undertake digital transformation. The first is an offensive approach that involves investments in digital transformation portfolios, while the second is a defensive approach that focuses on cultivating new capabilities within the organization.

The COBIT model classifies the concepts into five types:

a) Organization and Culture: variables related to the culture and management of the organization. The concepts of this variable used in this work are:

- *Strategy*, it refers to having an action plan for digital transformation and an understanding of the competitive environment,
- *Culture and management* refer to good practices at work, in the management of organizational processes and in the development of leadership, promoting human talent in its different areas and levels, creating an agile and innovative culture,
- *Way of working* refers to the type of face-to-face work and/or teleworking, and the practices and procedures implicit in them,
- *Leadership* refers to the managerial activities that organizational leaders have to influence the work team, and
- *Ecosystems* refers to the environment, not only internal, but also close external to the organization, which impacts the digital transformation process, to analyze its economic/social/cultural effect.

b) Customer: variables related to services (digital or not). The concepts of this variable used in this work are:

- *Digital channels* are the distribution channels of the different organizational strategic messages sent to interested parties.
- *Products*, services, and/or products offered by the organization as a result of its job or operation., and
- *Digital services* refer to the use of digital tools to improve the customer experience, based on a deep understanding of the customer and its needs.

c) Operations and Internal Processes: variables related to the organizational flow. The concepts of this variable used in this work are:

- *Supply chain* are the supply chains of goods and services that an organization requires for its operation,
- *Suppliers* refer to having providers of technological solutions, in addition to having advisors who accompany the digital transformation process of the organization, and
- *Business model* refers to the organizational capacity to explore new businesses based on technology, the development of new products that allow new markets to be captured, among other things.

d) Information Technology: variables related to the organization technology. The concepts of this variable used in this work are:

- *Innovation* refers to the business strategy of organizational transformation and human resource management based on reengineering, optimization, and process improvement,
- *Digitization* is the process of transforming analog or intangible processes and physical or tangible elements into digital ones,
- *Data and analytics* refer to the level of access to data that an organization has, the potential to monetize it by learning about its different uses, among other things, and
- Utilization of *technological solutions* involves implementing methods, systems, or tools that aid in the execution of organizational tasks with greater efficiency, as well as automating certain processes.

B. Fundamentals of FCMs

In 1986, Kosko developed FCMs [14] from the cognitive maps of Axelrod's work in 1976 [27]. FCMs are a popular way of modeling complex systems because of their straightforward construction and interpretation. They consist of a network of concepts and their interrelationships [14], [15], [18].

1. **Basic definitions of an FCM**

In Figure 1, an FCM is shown, which consists of eight concepts (C1 to C8). Each concept denotes a specific system variable of the system to study. The impact of one concept on another is displayed by the weight in the directed edges among them [16]. Different strategies can be employed to establish these relationships within FCMs. In [18], [22], [25] Aguilar et al. describe three ways to define the relationships in FCMs: using fuzzy rules, generic logical rules that describe the causal relationships, or mathematical models to describe the system under evaluation.

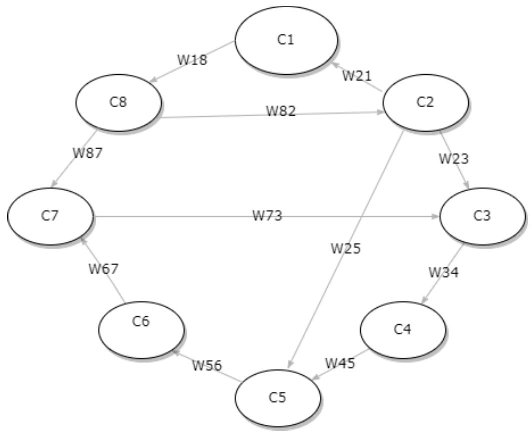

Figure 1. Example of an FCM.

The weight assignment in FCMs can be carried out using different methods, including those involving expert opinion or learning processes [18], [22]. The weights assigned represent the influence of one concept on another, with values ranging from -1 to +1 or 0 to 1, indicating either activating or inhibiting impacts [2]. Thus, different levels of causality are represented. An example of degrees of causality using linguistic terms can be found in Table 1. For example, if the value of the antecedent concept Ci is significantly low, and the value of the consequent concept Cj is substantially high, then the causal link between the two concepts can be categorized as negatively high. Mathematically, an FCM is represented as a tuple of four elements:

$$\Phi = \langle n, f(\cdot), r, W \rangle. \qquad (1)$$

where $n \in \mathbb{R}^m$ is the set of concepts $(n_1, \ldots, n_m)$, $f(\cdot)$ is the activation function that holds concept values in a determined range r, and $W \in \mathbb{R}^{m \times m}$ is the adjacency matrix employed to capture the connections among the concepts.

Table 1. Example of generic logic rules used to classify the causal relationships based on the values of concepts.

Linguistic term	Numerical value	Generic logic rule	
		Antecedent C_i	Consequent C_i
Positive complete	1	Very high	Very high
Positive high	0.75	High	High
Positive moderate	0.5	Medium	Medium
Positive low	0.25	Low	Low
Null	0	NA	NA

There are various activation functions that can be used to model the activation of each concept in an FCM (see Table 2). The selection of a specific activation function depends on the specific problem being addressed. For example, if the goal is to model disease symptoms, the sigmoid function may be more appropriate than the hyperbolic tangent function since the values of the concepts being modeled will typically range from 0 (absence of the symptom) to 1 (presence of the symptom), making negative values unnecessary [18], [19].

Table 2. Activation functions used in FCMs.

Activation function	Equation	Range
Bivalent	$f(x)=\begin{cases}1 & x>0\\ 0 & x\le 0\end{cases}$	$f(x)\in\{0,1\}$
Trivalent	$f(x)=\begin{cases}1 & x>0\\ 0 & x=0\\ -1 & x<0\end{cases}$	$f(x)\in\{-1,0,1\}$
Sigmoid	$f(x)=\frac{1}{1+e^{-\lambda x}}$	$f(x)\in[0,1]$
Hyperbolic tangent	$f(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$	$f(x)\in[-1,1]$

W is used to store the relationships between the concepts in an FCM. For example, see the adjacency matrix for the FCM of Fig 1 in Fig. 2.

$$
\mathbf{W}=\begin{array}{c}
\\ \boldsymbol{C_1}\\ \boldsymbol{C_2}\\ \boldsymbol{C_3}\\ \boldsymbol{C_4}\\ \boldsymbol{C_5}\\ \boldsymbol{C_6}\\ \boldsymbol{C_7}\\ \boldsymbol{C_8}
\end{array}
\begin{array}{c}
\begin{array}{cccccccc}\boldsymbol{C_1} & \boldsymbol{C_2} & \boldsymbol{C_3} & \boldsymbol{C_4} & \boldsymbol{C_5} & \boldsymbol{C_6} & \boldsymbol{C_7} & \boldsymbol{C_8}\end{array}\\
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{18}\\
w_{21} & 0 & w_{23} & 0 & w_{25} & 0 & 0 & 0\\
0 & 0 & 0 & w_{34} & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & w_{54} & 0 & w_{56} & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & w_{67} & 0\\
0 & 0 & w_{73} & 0 & 0 & 0 & 0 & 0\\
0 & w_{82} & 0 & 0 & 0 & 0 & w_{87} & 0
\end{pmatrix}
\end{array}
$$

Figure 2. Example of Adjacency Matrix.

2. Reasoning in FCMs

To mathematically define the inference process in an FCM, three key components are required: the weight matrix W, the activation function, and the current state vector of the FCM $a \in \mathbb{R}^m$, which indicates the current degree of activation of each concept. The activation degree of a concept determines if the concept is activated or stimulated in a particular iteration. Thus, in the case of the range [0, 1], a concept is "activated" when its value is 0 in one iteration (t) but becomes greater than 0 in the next iteration (t + 1).

To infer the output in an FCM, it is necessary a process that involves calculating the state vector through successive iterations of multiplying it with the weights matrix until the system reaches a steady state, as is defined by Equation 2 [14]:

$$a_j(t+1)=f\left(\sum_{i=1,i\neg j}^{m} W_{ij}a_i(t)\right) \qquad (2)$$

where a_j (t + 1) is the value of concept C_j at iteration t + 1, m is the number of concepts, W_{ij} is the value for the relationship from concept C_i to concept C_j, and $a_j(t)$ is the value of concept C_j at iteration t. The point of equilibrium (steady state) is reached when aj (t) = aj(t − 1) or aj (t) = aj(t − 1)<0.001. Thus, the system is considered to have reached a steady state when the difference between the state vectors in consecutive iterations is less than or equal to 0.001.

This is the base of the reasoning mechanism of an FCM. FCMs are a crucial tool in simulation scenarios because they enable experts to analyze the system's performance from various starting points. An initial condition is defined by the set of concepts in the iteration (0) and is denoted as:

$$a(0)=\left[a_1(0),a_2(0),\ldots,a_m(0)\right] \qquad (3)$$

where $a_1(0)$ is the value of concept C_1 at iteration = 0.

III. OUR FCM TO EVALUATE THE LEVEL OF DIGITAL TRANSFORMATION IN COMPANIES.

This section details the process to construct our evaluation model about the level of digital transformation in organizations utilizing FCMs. Our FCM is based on experts in the field to assign the relationships. Thus, the construction of the FCM is based on a methodology that consists of six steps [20], [23], [24] as depicted in Figure 3
.

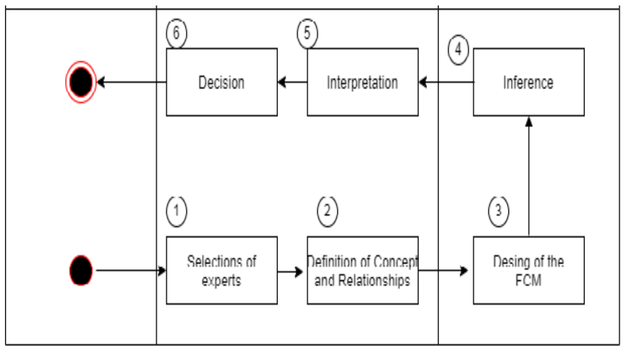

Figure 3. Methodology to build the FCM.

The rest of this section briefly describes these steps applied to our FCM.

A. *Selection of experts*

This study included the participation of five experts with more than five years of experience in the area of digital transformation. The experiences they have are in the fields of administration of digital transformation processes in different companies and government entities, in the study of the Cobit Model, in the field of new business models in the field of the Digital Economy, Government Digitization, among other areas.

B. *Concepts and relationships*

The initial phase of this process is carried out by experts in digital transformation. Based on the established concepts of the COBIT model (see Table 3), the experts constructed W with values ranging from 0 to 1 to indicate the relationships between the concepts. Table 3 shows an additional concept to those used in the COBIT model, which is the concept of the digital transformation level. This concept is the concept that is inferred using the FCM.

Using Table 1 as a guide, the experts selected the linguistic term that best represented the relationship between each concept, and then assigned numerical values to these terms to create the weight matrix. Table 1 allowed establishing the relationships of the concepts between each category and between categories. For example, digital services of the customer category were related to strategies and ecosystems of the organization, and culture category with product of the customer category, business model of the operations and internal processes category, among others.

Table 3. FCM Concepts for Digital Transformation.

Node	Name Concept
C1	Strategy
C2	Culture and Management
C3	Way of working
C4	Leadership
C5	Ecosystems
C6	Digital Services
C7	Products
C8	Digital Channels

C9	Suppliers and Supply Chain
C10	Business Model
C11	Processes
C12	Innovation
C13	Digitalization
C14	Technological Solutions
C15	Data and Analytics
End Node	**Digital Transformation**

C. FCM design

In our approach, the FCMs is developed by the experts, so their knowledge should be mixed into a single overall map to define the FCM. The process for producing a consolidated global map is described by the equation below [21], [22]:

$$E_{ij}^{G} = \sum_{e=1}^{NE} \frac{E_{ij}^{e}}{NE} \qquad (4)$$

This equation calculates the overall weight for the FCM, represented by E^{G}_{ij}. The opinion of expert e regarding the causal link between concepts C_i and C_j is represented by E^{e}_{ij} and NE indicates the total number of experts who participated. Finally, the overall FCMs are constructed. In our case, we have used the FCM Expert package (version 1.0.0) [2] but in the literature, there are others like [22], [25], [26]. Figure 4 shows the final FCMs. The construction of the FCMs with the experts has an implicit bias due to their subjectivity in assigning causal relationships [16], [18], [22], [25]. Addressing bias in FCMs is beyond the scope of our study, but it will be considered in future work.

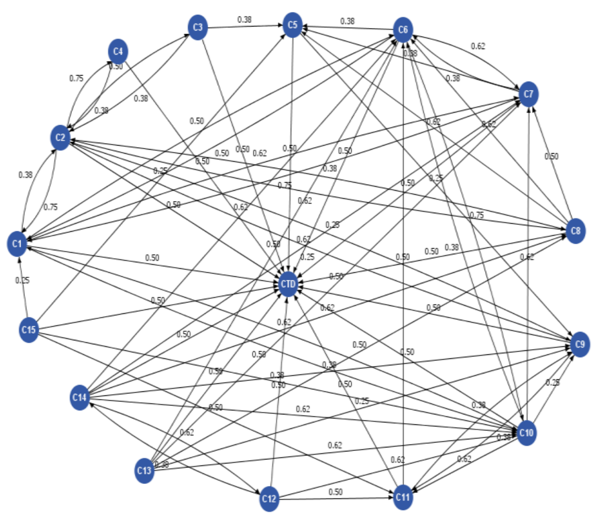

Figure 4. General FCM for evaluating the level of digital transformation in companies.

D. Inference

The fourth step involves the utilization of the FCM defined in the study to analyze different scenarios (for example, in our case, the level of digital transformation of different organizations) [18], [21]. In this step, the inference process is carried out for each case study, following the reasoning process outlined in section III-B-2. The experiment design and the results of this stage are shown in section 4.

Interpretation

The interpretation of FCM inference results for each case study is carried out in this step and is detailed in section 4.

Decision

The last stage is the decision that the user of our FCM makes to improve the digital transformation process in the organization. Since FCM is a method that provides an analysis of explainability based on causal relationships between concepts, this information can be used in the decision-making process to identify what to do.

IV. EXPERIMENTS

In this section, we explore the diagnostic capability of our FCM in different scenarios. We utilize the FCM to categorize digital transformation levels in several organizations. Particularly, we used two organizations as case studies. Specifically, the COBIT concepts in Table 3 are defined after an exhaustive evaluation of the respective concepts in each organization, so that from them, the FCM infers the concept of the level of digital transformation in that organization. No data was collected, what was done was instantiate the FCM concepts in each scenario (organization), for which the compliance of each one of them in said organizations was analyzed, studying their information, to see the behavior of the FCM.

Case 1. For the company Chamber of Commerce.

The Chamber of Commerce is an organization formed by entrepreneurs, and owners of small, medium, or large companies. In order to support business efficiency, the Chambers of Commerce are expanding their portfolio to virtual services, which will allow their users greater agility in their procedures.

For the data collection process, input values of the concepts are obtained in a range of (0.0, 0.25, 0.50, 0.75, 1.0), where 0 is the absence of this concept and 1 is the maximum presence of this concept in the organization, these values were provided by the experts of the organization see table 4. Thus, in this organization, after an evaluation of the current state of each concept linked to the COBIT model of digital transformation, the following values were determined by the organizational experts: the concept of Strategy (C1) with a value of 0.7, Culture and Management (C2) of 0.7, Way of Working (C3) of 0.5, Leadership (C4) of 0.3, Ecosystems (C5) of 0.5, Digital Services (C6) of 0.7, Products (C7) of 0.6, Digital Channels (C8) of 0. 8, Suppliers and Supply Chain (C9) of 0.4, Business Model (C10) of 0.6, Processes (C11) of 0.4, Innovation (C12) of 0.8, Digitalization (C13) of 0.8, Technology Solutions (C14) of 0.9, and Data and Analytics (C15) of 0.4.

Table 4. Input values for case study 1.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0.7	0.7	0.5	0.3	0.3	0.7	0.6	0.8	0.4	0.6	0.4	0.8	0.8	0.9	0.4

Table 5 shows the results of the inference process. Particularly, the initial vector of concepts (Table 4) is passed to the inference rule (Eq. 2) and the final vector can be seen in Table 5, where each iteration is the dynamic evolution of the concepts. In the final vector, the digital transformation concept (CTD) reached a value of 0.9954. Based on this result, we can see that the variables related to digitization (0.638) and data and analytics (0.659) are the ones that most need to be strengthened in this organization.

On the other hand, the causal relationships that affect the digitization variable are those of the customer and operation groups, while the variables that affect the data and analysis variable are those of the organization and culture and operation groups. It is necessary to determine which variables of these groups to improve in order to have a positive impact on these two variables. Now, these two variables, despite not having such high values, do not impact the perceived level of digital transformation in the organization, which means that the rest of the variables with values greater than 0.7 allow inferring in the organization a high level of digital transformation.

Table 5. Output vector for case study 1.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0,936	0,8878	0,7417	0,7754	0,889	0,9532	0,9539	0,8336	0,9502	0,9502	0,8832	0,741	0,638	0,7129	0,659	0,9954

Case 2. Galavis Coffee Company.

It is a north Santander Company, founded in 1918, dedicated to the roasting and distribution of Colombian

Coffee.
Table 6 shows the input values of the COBIT concepts for this company, after an evaluation of their current states by organizational experts. The concept of Strategy (C1) has a value of 0.5, Culture and Management (C2) of 0.4, Way of Working (C3) of 0.3, and so for the rest.

Table 6. Input table for case study 2.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0,5	0,4	0,3	0,5	0,3	0,4	0,4	0,4	0,3	0,4	0,5	0,3	0,3	0,3	0,3

Table 7 shows the results of the inference process. The final vector of this organization can be seen in Table 7. In the final vector, the CTD decision concept reached a value of 0.9878. Also, in Table 7 we can see that variables related to innovation and technological solutions are the ones that most need to be enhanced in this organization.

As before, an explicability analysis could be done to determine for those two variables, which are the ones that affect them, and based on that, how to influence to improve them. But again, these two variables do not affect by themselves the high level of transformation inferred, since the rest of the variables have a value greater than 0.7, which contributes to that value of digital transformation.

Table 7. Output vector for case study 2

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0,963	0,8877	0,7417	0,7754	0,8891	0,9531	0,9539	0,8335	0,8718	0,9501	0,8831	0,657	0,741	0,652	0,7125	0,9878

Case 3. A Company with high technology.
In this case, we are going to suppose a company where all the Information Technology concepts have a value of 1 and the rest in 0. The results can see in Table 8.

Table 8. Output vector for case study 3

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0,9723	0,9358	0,7761	0,8208	0,9358	0,9822	0,9829	0,8833	0,9185	0,9807	0,931	0,9148	0,9278	0,8964	0,937	0,975

In the final vector, the CTD variable (decision concept) reached a value of 0.975. Based on this result, we can determine that the variables of the information technology group have a high impact to achieve a high level of digital transformation. Also, we can see that the variable related to the way of working, which is the lowest variable (0.776), needs to be boosted in this case. Finally, other variables can also be identified that can be strengthened, such as leadership with 0.82 and digital channels with 0.8964.

Case 4. A Company without technology.
In this case, we are going to suppose a company where all the Information Technology concepts have a value of 0 and the rest in 1. The results can see in Table 9.

Table 9. Output vector for case study 4.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0.9723	0.9359	0.7763	0.821	0.9359	0.9822	0.9828	0.8832	0.9184	0.9807	0.9312	0,692	0.7745	0,689	0.7362	0.9271

In the final vector, the CTD variable (decision concept) reached a value of 0.9271. Based on this result, we can determine that the rest of the variables alone can contribute to a high digital transformation process. Also, the variables related to innovation and technological solutions (which belong to the group of information technology variables) are the ones that need to be promoted in this case. Other variables that may require repowering such as data and analytics, ways of working, digitalization are also identified.

COMPARISON WITH OTHER WORKS IN DIGITAL TRANSFORMATION

To compare this study with other works, a set of four criteria were established, which are outlined below:

Cri1: It uses non-intrusive schemes for diagnosis.
Cri2: It uses machine learning in the diagnostic model.
Cri3: It explains the reason for the level of deduced digital transformation.

Table 10 presents a comparison of this study with previous works based on these criteria that are deemed pertinent in the context of agroindustrial MSMEs and their digital transformation. These criteria are significant as they pertain to the application of automated technologies in agroindustry and the utilization of machine learning to enhance the agroindustrial production process. By using machine learning in the diagnostic model, it allows identifying patterns among the data to make predictions. By explaining the reason for the level of digital transformation inferred, it will allow organizations to assess their current digital capability and establish an improvement plan to move to a higher level.

Table 10. Comparison with other works.

	Cri1	Cri2	Cri3
[3]	√	√	X
[4]	X	X	X
[5]	X	X	X
[6]	X	X	√
[7]	X	X	√
[8]	X	X	√
[9]	X	X	√
This work	√	√	√

The first criterion is met by the studies [3] and ours since both use computerized methods and models for diagnosis. The second criterion is met by [3] and our study, which uses machine learning for different purposes, such as information processing for dengue diagnosis [3]. The third criterion is met by studies [6], [7], [8], [9]. They explain the relationship between digital transformation and MSME performance, conduct digital transformation maturity assessments, and design a comprehensive understanding of digital transformation and its impact. and our study since it focuses on the digital transformation of agro-industrial MSMEs.

Our study meets all the requirements of a *diagnostic model* to analyze the digital transformation in an organization, in our case, based on FCMs. Our model is *not intrusive* because it uses organizational expertise to assess the current status of certain variables in the organization, and from there, it uses a computerized tool to make the diagnosis. It uses the FCM theory as the *machine learning technique*, which is an explanatory technique because an FCM allows *explaining the conclusions* that are reached through the causal relationships that exist between the concepts/variables, the basis of its inference process.

VI. CONCLUSION

The purpose of this study was to develop a FCM to determine the level of digital transformation in MSMEs. A computational tool is proposed to analyze the variables, which are divided into five specific types: i) Organization and Culture, ii) Customer, iii) Operations and Internal Processes, iv) Information Technologies, and v) Objective, which indicates the level of transformation digital in an organization according to the values in the previous variables in this organization.

The FCM allowed analyzing the behavior of factors associated with digital transformation in MSMEs. Our approach is an explicable and interpretive technique that permits the explanation of results based on the diverse concepts in the digital transformation of the COBIT model.

This approach is very useful because it allows detecting the concepts that need to be strengthened or improved in an organization to strengthen its digital transformation process to be more competitive. The developed model is easily customizable and flexible, making it simple to integrate additional concepts and relationships. For its use, only an initial vector of the current state of the variables related to the digital transformation and the weight matrix are required.

In the case studies of agroindustrial MSMEs, our FCM-based tool allowed determining the variables related to digital transformation to improve. It was found that the variables associated with digitization, data and analysis, innovation, and technological solutions, were those that should be improved.

In turn, our model allowed determining for the MSMEs of the agro-industrial sector analyzed, the variables that should have a high value at the end because they influence the most to have a high level of digital transformation in companies. According to the results obtained, they are the variables of Strategy (C1), Digital Services (C6), Products (C7) and Business Model (C10). If these variables have a high value, they guarantee to reach a high level of digital transformation in said organizations.

The FCM developed in this study has limitations. Firstly, it did not consider the Digital Maturity Model, which identifies the level of digitization of processes, installed capabilities, and weaknesses in MSMEs. Secondly, FCM models based on human knowledge may result in a biased evaluation of the accuracy of these models. Particularly, both the FCM topology and the weights were defined by the experts.

Future research should focus on minimizing the bias of the digital transformation evaluation model in an organization through a learning process of the FCM topology (and its weights) based on learning techniques using organizational historical data linked to the COBIT model variables. Furthermore, since the dataset used in this study is from a single city in Colombia, distributed strategies such as federated learning should be considered to create a global model, in this case of a country, that includes data from different cities.

ACKNOWLEDGMENT

All the authors wish to thank the EAFIT University's Vice President for Discovery and Creation for their support in this research. Likewise, the authors would like to thank the HAMADI 4.0 project, code 22-STIC-06, of the STIC-AmSud regional program, for supporting this work.

REFERENCES

[1] M. Bach, M. Spremić, D. Vugec, "Integrating Digital Transformation Strategies into Firms: Values, Routes and Best Practice Examples," *Manag. Technol. Challenges Digit. Age*, n, pp. 107–128, 2018, doi: 10.1201/9781351238922-5.
[2] G. Nápoles, M. L. Espinosa, I. Grau, K. Vanhoof, "FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps," *Int. J. Artif. Intell. Tools*, vol. 27, no. 7, 2018, doi: 10.1142/S0218213018600102.
[3] W. Hoyos, J. Aguilar, M. Toro, "A clinical decision-support system for dengue based on fuzzy cognitive maps," *Health Care Manag. Sci.*, vol. 25, no. 4, pp. 666–681, 2022, doi: 10.1007/s10729-022-09611-6.
[4] I. Bolaños-Portilla, I. Hurtado-Sánchez, I. Restrepo-Tarquino, "Diffuse cognitive maps for analysis of vulnerability to climate variability in andean rural micro-watersheds," *DYNA*, vol. 87, no. 212, pp. 38–46, 2020, doi: 10.15446/dyna. v87n212.79943.
[5] A. Hosseini, J. Heidary, J. Antucheviciene, M. Azari, S. Razavi, "Supplier selection in the industry 4.0 era by using a fuzzy cognitive map and hesitant fuzzy linguistic VIKOR methodology," *Environ. Sci. Pollut. Res.*, 2023, doi: 10.1007/s11356-023-26004-6.
[6] N. Yuliantari, N. Pramuki, "The Role of Competitive Advantage in Mediating the Relationship Between Digital Transformation and MSME Performance in Bali," *J. Ekon. Bisnis JAGADITHA*, vol. 9, no. 1, pp. 66–75, 2022, doi: 10.22225/jj.9.1.2022.66-75.
[7] E. Gökalp, V. Martinez, "Digital transformation maturity assessment: development of the digital transformation capability maturity model," *Int. J. Prod. Res.*, vol. 60, no. 20, pp. 6282–6302, 2022, doi: 10.1080/00207543.2021.1991020.
[8] T. Aguiar, S. Gomes, P. Da Cunha, M. Da Silva, "Digital transformation capability maturity model framework," *Proc. - IEEE 23rd Int. Enterp. Distrib. Object Comput. Conf.*, pp. 51–57, 2019, doi: 10.1109/EDOC.2019.00016.
[9] V. Veldhoven, J. Vanthienen," Designing a Comprehensive Understanding of Digital Transformation and its Impact". *Bled eConference*. 2019.
[10] D. Sen, M. Ozturk, O. Vayvay, "An Overview of Big Data for Growth in SMEs," *Procedia - Soc. Behav. Sci.*, vol. 235, pp. 159–167, 2016, doi: 10.1016/j.sbspro.2016.11.011.
[11] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. Gutierrez De Mesa, J. Gomez-Pulido, "Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings," IEEE Access, vol. 8, pp. 16111-16126, 2020, doi: 10.1109/ACCESS.2020.2966545.
[12] A. Singh, T. Hess, "How chief digital officers promote the digital transformation of their companies," *MIS Q. Exec.*, vol. 16, no. 1, pp. 1–17, 2017, doi: 10.4324/9780429286797-9.
[13] A. Margiono, "Digital transformation: setting the pace," *J. Bus. Strategy*, vol. 42, no. 5, pp. 315–322, 2020, doi: 10.1108/JBS-11-2019-0215.
[14] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1. pp. 65–75, 1986, doi: 10.1016/S0020-7373(86)80040-2.
[15] J. Aguilar, "A fuzzy cognitive map based on the random neural model," *Lect. Notes in Comput. Sci.*, vol. 2070, pp. 333–338, 2001, doi: 10.1007/3-540-45517-5_37.
[16] G. Nápoles, I. Grau, L. Concepción, L. Koutsoviti Koumeri, J. Papa, "Modeling implicit bias with fuzzy cognitive maps," *Neurocomputing*, vol. 481, pp. 33–45, 2022, doi: 10.1016/j.neucom.2022.01.070.
[17] G. Xirogiannis, M. Glykas, C., Staikouras "Fuzzy Cognitive Maps in Banking Business Process Performance Measurement". In *Fuzzy Cognitive Maps. Studies in Fuzziness and Soft Computing* (Glykas M. ed.), Springer, pp. 161–200, 2010.
[18] H. Sánchez, J. Aguilar, O. Terán, J. Gutiérrez de Mesa, "Modeling the process of shaping the public opinion through Multilevel Fuzzy

Cognitive Maps", *Applied Soft Computing*, vol. 85, 2019, doi: 10.1016/j.asoc.2019.105756.
[19] J. Salmeron, S. Rahimi, A. Navali, A. Sadeghpour "Medical diagnosis of Rheumatoid Arthritis using data driven PSO-FCM with scarce datasets". *Neurocomputing,* pp.:104–112. 2017, doi: 10. 1016/j. neucom. 2016. 09. 113
[20] J. Waissman, R. Sarrate, T. Escobet, J. Aguilar, B. Dahhou, "Wastewater treatment process supervision by means of a fuzzy automaton model", Proc *IEEE International Symposium on Intelligent Control*, pp. 163-168, 2000, doi: 10.1109/ISIC.2000.88291
[21] N. Perozo, J. Aguilar, O. Terán and H. Molina, "A Verification Method for MASOES," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 64-76, 2013, doi: 10.1109/TSMCB.2012.2199106.
[22] J. Aguilar, J. Contreras. "The FCM Designer Tool". In *Fuzzy Cognitive Maps. Studies in Fuzziness and Soft Computing* (Glykas M. ed.), Springer, pp. 71–87, 2010, doi: 10.1007/978-3-642-03220-2_4
[23] J. Aguilar, I. Bessembel, M. Cerrada, F. Hidrobo, F. Narciso, "Una Metodología para el Modelado de Sistemas de Ingeniería Orientado a Agentes", *Revista Iberoamericana de Inteligencia Artificial*, vol. 12, no. 38, pp.39-60, 2008.
[24] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, L. Chamba-Eras, "Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service". In: New Advances in Information Systems and Technologies (Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M., eds.), vol. 444. Springer, pp. 993–1002, 2016
[25] J. Aguilar, "Multilayer Cognitive Maps in the Resolution of Problems using the FCM Designer Tool", *Applied Artificial Intelligence,* vol. 30, no. 7, pp. 720-743, 2016.
[26] G. Csardi, T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Syst.*, no. 1695, 2006.
[27] R. Axelrod *Structure of decision: The cognitive maps of political elites*. Princeton University Press, 1976.

Appendix E. Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps.

Explainability Analysis of the Evaluation Model of the Level of Digital Transformation in MSMEs based on Fuzzy Cognitive Maps

Jairo Fuentes
Universidad Francisco de Paula Santander, Cúcuta, Colombia
GIDITIC, Universidad EAFIT, Medellín, Colombia
fuentesjairo@gmail.com
and
Jose Aguilar
GIDITIC, Universidad EAFIT, Medellín, Colombia
CEMISID, Universidad de Los Andes, Mérida, Venezuela
IMDEA Networks Institute, Madrid, Spain
aguilar@ula.ve, jose.aguilar@imdea.org
and
Edwin Montoya
GIDITIC, Universidad EAFIT, Medellín, Colombia
emontoya@eafit.edu.co
and
William Hoyos
Grupo de Investigación ISI
Universidad Cooperativa de Colombia
Montería, Colombia
GIDITIC, Universidad EAFIT, Medellín, Colombia
william.hoyos@campussucc.edu.co
and
Diego Benito
IMDEA Networks Institute, Madrid, Spain
diego.benito@imdea.org

Abstract — the concept of digital transformation involves exploiting digital technologies to generate new ways of doing things in organizations, including the creation of new processes, models, and services that produce value based on the digitization of data and processes. The application of digital technologies enables organizations to develop capabilities for innovation, automation, etc., utilizing both established and emerging technologies, including the widespread use of artificial intelligence. This article proposes the implementation of Fuzzy Cognitive Maps (FCMs) based on experts and data for the evaluation of the level of digital transformation in MSMEs (Micro, Small and Medium Enterprises). Additionally, this work carries out an explainability analysis of the evaluation models based on FCMs. The main digital transformation variables used to define our FCMs were classified into five groups, based on the COBIT standard: i) Organization and Culture variables related to strategies, way of working, and ecosystems, ii) Customer variables related to services and digital channels and products, iii) Operations and Internal Processes variables related to supply chain, suppliers, and business model, iv) Information Technologies variables related to innovation, digitization, data and analytic. Finally, the fifth type of variable is the target, which indicates the level of digital transformation of the organization. Our model managed to infer the level of digital transformation of the organizations studied. Furthermore, the explanatory capacity of the FCMs developed in this work was explored using different explainability methods, some general and others specific to the FCMs. In general, the results obtained in the work are very encouraging since the quality metrics obtained with the prediction methods are very good, almost always higher than 98%; and the variability in the explainability techniques (very different between them), allow us to analyze in depth the behavior of the variables in the results obtained, something very important to understand how to improve the levels of digital transformation in organizations.

Keywords: Digital Transformation, Fuzzy Cognitive Maps, Explainability Analysis, Machine Learning.

1 Introduction

Digital transformation is the process of integrating digital technologies into the business model, processes, and culture of an organization to improve its strategies, and performance, among other things [1]. Digital transformation is a long-term endeavour that entails determining how technology is employed within an organization to affect its behaviour. Thus, digital transformation involves envisioning how digital technologies can improve service levels, reduce costs, and optimize inventory management, among other things, through the implementation of organizational changes using these technologies to achieve operational excellence. Advanced technologies, such as artificial intelligence, data analytics, and the Internet of Things, are integrated into the digital transformation process to enable organizations to collect and analyse real-time data, automate processes, and improve the supply chain.

On the other hand, FCMs are graph-based structures that support causal reasoning and enable the propagation of causality in both forward and backward directions [2]. They are particularly useful in domains where the concepts and relationships between them are inherently fuzzy, such as politics, military affairs, and historical events. They have been used in different areas where the modelling of the problem is very complex, such as health [3], analysis of social networks [4], or emerging processes [5], [21], among others. FCMs have had a lot of interest in the area of machine learning due to their explainability capabilities. In the field of explainability analysis methods, a classification can be made according to whether the explanation relates to the model and/or its output. This implies that explainability can be global, providing information on the inner workings of the whole model for a data set, or local, focusing on a specific input A and its corresponding output Z. In general, many explainability analysis methods are currently being developed, but one of the challenges is knowing which ones are useful for each machine learning technique, such as which ones could be used in FCMs.

1.1 Related Works

At the level of FCM in classification and prediction tasks, the approach described in [3] combines different indicators used in the standard diagnosis of dengue, such as clinical symptoms, laboratory tests, and physical findings, to determine the severity of dengue. The model achieved an accuracy of 89.4% in diagnosing dengue, indicating performance very close to other well-established machine learning models. Also, to study the vulnerability factors to climate variability in Andean rural micro-watersheds, an FCM was employed in [4], the results of the FCM application did not present an important difference between the two climatic periods to which it was applied. Another work based on an FCM, and a fuzzy linguistic methodology, called VIKOR, determines the vendor selection in Industry 4.0 [5]. VIKOR enables the ranking and selection of suitable suppliers for specific tasks.

The investigation carried out in [6] analyzes the effect of digital transformation on the performance of MSMEs in terms of competitive advantage. The results of this research highlight the positive influence of digital transformation on competitive advantage and the overall performance of MSMEs. In [7], the goal is to develop a holistic maturity model for digital transformation using a theory-based approach, known as DX-CMM. This model aims to help organizations assess their current level of capability and maturity in digital transformation, identify gaps, and create a standardized roadmap for enhancing their digital transformation efforts. According to [6], technological advancements and digital information are expected to be utilized by traditional entrepreneurs. However, this study concluded that all proposed hypotheses were acceptable, and that competitive advantage partially mediated the relationship between digital transformation and MSMEs' performance. In [8], the authors propose a staged digital transformation capability maturity model framework that enables organizations to assess their present digital capability and establish a plan of improvements to move to a higher level. In [7] conducted a systematic literature review and found that none of the 18 existing maturity models in the field of digital transformation met all the criteria of appropriateness, completeness, clarity, and objectivity.

With respect to explainability Analysis for FCMs, the interpretability of FCMs has been confined to the fact that both the concepts and the weights have a well-defined meaning for the problem being modeled. This rather naïve assumption oversimplifies the complexity behind an FCM-based model. The work of Apostolopoulos et al. [35] studies the explainability properties of FCMs and presents critical examples from the literature that demonstrate their superiority in explainability tasks in various domains such as medical precision agriculture, decision support systems, and public policy formulation. Nápoles et al. [36] propose a post-hoc explanation method to estimate the relevance of concepts in FCMs. The proposal considers the dynamic properties of the FCM and uses the values of SHApley additive explanations (SHAP) for scenario analysis. To improve the interpretability of FCMs, in [37] the authors use Liang-Kleeman information flow (L-K IF) analysis, which identifies actual causal relationships from the data using an automatic causal search algorithm. In [38] they propose a symbolic explanation module that allows extracting information and patterns from the FCM-based model, which can be seen as an inverse symbolic reasoning rule that infers the inputs that must be provided to the model to obtain the desired result. These relationships are then imposed as constraints on the learning process to rule out spurious correlations in the data to improve the aggregate predictive and explanatory power of the model. Mansouri et al. [39] propose using an FCM to develop simplified auxiliary models that can provide greater transparency to an LSTM model built to predict industrial bearing failures based on vibration sensor readings. The FCM mimics the performance of the baseline LSTM model and provides more information about the black-box model, such as (i) what variables contribute to the prediction result and (ii) what values could be controlled to avoid potential failures. Thus, the FCM is used to simplify the deep learning model and offer greater explainability. Finally, the article [40] proposes a temporal FCM (tFCM) model, which combines the predictive power of deep neural networks with the interpretability of FCMs. In the proposed tFCM model, cognitive states are modeled as fuzzy, multidimensional, and interrelated vectors, which are input into a short-term memory network for prediction. This hybrid model combines the ability of deep neural networks to discover latent factors with the ability of FCMs to reveal potential correlations.

According to the literature review, there is no digital transformation approach that allows companies to evaluate their existing digital competencies and create a roadmap to improve and reach a higher level. There are also no applications based on FCMs to analyze digital transformation processes in organizations, or that develop an explainability analysis process to understand the factors that may impact their organizational digital transformation. Therefore, there is a lack of research and products available for maturity models related to digital transformation.

1.2 Contributions

The main objective of this research is to use FCM to evaluate the levels of digital transformation in MSMEs and explain possible actions to be taken to improve these levels. In this work, the causal relationships between digital transformation concepts are defined in two different ways, using data or according to expert opinions. The FCM is responsible for analyzing the factors associated with digital transformation data and providing a diagnosis. The study is based on five groups of concepts issued by the Governance and Management Objectives in COBIT [7]. The concepts are classified into five types and reflect the behavior of the company at all levels such as: i) Organization and Culture, ii) Customer, iii) Operations and Internal Processes, iv) Information Technology, and v) Objective, which indicates the digital transformation level of the organization. Also, the study considers several explainability analysis algorithms in the context of FCMs. The main contributions of this paper are:

- Develop several FCMs to determine the level of digital transformation in MSMEs.
- Carry out an explainability analysis for the FCMs that allows studying their results in the context of digital transformation.

This paper is organized as follows: Section 2 discusses the theoretical framework employed, Section 3 defines our FCM for the evaluation of the digital transformation levels in MSMEs, and Section 4 details the computational experiments conducted. Section 5 compares our explainability work with other papers in digital transformation, and Section 6 presents the conclusions and future works.

2. THEORETICAL BACKGROUND

This section presents a background of Digital Transformation. In addition, it provides the basic concepts of FCMs and explainability analysis in machine learning.

2.1 DIGITAL TRANSFORMATION

The application of digital technologies to generate significant changes in an organization with the goal of improving its operations is what defines digital transformation. [9]. Digital transformation is based on the adoption of digital technologies in an organization to transform its business processes. The interpretation of "digital technology" in the context of digital transformation differs across literature. Some authors encompass new Information Technologies solutions [10], and others emphasize cutting-edge technologies like big data, cloud computing, artificial intelligence, blockchain, and the Internet of Things [11], [20]. Also, some authors include social media analytics, and embedded devices [12].

As described in [13], there are two possible strategies for companies to undertake digital transformation. The first is an offensive approach that involves investments in digital transformation portfolios, while the second is a defensive approach that focuses on cultivating new capabilities within the organization.

COBIT (Control Objectives for Information and Related Technology) is a model for the characterization of business information technologies, promoted by ISACA (https://www.isaca.org) since 1996. The COBIT model for digital transformation classifies the concepts into four types:

a) Organization and Culture: variables related to the culture and management of the organization. The concepts of this variable used in this work are:

- *Strategy*, it refers to having an action plan for digital transformation and an understanding of the competitive environment,
- *Culture and management:* it refers to good practices at work, in the management of organizational processes and in the development of leadership, promoting human talent in its different areas and levels, creating an agile and innovative culture,
- *Way of working:* it refers to the type of face-to-face work and/or teleworking, and the practices and procedures implicit in them,
- *Leadership:* it refers to the managerial activities that organizational leaders have to influence the work team, and
- *Ecosystems:* it refers to the environment, not only internal, but also close external to the organization, which impacts the digital transformation process, to analyze its economic/social/cultural effect.

b) Customer: variables related to services (digital or not). The concepts of this variable used in this work are:

- *Digital channels* are the distribution channels of the different organizational strategic messages sent to interested parties.
- *Products*, services, and/or products offered by the organization as a result of its

job or operation, and

- *Digital services* refer to the use of digital tools to improve the customer experience, based on a deep understanding of the customer and its needs.

c) Operations and Internal Processes: variables related to the organizational flow. The concepts of this variable used in this work are:

- *Supply chains* are the supply chains of goods and services that an organization requires for its operation,
- *Suppliers* refer to having providers of technological solutions, in addition to having advisors who accompany the digital transformation process of the organization, and
- *Business model* refers to the organizational capacity to explore new businesses based on technology, the development of new products that allow new markets to be captured, among other things.

d) Information Technology: variables related to the organization technology. The concepts of this variable used in this work are:

- *Innovation* refers to the business strategy of organizational transformation and human resource management based on reengineering, optimization, and process improvement,
- *Digitization* is the process of transforming analog or intangible processes, and physical or tangible elements, into digital ones,
- *Data and analytics* refer to the level of access to data that an organization has, the potential to monetize it by learning about its different uses, among other things, and
- Utilization of *technological solutions* involves implementing methods, systems, or tools that aid in the execution of organizational tasks with greater efficiency, as well as automating certain processes.

2.2 FUNDAMENTALS OF FCM.

In 1986, Kosko developed FCMs [14] from the cognitive maps of Axelrod's work in 1976 [27]. FCMs are a popular way of modeling complex systems because of their straightforward construction and interpretation. They consist of a network of concepts and their interrelationships [14], [15], [18].

2.2.1 BASIC DEFINITIONS OF AN FCM

In Figure 1, an FCM is shown, which consists of eight concepts (C1 to C8). Each concept denotes a specific variable of the system to study. The impact of one concept on another is displayed by the weight in the directed edges among them [16]. Different strategies can be employed to establish these relationships within FCMs. In [18], [22], [25] Aguilar et al. describe three ways to define the relationships in FCMs: using fuzzy rules, generic logical rules that describe the causal relationships, or mathematical models to describe the system under evaluation.

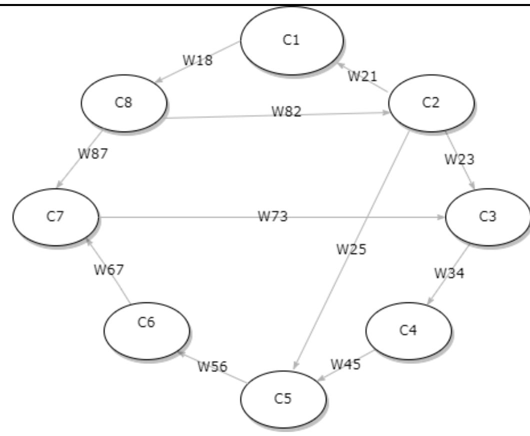

Figure 1. Example of an FCM.

The weight assignment in FCMs can be carried out using different methods, including those involving expert opinion or learning processes [16], [18], [22]. The weights assigned represent the influence of one concept on another, with values ranging from -1 to +1 or 0 to 1, indicating either activating or inhibiting impacts [2]. Thus, different levels of causality are represented. An example of degrees of causality using linguistic terms can be found in Table 1. For example, if the value of the antecedent concept Ci is significantly low, and the value of the consequent concept Cj is substantially high, then the causal link between the two concepts can be categorized as negatively high. Mathematically, an FCM is represented as a tuple of four elements:

$$\Phi = \langle n, f(\cdot), r, W \rangle. \qquad (1)$$

where $C \in \mathbb{R}^m$ is the set of m concepts (C_1, ..., C_m), f (·) is the activation function that holds concept values in a determined range r, and $W \in \mathbb{R}^{m \times m}$ is the adjacency matrix employed to capture the connections among the concepts.

Table 1. Example of generic logic rules used to classify the causal relationships based on the values of concepts.

Linguistic term	Numerical value	Generic logic rule	
		Antecedent C_i	Consequent C_i
Positive complete	1	Very high	Very high
Positive high	0.75	High	High
Positive moderate	0.5	Medium	Medium
Positive low	0.25	Low	Low
Null	0	NA	NA

There are various activation functions that can be used to model the activation of each concept in an FCM (see Table 2). The selection of a specific activation function depends on the specific problem being addressed. For example, if the goal is to model disease symptoms, the sigmoid function may be more appropriate than the hyperbolic tangent function since the values of the concepts being modeled will typically range from 0 (absence of the symptom) to 1 (presence of the symptom), making negative values unnecessary [18], [19].

Table 2. Activation functions used in FCMs.

Activation function	Equation	Range
Bivalent	$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$	$f(x) \in \{0, 1\}$
Trivalent	$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$	$f(x) \in \{-1, 0, 1\}$
Sigmoid	$f(x) = \frac{1}{1+e^{-\lambda x}}$	$f(x) \in [0, 1]$
Hyperbolic tangent	$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$	$f(x) \in [-1, 1]$

W is used to store the relationships between the concepts in an FCM. For example, see the adjacency matrix for the FCM of Fig. 1 in Fig. 2.

$$
\mathbf{W} = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \\ C_8 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 \end{array} \\
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & W_{18} \\
W_{21} & 0 & W_{23} & 0 & W_{25} & 0 & 0 & 0 \\
0 & 0 & 0 & W_{34} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & W_{54} & 0 & W_{56} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & W_{67} & 0 \\
0 & 0 & W_{73} & 0 & 0 & 0 & 0 & 0 \\
0 & W_{82} & 0 & 0 & 0 & 0 & W_{87} & 0
\end{bmatrix}
\end{array}
$$

Figure 2. Example of Adjacency Matrix.

2.2.2 Reasoning in FCMs

To mathematically define the inference process in an FCM, three key components are required: the weight matrix W, the activation function, and the current state vector of the FCM ($a \in \mathbb{R}^m$) which indicates the current degree of activation of each concept. The activation degree of a concept determines if the concept is activated or stimulated in a particular iteration. Thus, in the case of the range *[0, 1]*, a concept is "activated" when its value is *0* in one iteration (t) but becomes greater than *0* in the next iteration *(t + 1)*.

To infer the output in an FCM, it is process is necessary to define a process that involves calculating the state vector through successive iterations of multiplying it with the weights matrix until the system reaches a steady state, as is defined by Equation 2 [14]:

$$a_j(t+1) = f\left(\sum_{i=1, i \neg j}^{m} W_{ij} a_i(t) \right) \qquad (2)$$

where $a_j\ (t + 1)$ is the value of concept C_j at iteration t + 1, m is the number of concepts, W_{ij} is the value for the relationship from concept C_i to concept C_j, and $a_j(t)$ is the value of concept C_j at iteration t. The point of equilibrium (steady state) is reached when $a_j(t) = a_j(t-1)$ or $a_j(t) = a_j(t-1) < 0.001$. Thus, the system is considered to have reached a steady state when the difference between the state vectors in consecutive iterations is less than or equal to *0.001*.

This is the base of the reasoning mechanism of an FCM. FCMs are a crucial tool in simulation scenarios because they enable experts to analyze the system's performance from various starting points. An initial condition is defined by the set of concepts in the iteration (*0*) and is denoted as:

$$a0)=[a_1(0), a_2(0), \ldots., a_m(0)] \quad (3)$$

where $a_1(0)$ is the value of concept C_1 at iteration = *0*.

2.3 Explicability in FCM

In the current landscape, the ability to provide explanations for results obtained using machine learning techniques is crucial, especially in critical sectors. Literature has developed various explainability methods [28]. Some of these methods are Local Interpretable Model-agnostic Explications (LIME) and feature importance, and the Weight-Filter FCM method proposed in [34]. In this work, we propose an explainability analysis based on these methods.

2.3.1 Explicability Analysis based on Weight-Filter FCM

Much of the existing research on FCM focuses mainly on the constructions of the models and examining the causal relationships between the concepts. There is a notable scarcity of work that delves into a more comprehensive analysis of explainability in FCM. Moreover, the limited studies that do exist often concentrate on employing local sensitivity to quantify the impact of slight perturbations, or changes in input values and weights on the model's predictive capacity.

In the article [34], the authors present a different method. The fundamental concepts of this approach are clarified below. The evaluation of sensitivity between concepts is carried out by eliminating low-impact causal connections between input concepts and target concepts. Thus, this approach aims to examine the causal impact exerted by less impactful concepts on the target concepts to which they are connected. The goal is to determine if this influence significantly affects the performance of the model. To achieve this, the least impactful causal connections (with fewer weights) are eliminated, establishing weight thresholds for this decision. Thus, for a given threshold value *u*, each causal relationship is characterized by a weight w_{it} where C_i represents an input concept and C_t is a target concept, if the weight w_{it} falls below the specified threshold *u*, then this causal relationship is eliminated in the model. Once these concepts have been eliminated, the quality of the FCM model with that new configuration is evaluated, and if the performance results do not degrade with that new model, then it is definitively eliminated.

Thus, if notable variations in results emerge between the initial model and the adjusted model, then it can be inferred that the removal of less impactful causal relations influences the performance of the model.

2.3.2 Explicability analysis based on LIME

LIME, introduced by Ribeiro et al. [32], is a method designed to elucidate individual predictions made by a model through the construction of local surrogate models for each specific prediction. The surrogate models are formulated by generating a set of instances in proximity to the target instance and assigning labels to them using the original machine learning model. Subsequently, these instances and labels are employed to train the surrogate model, which tends to be more interpretable. The primary objective of LIME is to provide an explanation for the prediction of a given instance. For each instance requiring an explanation, LIME generates a surrogate model. Given a specific instance, LIME investigates the impact on the prediction when variations occur within the variables of that instance. These variations constitute the neighborhood of the instance. Equation 4 illustrates the bases of LIME to obtain an explanation for a data point x.

$$\varepsilon(X) = argmin_g \,(L\,(f,g,\pi_x) + \theta(g)) \qquad (4)$$

Where g is the family of possible explanations, L is the loss function and it measures how close g (surrogate model) is to the prediction of f (original model) in its vicinity π_x, and θ(g) measures the complexity of the surrogate model

2.3.3 Explicability analysis based on Feature importance

The Feature Importance is a method that calculates the relative importance of each descriptor/feature, given a dataset, in the building of a model. The feature importance method is useful in several contexts such as Feature selection, Model interpretability, Model debugging, and Business decision-making, among others. In this work, we will use it to evaluate the model interpretability. The method provides several ways to classify features according to their contributions to the inference of the target variable. It is possible to use different strategies to analyze the importance of the features, such as the correlation with the target variable, the quality of the model according to each variable, among others. In this work, the sensitivity of the results to the permutation of their values will be used due to its simplicity, easy understanding and low computational cost. The conceptual bases of this method (permutation criterion) are [33]:

Let $X = [x_1 . x_2, \dots x_j, \dots x_m]$ be an individual with m features, where x_j is the descriptor j of the individual X. Let $C = [c1_, c2_, \dots ck_, \dots . c_p]$ be an array that contains the labels of the classes that any individual X belongs. In this sense, a possible structure of the dataset is shown in Figure 3.

$$\begin{pmatrix} x_1^1(1) & x_2^1(1) & \cdots & x_j^1(1) & \dots & x_m^1(1) & c_1 \\ x_1^2(2) & x_2^2(2) & \dots & x_j^2(2) & \dots & x_m^2(2) & c_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_j^i(k) & \dots & \dots & x_j^i(k) & \dots & x_m^2(k) & c_k \\ x_j^i(p) & \dots & \cdots & x_n^j(p) & \dots & x_n^i(p) & c_p \end{pmatrix}$$

Figure 3. Structure of the dataset.

Where $x_j^n(p)$ represents the descriptor j of the individual n in the class p. The permutation criterion algorithm consists of modifying the values of descriptor j, leaving the rest of the values fixed, to determine the performance metric of the model. Thus, each time the model is trained for a new value of the analysed characteristic (e.g., descriptor *j*) its performance metrics are calculated. This is done repetitively to analyse the sensitivity of the model to the variable.

From the above, the sensitivity of the model to each variable is obtained, such that if the results change a lot, then the model is very sensitive, otherwise, it is not very sensitive. This allows establishing a ranking between the variables, with the most sensitive being the most highly ranked.

3 OUR FCM TO EVALUATE THE LEVEL OF DIGITAL TRANSFORMATION IN COMPANIES.

This section details the process of constructing our evaluation model about the level of digital transformation in organizations utilizing FCMs. Our FCMs are based on experts and data to define the relationships. The construction of the FCMs is based on a methodology that consists of six steps [3], [23], [24]. The rest of this section briefly describes these steps applied to our FCM.

3.1 SELECTION OF EXPERTS AND PREPARATION OF THE DATASET

For the construction of the FCM based on experts, this study included the participation of five experts with more than five years of experience in digital transformation. They have experiences in the fields of management of digital transformation processes in different companies and government entities, in the study of the COBIT Model, in the field of business models and Digital Economy, Government Digitization, among other areas.

On the other hand, the dataset used to build the FCM is from the National Administrative Department of Statistics -DANE-, a Colombian entity responsible for the planning, collection, processing, analysis, and dissemination of statistics of companies and official organizations, at the national and international level [41]. The original dataset has 635 variables and corresponds to data from Colombian MSMEs. We selected from the dataset the 15 variables that the experts defined that are like the concepts of the COBIT model, so they are in turn relevant to characterize the Colombian MSMEs. These variables from the dataset are organizational methods, agile culture, collaborative work, work modality, work team, digital utility, products offered, distribution routes, supply chain, procedure, business pattern, analytics, novelty, digitalization and technologies. Finally, records with more than 80% missing values, for variables, are dropped. Additionally, we used the Synthetic Minority Oversampling Technique (SMOTE) to balance the classes [31].

3.2 CONCEPTS AND RELATIONSHIPS

In the case of the FCM based on experts, the initial phase of this process is carried out by experts in digital transformation. Based on the established concepts of the COBIT model (see Table 3), the experts constructed W with values ranging from 0 to 1 to indicate the relationships between the concepts. Table 3 shows an additional concept to those used in the

COBIT model, which is the concept of the digital transformation level. This concept is the concept that is inferred using the FCM. Using Table 1 as a guide, the experts selected the linguistic term that best represented the relationship between each concept, and then assigned numerical values to these terms to create the weight matrix.

Table 3. FCM Concepts for Digital Transformation.

Node	**Name Concept**
C1	Strategy
C2	Culture and Management
C3	Way of working
C4	Leadership
C5	Ecosystems
C6	Digital Services
C7	Products
C8	Digital Channels
C9	Suppliers and Supply Chain
C10	Business Model
C11	Processes
C12	Innovation
C13	Digitalization
C14	Technological Solutions
C15	Data Analytics
End Node	**Digital Transformation**

In the case of the learning approach (FCM based on data), we have used the FCM expert tool [2] to find the best FCM, which is based on PSO.

3.3 FCM DESIGN

In the FCMs developed by the experts in MSMEs, their knowledge is mixed into a single overall map to define the FCM. The process for producing a consolidated global map is described by the equation below [21], [22]:

$$E_{ij}^{G} = \sum_{e=1}^{NE} \frac{E_{ij}^{e}}{NE} \qquad (4)$$

This equation calculates the overall weight for the FCM, represented by $E^{G}{}_{ij}$. The opinion of each expert e regarding the causal link between concepts C_i and C_j is represented by $E^{e}{}_{ij}$, and NE indicates the total number of experts who participated. For the construction of this FCM, we also have used the FCM Expert package (version 1.0.0) [2] but in the literature, there are others like [22], [25], [26]. Figure 4 shows the final FCMs.

Also, as we said before, in the case of the FCM based on data, we have used the FCM expert package [2] to train the FCM, which uses a PSO for the training. Specifically, the dataset is divided into a part (80%) for the training of the FCM and the rest for the test phase.

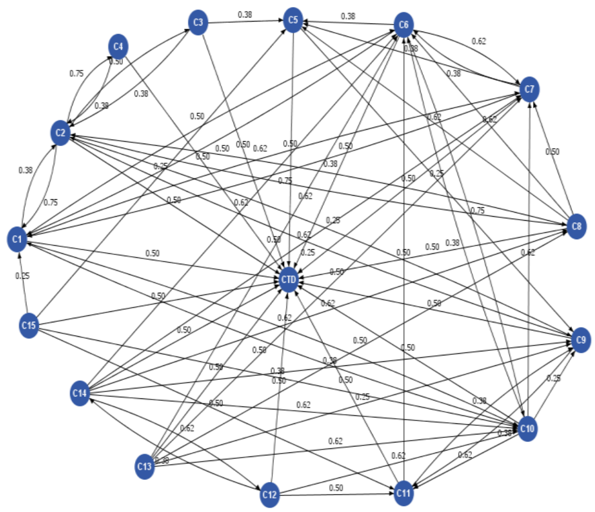

Figure 4. General FCM for evaluating the level of digital transformation in companies.

3.4 Inference

The fourth step involves the utilization of the FCMs defined in this study to analyze different scenarios (for example, in our case, the level of digital transformation of different organizations) [18], [21]. In this step, the inference process is carried out for each case study, following the reasoning process of the FCMs. The experiment design and the results of this stage are shown in section 4.

3.5 Interpretation

The interpretation of FCM inference results for each case study is carried out in this step and is detailed in section 4. In our case, we are going to use four explainability analysis methods, which are LIME, feature importance, Weight-Filter FCM, and an explainability analysis based on Classic FCM. In the latter case, since the way in which FCM establishes influence relationships between the input concepts and the target concepts is through the use of weights assigned to the causal relationships, they are used to identify the most influential characteristics in the model decision-making process.

3.6 Decision

The last stage is the decision that the user of the FCMs can make to improve the digital transformation process in the organization based on their results. Since FCMs are now extended with explainability analysis methods, this new information can be used in the decision-making process to identify exactly what to do.

4 EXPERIMENTS

This section aims to evaluate the efficacy of our FCMs in different scenarios to categorize the levels of digital transformation in multiple organizations. Two organizations are selected as case studies for this purpose and two FCMs are defined, one based on experts and another on data. Initially, the COBIT concepts are subject to an evaluation within each organization, to generate the input of the FCM inference process to ascertain the respective digital transformation levels.

4.1 COMPARISON OF THE RESULTS

4.1.1 USING THE TEST DATASET.

The results obtained from the application of the data-driven FCM on the dataset described above correspond to a precision of 1, an accuracy of 1, a recall of 1 and an F1 score of 1, which indicates an excellent performance for the prediction of the concept that represents the digital transformation. The hyperparameter configuration that allowed us to achieve this result consists of an initial population size of 70 individuals, the sigmoid function as the activation function, and the modified Kosko function as the reasoning inference function. In the case of the FCM based on experts, the results are an accuracy of 0.9835, precision of 0.98, recall of 1 and F1-Score of 0.99, which indicates a lower performance for the prediction of the concept that represents the level of digital transformation (see Table 4).

Table 4. Quality of the proposed FCMs.

	accuracy	**precision**	**recall**	**F1-Score**
Data-driven FCM	1	1	1	1
Expert-based FCM	0.9835	0.98	1	0.99

4.1.2 CASE 2. FOR THE CHAMBER OF COMMERCE.

The chamber of commerce acts as a collective body, bringing entrepreneurs and business owners together. In an effort to streamline business operations, Chambers of Commerce are broadening their range of offerings to encompass virtual services, allowing users to conduct their tasks with enhanced speed and convenience.

Table 5 shows the input values for concepts, where 0 is the absence of value in that concept and 1 is the maximum presence of that concept. In this organization, the current values of each concept of the COBIT model were determined as the average of what the company's experts thought. The final values are mentioned below: the concept of Strategy (C1) with a value of 0.7, Culture and Management (C2) of 0.7, Way of Working (C3) of 0.5, Leadership (C4) of 0.3, Ecosystems (C5) of 0.5, Digital Services (C6) of 0.7, Products (C7) of 0.6, Digital Channels (C8) of 0.8, Suppliers and Supply Chain (C9) of 0.4, Model of Business (C10) of 0.6, Processes (C11) of 0.4, Innovation (C12) of 0.8, Digitalization (C13) of 0.8, Technological Solutions (C14) of 0.9 and Data Analytics (C15) of 0.4. With the weights averaged by the experts, we proceeded to infer the results in each FCM (see Table 6).

Table 5. Input values for the case study 2.

C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**

0,7	0,7	0,5	0,3	0,3	0,7	0,6	0,8	0,4	0,6	0,4	0,8	0,8	0,9	0,4

Table 6 shows the results of the inference processes. In the final vector, the digital transformation concept (CTD) reached a value of 0.9954 and 0.9979, respectively.

Table 6. Output vector for the case study 2.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0,936	0,8878	0,7417	0,7754	0,889	0,9532	0,8539	0,8336	0,9502	0,9502	0,8832	0,741	0,638	0,7129	0,659	0,9954
Data driven FCM	0,9854	0,9626	0,9514	0,916	0,9691	0,9863	0,9793	0,9829	0,9761	0,9939	0,9778	0,9834	0,9666	0,971	0,977	0,9979

4.1.3 Case 3. Galavis Coffee Company.

Galavis Coffe is a north Santander Company, founded in 1918, dedicated to the roasting and distribution of Colombian Coffee. Table 7 shows the input values for the concepts of the COBIT model of digital transformation. These values describe the current state of these concepts in the organization according to the company's experts. With this input, the results of the inference process in each FCM is defined in Table 8.

Table 7. Input table for the case study 3.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0,5	0,4	0,3	0,5	0,3	0,4	0,4	0,4	0,3	0,4	0,5	0,3	0,3	0,3	0,3

Thus, Table 8 shows the final vector of concepts. In this case, the CTD decision concept reached a value of 0.9954 and 0.9978, respectively.

Table 8. Output vector for the case study 3.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0,936	0,8878	0,7417	0,7754	0,889	0,9532	0,8539	0,9636	0,9502	0,9502	0,8832	0,741	0,638	0,7129	0,659	0,9954
Data driven FCM	0,9821	0,9589	0,9501	0,9002	0,9622	0,9791	0,9703	0,9754	0,9712	0,9905	0,9708	0,9788	0,9616	0,9699	0,971	0,9978

The two FCMs reach a very close CTD value. The inference process (causal relationship) of both models in this case leads to very close values even though the behavior of the other concepts is not the same in each model (for the same input). For example, in the expert-based FCM, the relevant variables at the end are C1, C6, C8, C9, and C10, while in the Data-driven FCM are all.

4.2 Explicability Analysis

For the explainability analysis, LIME, Feature Importance and classic FCM methods were used for the data-based FCM. For the expert-based FCM, the weight filter FCM was used. In the case of expert-based FCM, the classic FCM method cannot be used since the weights that reached the target variable were all the same. Additionally, the LIME and

Feature Importance methods require data in their analysis to update the FCM model after a training phase, which is not possible in expert-based FCM. In the case of the data-based FCM, the weight filter FCM method was not used because in this case, the model depends on all the variables with a certain relevance, in such a way that when applying any type of filter to eliminate some of them, the model no longer behaved adequately.

4.2.1 Explicability Analysis based on Weight-Filter FCM for expert-based FCM

Table 9 shows the adjacency matrix given by the average of the weights given by the experts in each of the relationships between the concepts. In this case, the most important characteristic is "C14" with an influence of 0 .44, followed by "C13" with 0.40.

Table 9. Adjacency Matrix Average weights given by Experts.

		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD	Average
C1	Strategy	0	0,312		0,625		0,375	0,25	0,25		0,625	0,125	0,375	0,25		0,562	0,5	0,341
C2	Culture and Management	0,125	0	0,875	0,625	0,25			0,438	0,18	0,25	0,375	0,375		0,25	0,125	0,5	0,322
C3	Way of working	0,25	0,75	0	0,375	0,125	0,375	0,25	0,125		0,25	0,375	0,25	0,4375	0,25		0,5	0,293
C4	Leadership	0,5	0,5	0,25	0	0,125					0,25	0,1875	0,25			0,375	0,5	0,271
C5	Ecosystems	0,25				0	0,375		0,625	0,375	0,125	0,25	0,375	0,125	0,125	0,125	0,5	0,250
C6	Digital Services	0,5		0,25		0,625	0	0,5	0,375	0,25	0,75		0,375	0,25	0,25	0,25	0,5	0,365
C7	Products	0,375		0,25		0,25	0,375	0	0,437	0,375	0,125		0,25	0,437	0,25	0,25	0,5	0,281
C8	Digital Channels	0,25	0,625	0,25		0,75	0,375	0,625	0		0,375		0,375	0,437	0,375	0,187	0,5	0,385
C9	Suppliers and Supply Chain		0,375	0,25				0,25		0	0,125	0,75	0,125	0,375	0,437		0,5	0,299
C10	Business model	0,625	0,25	0,25	0,25		0,625	0,375	0,25	0,5	0	0,625	0,625	0,125	0,25	0,375	0,5	0,366
C11	Processes	0,25	0,25	0,25		0,25	0,625	0,25	0,25	0,875	0,25	0	0,25	0,187	0,375	0,375	0,5	0,317
C12	Innovation	0,25	0,25			0,25	0,5		0,25	0,25	0,5	0,625	0		0,375	0,25	0,5	0,318
C13	Digitization					0,25	0,625	0,375	0,375	0,375	0,75		0,25	0	0,375	0,687	0,5	0,406
C14	Technological solutions			0,25		0,25	0,5	1	0,75	0,75	0,5		0,25	0,25	0	0,375	0,5	0,443
C15	Data and Analytics	0,75			0,25	0,375		0,25	0,25		0,5	0,75	0,25	0,5	0,25	0	0,5	0,375

The thresholds have been defined considering the distribution of the weights of the causal relationships given by the experts. This analysis allows us to evaluate the influence of concepts with lower weights by observing whether the behavior of the model changes or remains the same. Table 10 summarizes the filters selected. Each filter is more restrictive than the previous one. For example, Filter 1 is responsible for removing all causal relationships associated with the target concepts (classes) whose weight is less than or equal to 0.125. Filter 2 suppresses causal relationships with weights less than 0.25, Filter 3 eliminates all causal relationships with weights less than 0.50, while Filter 4 suppresses causal relationships with weights less than 0.75.

Table 10. Filters based on the average of the weights of the causal relationships with respect to the target variable.

Filter 1	<= 0.125
Filter 2	< 0.25

Filter 3	< 0.50
Filter 4	< 0.75

For filters 1, 2, and 3, the results of the inference process have a small variation between the relationships of the concepts (see the value of CTD in Table 11 for filter 1). This tells us that we cannot clearly extract a feature that is more relevant than the others.

Table 11. Results with filter 1.

Inference results

Plotter | Table

Step	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	0.8741	0.8081	0.7186	0.7549	0.7663	0.8741	0.8267	0.7663	0.8267	0.9284	0.867	0.6926	0.6792	0.6225	0.7548	0.9859
3	0.9635	0.9006	0.8062	0.8589	0.8684	0.956	0.9141	0.8471	0.9253	0.9843	0.9555	0.7812	0.742	0.6508	0.8472	0.999
4	0.9735	0.9216	0.8312	0.8833	0.891	0.966	0.9291	0.8674	0.9394	0.9889	0.9673	0.8016	0.7624	0.6572	0.8697	0.9994
5	0.9748	0.9261	0.8372	0.8877	0.8952	0.9675	0.9316	0.8718	0.9412	0.9896	0.969	0.8053	0.768	0.6586	0.8744	0.9995
6	0.975	0.9271	0.8386	0.8885	0.896	0.9678	0.9321	0.8727	0.9415	0.9897	0.9693	0.8059	0.7695	0.659	0.8754	0.9995
7	0.9751	0.9273	0.8389	0.8887	0.8962	0.9679	0.9322	0.8729	0.9415	0.9897	0.9694	0.806	0.7698	0.659	0.8757	0.9995

For filter 4, when performing the inference, it results in a significant variation in the results of CTD. At the end, the most important feature is “C10” with an influence of 0.91, “C9” with an influence of 0.89, followed by “C11” with an influence of 0.88 (see Table 12). In general, it is observed that the other concepts (“C4”, “C5”, “C6”, “C8”, “C12”, “C13”, “C14” and “C15”) do not have a relevant contribution to the prediction.

Table 12. Result of weights with filter 4.

Inference results

Plotter | Table

Step	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
2	0.7549	0.7058	0.7186	0.6225	0.7058	0.6225	0.7311	0.7058	0.7879	0.8176	0.7773	0.6225	0.6225	0.6225	0.6225	0.6225
3	0.8319	0.7764	0.7919	0.6508	0.7747	0.6508	0.7947	0.7636	0.8738	0.8966	0.8624	0.6508	0.6508	0.6508	0.6508	0.6508
4	0.8502	0.7974	0.8132	0.6572	0.7937	0.6572	0.8093	0.7776	0.8925	0.9106	0.8814	0.6572	0.6572	0.6572	0.6572	0.6572
5	0.8542	0.8033	0.8192	0.6586	0.7985	0.6586	0.8125	0.7808	0.8963	0.9131	0.8853	0.6586	0.6586	0.6586	0.6586	0.6586
6	0.8551	0.805	0.8209	0.659	0.7997	0.659	0.8132	0.7816	0.8971	0.9135	0.8861	0.659	0.659	0.659	0.659	0.659
7	0.8553	0.8054	0.8213	0.659	0.7999	0.659	0.8134	0.7817	0.8972	0.9136	0.8863	0.659	0.659	0.659	0.659	0.659

4.2.2 Explicability Analysis based on Classic FCM for data-based FCM

According to the FCM model generated with the dataset (see Figure 5), the most important feature is "C1" with an influence of 0.52, followed by "C5' 'with 0.44. "C12" and "C15" both have an influence of 0.42 and 0.40, respectively. Finally, "C8" has an influence of 0.36. We see that there are no common relevant concepts with respect to those defined in section 4.2.1.

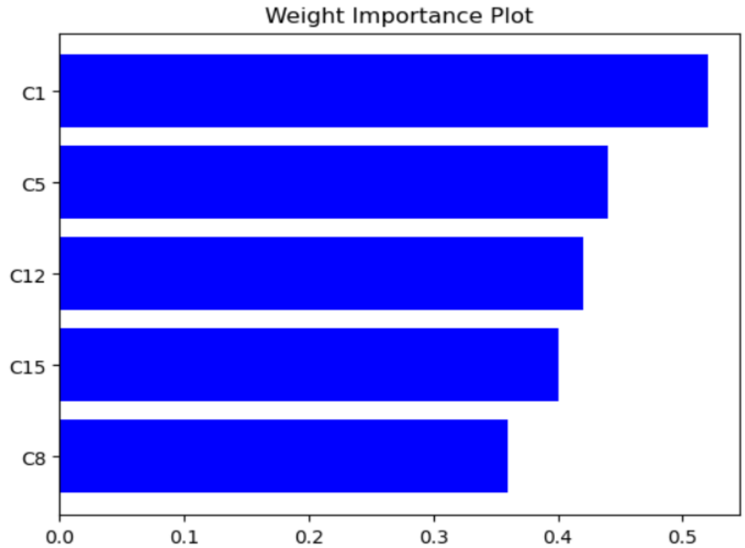

Figure 5. Weight importance Plot.

It is good to highlight that both explainability methods for FCM give different results in the explainability analysis. Now, the reason is that one is based on eliminating the concepts with less weight in their relationship with CTD (section 4.2.1) such that it does not affect the quality of the inferences, while the one in this section is left the concepts with greater weight in the relationship with CTD. Thus, opposite processes follow, but despite that, they agree on some relevant concepts, for example, C1 and C15.

4.2.3 Explicability analysis based on LIME for data-based FCM

Next, we present an example of the analysis that can be done for each class in each dataset, to study the behavior of each variable in a given model. Figure 6 illustrates the calculation of the LIME values for the prediction of the class 1.

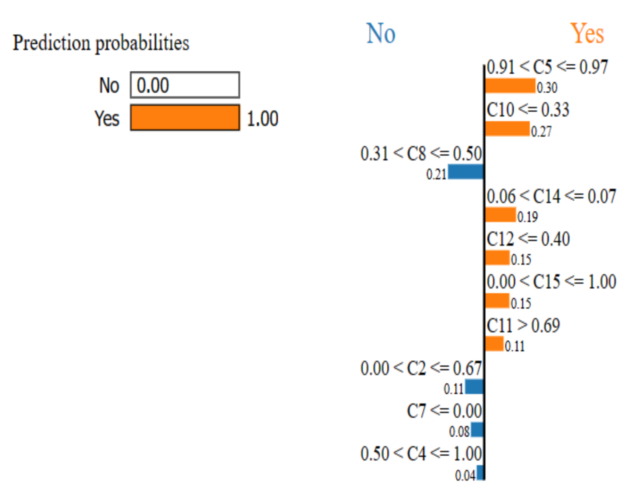

Figure 6. Feature Importance based on LIME for class 1 for data-based FCM.

```
[('0.91 < C5 <= 0.97', 0.3001793956564),
 ('C10 <= 0.33', 0.26649081469982155),
 ('0.31 < C8 <= 0.50', -0.2120230535399748),
 ('0.06 < C14 <= 0.07', 0.1858697854062809),
 ('C12 <= 0.40', 0.1486919740988927),
 ('0.00 < C15 <= 1.00', 0.14711802758805537),
 ('C11 > 0.69', 0.11321779302762842),
 ('0.00 < C2 <= 0.67', -0.11210549710522849),
 ('C7 <= 0.00', -0.07796034786804207),
 ('0.50 < C4 <= 1.00', -0.04413467707423912)]
```

Figure 7. LIME for class 1 of the data-based FCM

To analyze the results presented in Figure 6, the following is taken into consideration:

- The order of the features reflects their weight in the model’s prediction. This shows a ranking of which features are more important for obtaining that prediction. In this case, the variable “C5” has 30% importance, “C10” with 27%, “C8” with 21%, “C14” with 18%. and “C12” has 15%. These are the most influential features.
- The colour of each feature and the x-axis indicate how it positively or negatively influences the prediction. The variables "C5", "C10" and "C14" (orange color) have a positive influence on the prediction, while the variables "C8", and "C2" (blue color) have a negative influence.
- The value next to each feature indicates the threshold from which it positively or negatively contributes to the prediction of the class. For example, values between 0.91 and 0, 97 for “C5” have a positive impact on the prediction. Also, values within < 0.31 and <= 0.50 of “C8” have a negative contribution to the prediction.
- Figure 7 summarizes all the information shown in Figure 6, showing the list of variables, from which values impact the results, and the degree of importance of each variable.

Based on the information explained above and Figure 7, we can extract the following conclusions in terms of explainability:

- ”C5” is the variable with the highest positive weight in the prediction. Values in the interval (0.91, 0.97] contribute to the prediction of class 1 of the dataset.
- ”C10” is the second most positively weighted feature in the prediction. Values lower than the 0.33 threshold contribute to the prediction of class 1.
- “C8” is the variable with the highest negative weight in the prediction, Values in the interval (0.31,0.50] contribute to a prediction of class 0.
- “C14” and” C15” are the third and fourth features with the highest positive weight in predicting class 1. Values in the intervals (0.06,0.07] and (0,1], respectively, contribute to the prediction of class 1.
- “C2”, “C7” and “C4” have a significant negative impact on the prediction of class 1; however, their influence and weight are lower than the rest of the variables already discussed.
- The rest of the variables have a little significant impact on the prediction.

Based on a similar feature importance analysis for the rest of the classes, we establish the overall importance of the features in each of the datasets. Figure 8 shows the ranking of the most important features for the FCM model.

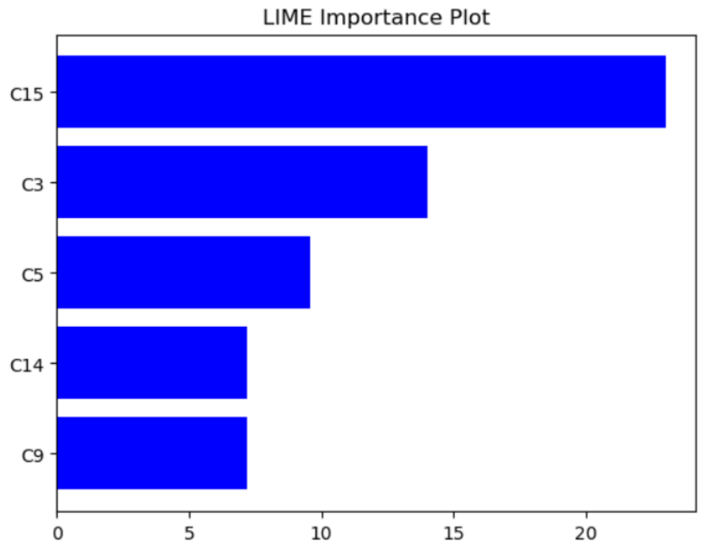

Figure 8. Ranking of the most important features for the FCM model.

It is observed that the "C15" characteristic is the most important for the model, with 23%. It is followed by "C3" with 14%. Next are "C5" with 9.6%, followed by "C9" and "C14" both with 7.2%

4.2.4 Explicability analysis based on Feature Importance for data-based FCM

In this section, we proceed with the explainability analysis using the feature importance method based on permutation. Figure 9 shows the most relevant characteristics according to this analysis.

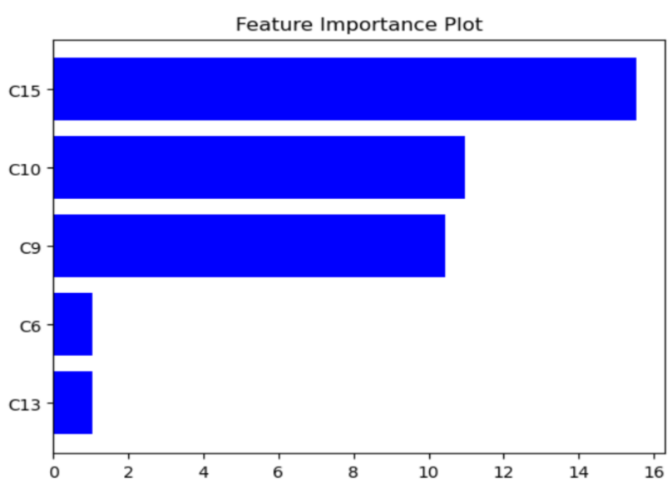

Figure 9. Most relevant characteristics according to the analysis.

In this case, the most important feature is “C15” with a value of 15. It is followed by “C10” with 11, and “C9” with a value of 10. Finally, “C6” with 10, and “C13” with 1.04.

4.2.5 Comparison between Explainability Methods

In Table 13, we see the most important features that each method provides. First, we note that the FCM models give a different ranking of the most relevant characteristics, but even the methods for the same model as well (for example, in the case of FCM based on Data). Now, the only feature that appears in all cases, and even in two with the highest value, is C15. Other features that appear several times are C10 and C5, twice between the first 3, and C3, C9, C6 and C14. In that sense, C15 (Data Analytics) mainly, and C5 (Ecosystems) and C10 (Business Model), are essential according to our analyzes for a high level of digital transformation in an organization.

Table 13. Feature Importance by Method.

Method	First Feature	Second Feature	Third Feature	Others
Expert-based FCM and Weight-Filter FCM	C10	C8	C11	C4, C5, C6, C8, C12, C13, C14 and C15.
Data-based FCM and Classic FCM	C1	C5	C12	C15
Data-based FCM and LIME	C15	C3	C5	C9 and C14
Data-based FCM and Feature Importance	C15	C10	C9	C6 and C13

Explainability methods can produce different results regarding feature importance due to variations in their approaches, assumptions, and limitations. These discrepancies can lead to differences in determining which features are considered most crucial to the model. Numerous authors, including Lundberg and Lee [42], and Ribeiro et al.[43], provide detailed comparisons illustrating how different explainability methods produce variable results. Given these differences in results between explainability methods, an analysis of feature importance performance was performed for each method. The RemOve And Retrain (ROAR) method was used [44]. ROAR is a method used to evaluate the importance of features, in which the subset of the input features considered most critical in a given explainability method is removed, and replaced with a constant average value. By observing how model performance changes when these features are removed and replaced, ROAR provides information about the relative importance of different input features in the model's predictions. A degradation rate is calculated, defined as the difference in performance between the model trained with all features and one trained without a proportion of the features considered most important by each explainability method. In our case, the 3 most important characteristics were considered according to each method, as shown in Table 13. Precision was used as a metric to calculate the degradation rate. The method with the highest degradation rate is considered to suggest the importance of the features that best fit the predictive behavior of the model.

Table 14. ROAR results by method.

Method	Accuracy	ROAR Accuracy	Degradation Rate
Expert-based FCM and Weight-Filter FCM	1	0.6236	0.3764
Data-based FCM and Classic FCM	0.9835	0.4681	0.5154
Data-based FCM and LIME		0.6910	0.2925
Data-based FCM and Feature Importance		0.8357	0.1478

Table 14 shows the degradation indices obtained when evaluating each of the explainability methods. It is noteworthy that *Weight-Filter FCM for Expert-based FCM* and *Classic FCM for Data-driven FCM* have the most important degradations, 0.3764 and 0.5154, respectively. This suggests that these methods provide the strongest explanations in terms of ROAR. This result is consistent since they are methods that arise directly from FCM.

5 COMPARISON WITH OTHER WORKS IN DIGITAL TRANSFORMATION

To compare this study with other works, a set of four criteria were established, which are outlined below:

- Cri1: It uses non-intrusive schemes for the evaluation.
- Cri2: It uses machine learning in the evaluation model.
- Cri3: It explains the reasons for the level deduced of digital transformation.
- Cri4: It uses different types of explainability methods.

Table 15 presents a comparison of this study with previous works based on these criteria that are deemed pertinent in the context of agroindustrial MSMEs and their digital transformation. These criteria are significant as they pertain to the application of automated technologies in agroindustry and the utilization of machine learning to enhance the agroindustrial production process. Using machine learning in the evaluation model allows identifying patterns among the data to make predictions. By explaining the reason for the level of digital transformation inferred will allow organizations to assess their current digital capability and establish an improvement plan to move to a higher level.

Table 15. Comparison with other works.

	Cri1	Cri2	Cri3	Cri4
[3]	√	√	X	X
[4]	X	X	X	X
[5]	X	X	X	X
[6]	X	X	√	X

[7]	X	X	√	X
[8]	X	X	√	X
[9]	X	X	√	X
[29]	X	√	X	√
[30]	X	√	X	√
[31]	√	√	X	√
This work	√	√	√	√

The first criterion is met by the studies [3], [31] and ours since both use computerized methods and models for evaluation. The second criterion is met by [3],[29], [30], [31] and our study, which uses machine learning for different purposes, such as information processing for dengue diagnosis [3]. The third criterion is met by studies [6], [7], [8], [9]. They explain the relationship between digital transformation and MSME performance, conduct digital transformation maturity assessments, and design a comprehensive understanding of digital transformation and its impact.

With respect to the fourth criterion, in [29] was established that the results obtained through FCMs are strongly linked to relationships with high weights. The work [30] addresses the evaluation of the robustness of the FCM using different uncertainty propagation analysis methods. Both studies use the local sensitivity of the concepts as an indicator of the quality of the robustness, measuring the variations in the predictions by changing the weights of the relationships. Within the scope of fuzzy models, [31] presents a rule-based fuzzy explanatory system designed especially for implementation in deep neural networks.

Our study meets all the requirements of an evaluation model to analyze the digital transformation in an organization. Our model is not intrusive because it uses organizational experience to evaluate the current state of certain organizational variables and, from there, uses a computer tool to make the evaluation. It uses the FCM theory as a machine learning technique, and different methods for an explainability analysis. Using our explainability analysis methods proposed in section 4.2, FCM explains the conclusions reached through the causal relationships that exist between the concepts/variables.

6 CONCLUSIONS

The purpose of this study was to develop FCMs to determine the level of digital transformation in MSMEs, and to carry out an explainability analysis process to understand the results returned by the FCMs. A computational tool is proposed to analyze the variables, which are divided into five specific types: i) Organization and Culture, ii) Customer, iii) Operations and Internal Processes, iv) Information Technologies, and v) Objective, which indicates the level of transformation digital in an organization according to the values in the previous variables in this organization.

The FCM allowed analyzing the behavior of factors associated with digital transformation in MSMEs. Our approach is an explicable and interpretive technique that permits the explanation of results based on the diverse concepts in the digital transformation of the COBIT model. This approach is very useful because it allows the detection of the concepts that need to be strengthened or improved in an organization to strengthen its digital transformation process to be more competitive. The developed model is easily customizable and flexible, making it simple to integrate additional concepts and relationships. For its use, only an initial vector of the current state of the variables related to the digital transformation and the weight matrix is required.

In the case studies of agroindustrial MSMEs and the chamber of commerce, our FCM-based tool allowed determining the variables related to digital transformation to improve. The quality of the FCM-based models for predicting the level of digital transformation in an organization was greater than 98% in all quality metrics. Specifically, it was found that the variables associated with digitization, data and analysis, innovation, and technological solutions, were those that should be improved. In turn, our model allowed determining for the MSMEs of the agro-industrial sector analyzed, the variables that should have a high value at the end because they influence the most to have a high level of digital transformation in companies.

Additionally, an explainability analysis was carried out in the different FCM models developed. In the case of expert-based FCM, the most relevant variables are C10, C8 and C11. In the case of the data-based FCM, where more explainability methods were used, the most relevant variables in general given by them were C5 and C15. We see that the explainability results do not coincide between the methods, and of course, even less so when the different FCMs developed are considered. Due to that, we analyzed the quality of the explainability methods using the ROAR method, and the best were *Classic FCM for Data-based FCM* and *Weight-Filter FCM for Expert-based FCM*.

Among the main limitations of this work is the lack of large databases for the learning process of the Fuzzy Cognitive Map (they were tested with those obtained in the literature). Another limitation has to do with the complexity of testing the model in an organization to assess its level of digital transformation due to the lack of methodologies that allow capturing the required information.

Future work should develop new metrics to evaluate the quality of explainability provided by the methods. With respect to their quality evaluation, Table 14 is one example of that comparison based on a method of evaluating their quality, which should be extended in the future. Also, future work should develop methods that allow hybridizing the analysis provided by them. Other future research should focus on developing methodologies to evaluate the levels of digital transformation in organizations using tools like ours. Furthermore, since the dataset used in this study is from a single city in Colombia, distributed strategies, such as federated learning, must be considered to create a global model, in this case of a country, or world region, that includes data from different sites.

Acknowledgments

All the authors wish to thank the EAFIT University's Vice President for Discovery and Creation for their support in this research. Likewise, the authors would like to thank the HAMADI 4.0 project, code 22-STIC-06, of the STIC-AmSud regional program, for supporting this work.

References

[1] M. Bach, M. Spremić, D. Vugec, “Integrating Digital Transformation Strategies into Firms: Values, Routes and Best Practice Examples,” *Manag. Technol. Challenges Digit. Age*, pp. 107–128, 2018, doi: 10.1201/9781351238922-5.

[2] G. Nápoles, M. L. Espinosa, I. Grau, K. Vanhoof, “FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps,” *Int. J. Artif. Intell. Tools*, vol. 27, no. 7, 2018, doi: 10.1142/S0218213018600102.

[3] W. Hoyos, J. Aguilar, M. Toro, “A clinical decision-support system for dengue based on fuzzy cognitive maps,” *Health Care Manag. Sci.*, vol. 25, no. 4, pp. 666–681, 2022, doi: 10.1007/s10729-022-09611-6.

[4] I. Bolaños-Portilla, I. Hurtado-Sánchez, I. Restrepo-Tarquino, “Diffuse cognitive maps for analysis of vulnerability to climate variability in andean rural micro-watersheds,” *DYN*A, vol. 87, no. 212, pp. 38–46, 2020, doi: 10.15446/dyna. v87n212.79943.

[5] A. Hosseini, J. Heidary, J. Antucheviciene, M. Azari, S. Razavi, “Supplier selection in the industry 4.0 era by using a fuzzy cognitive map and hesitant fuzzy linguistic VIKOR methodology,” *Environ. Sci. Pollut. Res*., 2023, doi: 10.1007/s11356-023-26004-6.

[6] N. Yuliantari, N. Pramuki, “The Role of Competitive Advantage in Mediating the Relationship Between Digital Transformation and MSME Performance in Bali,” *J. Ekon. Bisnis Jagaditha*, vol. 9, no. 1, pp. 66–75, 2022, doi: 10.22225/jj.9.1.2022.66-75.

[7] E. Gökalp, V. Martinez, “Digital transformation maturity assessment: development of the digital transformation capability maturity model,” *Int. J. Prod. Res*., vol. 60, no. 20, pp. 6282–6302, 2022, doi: 10.1080/00207543.2021.1991020.

[8] T. Aguiar, S. Gomes, P. Da Cunha, M. Da Silva, “Digital transformation capability maturity model framework,” in *Proc. IEEE 23rd Int. Enterp. Distrib. Object Comput. Conf*., pp. 51–57, 2019, doi: 10.1109/EDOC.2019.00016.

[9] V. Veldhoven, J. Vanthienen," Designing a Comprehensive Understanding of Digital Transformation and its Impact". In proc. *Bled eConference*. 2019.

[10] D. Sen, M. Ozturk, O. Vayvay, “An Overview of Big Data for Growth in SMEs,” *Procedia - Soc. Behav. Sci.*, vol. 235, pp. 159–167, 2016, doi: 10.1016/j.sbspro.2016.11.011.

[11] L. Morales Escobar, J. Aguilar, A. Garcés-Jiménez, J. Gutierrez De Mesa, J. Gomez-Pulido, “Advanced Fuzzy-Logic-Based Context-Driven Control for HVAC Management Systems in Buildings," *IEEE Access*, vol. 8, pp. 16111-16126, 2020, doi: 10.1109/ACCESS.2020.2966545.

[12] A. Singh, T. Hess, “How chief digital officers promote the digital transformation of their companies,” *Mis. Exec.*, vol. 16, no. 1, pp. 1–17, 2017, doi: 10.4324/9780429286797-9.

[13] A. Margiono, “Digital transformation: setting the pace,” *J. Bus. Strategy*, vol. 42, no. 5, pp. 315–322, 2020, doi: 10.1108/JBS-11-2019-0215.

[14] B. Kosko, “Fuzzy cognitive maps,” *International Journal of Man-Machine Studies*, vol. 24, no. 1. pp. 65–75, 1986, doi: 10.1016/S0020-7373(86)80040-2.

[15] J. Aguilar, “A fuzzy cognitive map based on the random neural model,” *Lect. Notes in Comput. Sci*., vol. 2070, pp. 333–338, 2001, doi: 10.1007/3-540-45517-5_37.

[16] G. Nápoles, I. Grau, L. Concepción, L. Koutsoviti Koumeri, J. Papa, “Modeling implicit bias with fuzzy cognitive maps,” *Neurocomputing*, vol. 481, pp. 33–45, 2022, doi: 10.1016/j.neucom.2022.01.070.

[17] G. Xirogiannis, M. Glykas, C., Staikouras “Fuzzy Cognitive Maps in Banking Business Process Performance Measurement". In *Fuzzy Cognitive Maps. Studies in Fuzziness and Soft Computing* (Glykas M. ed.), Springer, pp. 161–200, 2010.

[18] H. Sánchez, J. Aguilar, O. Terán, J. Gutiérrez de Mesa, "Modeling the process of shaping the public opinion through Multilevel Fuzzy Cognitive Maps", *Applied Soft Computing*, vol. 85, 2019, doi: 10.1016/j.asoc.2019.105756.

[19] J. Salmeron, S. Rahimi, A. Navali, A. Sadeghpour "Medical diagnosis of Rheumatoid Arthritis using data driven PSO-FCM with scarce datasets". *Neurocomputing*, pp.:104–112. 2017, doi: 10. 1016/j. neucom. 2016. 09. 113.

[20] L. Morales, C. Ouedraogo, J. Aguilar, C. Chassot, S. Medjiah, K. Drira, " Experimental Comparison of the Diagnostic Capabilities of Classification and Clustering Algorithms for the QoS Management in an Autonomic IoT Platform", *Service Oriented Computing and Applications*, vol. 13, no. 3, pp 199–219, 2019.

[21] N. Perozo, J. Aguilar, O. Terán and H. Molina, "A Verification Method for MASOES," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 64-76, 2013, doi: 10.1109/TSMCB.2012.2199106.

[22] W. Hoyos, J. Aguilar, M. Toro, "Dengue models based on machine learning techniques: A systematic literature review", *Artificial Intelligence in Medicine*, vol. 119, 2021, doi: /10.1016/j.artmed.2021.102157.

[23] J. Aguilar, I. Bessembel, M. Cerrada, F. Hidrobo, F. Narciso, "Una Metodología para el Modelado de Sistemas de Ingeniería Orientado a Agentes", *Revista Iberoamericana de Inteligencia Artificial*, vol. 12, no. 38, pp.39-60, 2008.

[24] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, L. Chamba-Eras, "Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service". In: *New Advances in Information Systems and Technologies* (Rocha, Á., Correia, A., Adeli, H., Reis, L., Mendonça Teixeira, M., eds.), vol. 444. Springer, pp. 993–1002, 2016.

[25] J. Aguilar, "Multilayer Cognitive Maps in the Resolution of Problems using the FCM Designer Tool", *Applied Artificial Intelligence,* vol. 30, no. 7, pp. 720-743, 2016.

[26] G. Csardi, T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Syst.*, no. 1695, 2006.

[27] R. Axelrod, "Structure of decision: The cognitive maps of political elites". Princeton University Press, 1976.

[28] W. Stach, L. Kurgan, W. Pedrycz and M. Reformat, "Genetic learning of fuzzy cognitive maps", *Fuzzy Sets Syst.*, vol. 153, no. 3, pp. 371-401, 2005.

[29] C. Murungweni, M. van Wijk, JA. Andersson, E. Smaling, and K. Giller. "Application of fuzzy cognitive mapping in livelihood vulnerability analysis". *Ecology and Society*, vol. 16, no. 4, 2011.

[30] A. Al Farsi, D. Petrovic, and F. Doctor. "A non-iterative reasoning algorithm for fuzzy cognitive maps based on type 2 fuzzy sets*". Information Sciences,* vol. 622, pp. 319–336, 2023.

[31] F. Aghaeipoor, M. Sabokrou, and A. Fernandez. "Fuzzy rule-based explainer systems for deep neural networks: From local explainability to global understanding". *IEEE Transactions on Fuzzy Systems*, vol. 31, no 9, pp. 3069–3080, 2023.

[32] M. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier". In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. 2016.

[33] X. Mi, B. B. Zou, F. Zou. "Permutation-based identification of important biomarkers for complex diseases via machine learning models". *NatCommun*, vol. 12, 2022.

[34] D. Benito, C. Quintero, J. Aguilar, J. M. Ramírez and A. Fernández Anta "Explainability Analysis: An In-Depth Comparison between Fuzzy Cognitive Maps and LAMDA", *Applied Soft Computing Journal*, 2024

[35] I. Apostolopoulos, P. Groumpos, "Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence. *Appl. Sci.* vol. 12, 2023, https://doi.org/10.3390/app13063412.

[36] G. Nápoles, N. Ranković, Y. Salgueiro, "On the interpretability of Fuzzy Cognitive Maps", *Knowledge-Based Systems,* vol. 281, 2023.

[37] M. Tyrovolas, X. Liang, C. Stylios, C. "Information flow-based fuzzy cognitive maps with enhanced interpretability". *Granul. Comput.* vol. 8, pp. 2021–2038, 2023.

[38] F. Hoitsma, A. Knoben, M. Espinosa, G. Nápoles, “Symbolic Explanation Module for Fuzzy Cognitive Map-Based Reasoning Models”. Lecture Notes in Computer Science, vol 12498. pp 21–34, Springer, 2020. https://doi.org/10.1007/978-3-030-63799-6_2

[39] T. Mansouri, S. Vadera, “Explainable fault prediction using learning fuzzy cognitive maps”. *Expert Systems*, vol. 40, no. 8, pp. e13316, 2023.

[40] Y. Wei and B. Jiang, "Interpretable Cognitive State Prediction via Temporal Fuzzy Cognitive Map," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 514-526, 2024.

[41] Departamento Administrativo Nacional de Estadística, “Colombia - Encuesta de Desarrollo e Innovación Tecnológica - EDIT - Industria 2017 - 2018”, [Online]. Available: http://microdatos.dane.gov.co/index.php

[42] Scott M. Lundberg y Su-In Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NIPS), 2017.

[43] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," AAAI Conference on Artificial Intelligence, 2018.

[44] Hooker, S., Erhan, D., Kindermans, P., & Kim, B. (2018). A Benchmark for Interpretability Methods in Deep Neural Networks. Neural Information Processing Systems.

Appendix F. Assessment of the Digital Transformation Level in MSMEs through Trained Fuzzy Cognitive Maps.

Assessment of the Digital Transformation Level in MSMEs through Trained Fuzzy Cognitive Maps

Jairo Fuentes [1,2], Jose Aguilar [2,3,4,*], William Hoyos [2,5,6]
[1] Universidad Francisco de Paula Santander, Cúcuta, Colombia
[2] GIDITIC, Universidad EAFIT, Medellín, Colombia
[3] CEMISID, Universidad de Los Andes, Mérida, Venezuela
[4] IMDEA Networks Institute, Madrid, Spain
[5] Grupo de Investigación ISI, Universidad Cooperativa de Colombia, Montería, Colombia
[6] GIMBIC, Universidad de Córdoba, Montería, Colombia
[*] Correspondence author

Abstract

The essence of digital transformation lies in harnessing digital technologies to foster novel approaches within organizations, leading to the creation of innovative processes, models, and services. This digitalization of data and operations generates significant value and empowers organizations to embrace innovation, automation, and diverse capabilities, encompassing both well-established technologies and cutting-edge advancements, like artificial intelligence. In this research, two innovative fuzzy cognitive maps are proposed, one led by experts and the other based on data to assess the degree of digital transformation in Micro, Small and Medium Enterprises. The essential digital transformation variables are classified into five distinct groups: i) Organization and Culture, which includes strategies, work methods, and ecosystems, ii) Customer, including services, digital channels, and products, iii) Operations and Internal Processes, which encompasses the supply chain, suppliers, and business model, and iv) Information Technology, comprising innovation, digitization, data, and analytics. The last category of variables is the target, which signifies the digital transformation level of the organization. Remarkably, our model demonstrated an impressive accuracy of 99.4% in determining an organization's digital transformation level. Consequently, our fuzzy cognitive map is proficient in forecasting and studying the factors linked to digital transformation in MSMEs. Additionally, the paper carries out a comprehensive

explainability analysis to assist MSMEs in comprehending their present situation regarding digital transformation based on the FCMs results.

Keywords: fuzzy cognitive maps, explainability analysis, digital transformation.

1. Introduction

The essence of digital transformation lies in integrating digital technologies seamlessly into an organization's framework, processes, and cultural fabric to bolster its strategies and overall effectiveness [1]. Embracing digital transformation entails envisioning how digital technologies can elevate service levels, streamline costs, and optimize inventory management, among other things, through the strategic implementation of organizational changes facilitated by these technologies. On the other hand, fuzzy cognitive maps (FCMs) are structured as graphs, which support causal reasoning in both forward and backward directions [2]. Their exceptional utility emerges in domains characterized by inherent fuzziness in concepts and relationships, such as politics, military affairs, and historical events. Moreover, FCMs have demonstrated their efficacy in tackling complex problem modeling in various domains, including healthcare [3], social network analysis [4], [26], and emergent processes [5], [6], among others. This research primarily aims to utilize FCMs for evaluating the digital transformation levels prevalent in Micro, Small and Medium Enterprises (MSMEs).

An FCM serves as a vital tool for analyzing data related to digital transformation drivers, leading to informative diagnostic results. Expert opinions play a key role in refining the causal relationships between different digital transformation concepts for greater precision. Also, a data-based FCM can be used, which will allow handling a large number of data related to the variables/concepts, to compare the results considering the experts or the data.

1.1 Related Works

Regarding FCM applications in classification and prediction tasks, [3] introduces an approach that integrates various indicators commonly utilized in diagnosing dengue, such as clinical symptoms, laboratory tests, and physical findings, to gauge the severity of dengue cases. Similarly, [4] leverages an FCM to explore vulnerability factors associated with climate variability in Andean rural micro-watersheds. Additionally, another study adopts FCM in conjunction with the VIKOR fuzzy linguistic methodology to facilitate vendor selection within the realm of Industry 4.0 [5].

On the other hand, in digital transformation, the research conducted in [15] investigates its impact on the competitive advantage of MSMEs. The study highlights that competitive advantage serves as a mediator between digital transformation and the performance of MSMEs. In [7], the objective is to construct a comprehensive maturity model for digital transformation, known as DX-CMM, using a theory-based approach. This model is designed to assist organizations in evaluating their current level of capability and maturity in digital transformation, and identifying areas for improvement. The findings of the study [15] revealed that competitive advantage played a partial mediating role in the relationship between digital transformation and MSME performance. The framework proposed in [16] introduces a phased approach to the maturity model for digital transformation capabilities, enabling organizations to assess their current digital capabilities and devise a roadmap for enhancements to attain higher levels of proficiency. In [7], a comprehensive review of existing literature on digital transformation maturity models was conducted, identifying that none of the 18 models examined met the criteria of appropriateness, completeness, clarity, and objectivity. The authors of [17] determined that digital transformation involves the strategic use of digital technologies to drive notable changes and enhancements in an organization's operations. Finally, the authors of [30] proposed a preliminary FCM based on experts for the evaluation of the level of digital transformation in MSMEs.

According to the review of existing literature, the current state of research lacks adequate maturity models that allow organizations to comprehensively evaluate their digital competencies and establish an improvement roadmap, which makes difficult to advance in their efforts in digital transformation. On the other hand, there are no robust computational tools for evaluating organizational maturity in digital transformation. A first incipient result presents a first model, but not adjusted to the environment data, nor following a methodological framework that can consider existing maturity models. It is in this context that this work is developed.

1.2 Contributions

This investigation is built upon four distinct groups of concepts originating from COBIT's Governance and Management Objectives [7]. These concepts are systematically categorized into four types, capturing the organization's behavior across diverse levels, encompassing Organization and Culture, Customer, Operations and Internal Processes, and Information Technology. We add a last concept, called Objective, representing the organization's digital transformation level. The primary contributions of this study are outlined below:

- A methodology to build FCMs to evaluate the degree of digital transformation in MSMEs.
- Two FCMs, one based on experts and the other on data, both built with our methodology.
- A comprehensive explainability analysis to assist MSMEs in comprehending their present situation regarding digital transformation based on the FCMs results.

The structure of this paper is as follows: Section 2 presents the theoretical framework used, Section 3 introduces our FCM model for assessing digital transformation levels in MSMEs, Section 4 provides details on the computational experiments conducted, Section 5 compares our work with other studies, and Section 6 offers conclusions and future research possibilities.

2. Theoretical background

This section presents the background of Digital Transformation in MSMEs. In addition, this section provides the basic concepts of the MCFs used in this research.

2.1 Background on Digital Transformation

Digital transformation refers to the process through which companies or organizations utilize digital technologies to bring about significant changes and improvements in their business activities [8]. The foundation and key to corporate digital transformation lie in the utilization of digital technologies [9]. The definition of "digital technology" in the context of digital transformation varies among different sources of literature. Some sources define it broadly and vaguely, encompassing new IT solutions and digitalization [10] [11] [12]. On the other hand, another category of literature emphasizes emerging and cutting-edge digital technologies, such as big data, cloud computing, mobile connectivity, artificial intelligence, the Internet of Things, and blockchain [13]. For instance, [13] consider new digital technologies to include social media, mobile connectivity, analytics, and embedded devices. Remane et al. [14] argue that digital technologies encompass cloud computing, mobile connectivity, and big data, among others.

2.1.1 COBIT Model of Digital Transformation

COBIT has been used as a framework specifically designed to analyze the digital transformation within an organization [7]. This framework's scope extends across the entire enterprise, encompassing all technological and information-processing endeavors undertaken by the organization to realize its objectives.

The model establishes some components or concepts to represent general aspects in the context of digital transformation in an organization. These concepts are classified into four types, capturing the behavior of the organization at various levels, encompassing Organization and Culture, Customer, Operations and Internal Processes, and Information Technology.

The Organization and culture component is made up of the following concepts:

- **Strategy:** it refers to the need for an alignment between purpose, aspiration and actions, to have a plan for digital transformation.
- **Culture and management:** it refers to the development of leadership and ways of working, enhancing talent different areas and levels, and creating an agile and innovative culture.
- **Ecosystems of digital transformation:** it refers to the context of the digital transformation in an organization, considering not only internal processes, but also in the external environment of the organization. Collaborative work with companies in the same sector or even allies that emerge from other sectors is key to promoting digitalization and promoting technology-based innovation.
- **Form of Work:** it refers to the in-person and/or teleworking modality**.**
- **Leadership** managerial sets/activities that an individual has to influence a workgroup.

The client component is made up of the following concepts:

- **Digital Services:** it refers to the use of digital tools to improve the customer experience.
- **Products** offered by the company as a result of its job or operation.
- **Digital Channels** are the distribution channels for the different strategic messages sent to interested parties.

The operations and internal processes component refers to the incorporation of technology in their different internal processes. Each sector has core processes that are a priority at the time of digitalization. The main concepts in this component are:

- **Suppliers and supply chain:** it refers to all the supply chains of goods and services that a company requires for its operation. Having technological solution providers is key, in addition to having digitalized suppliers that accompany the company's digital transformation process.

- **New business models, digital products and services:** it refers to the exploration of new businesses based on technology, and the development of new digital products or services that allow capturing new markets.
- **Business Model:** Positioning of products or services, market launch, potential clients, and forms of financing.

The Information Technology component is defined by the following concepts:

- **Data and analytics:** it refers to the level of access to data that a company has, and the potential to monetize it.
- **Innovation and digitalization processes** based on technologies.
- **Technological Solutions:** it refers to existing technological products and tools.

2.2 Fundamentals of FCMs

2.2.1 Introduction to FCM

Fuzzy cognitive maps (FCM), introduced by Kosko, offer a knowledge-centric approach to model and simulate dynamic systems [18]. FCMs can be seen as a synthesis of fuzzy logic, neural networks, and cognitive maps. They offer a means to depict the understanding of complex systems that exhibit traits such as uncertainty and causation.

Structurally, FCMs can be depicted as fuzzy directed graphs featuring feedback loops, resembling a network of neural processing units linked by signed-weighted associations (see Figure 1). Employing this approach, one can aptly model a system using concepts (e.g., variables, objects, or entities) and the causal relationships that bind these concepts. Each concept is characterized by its activation level, denoting the degree to which a particular variable impacts others. The incorporation of fuzziness allows for the portrayal of varying degrees of causality, represented as connections between concepts [1].

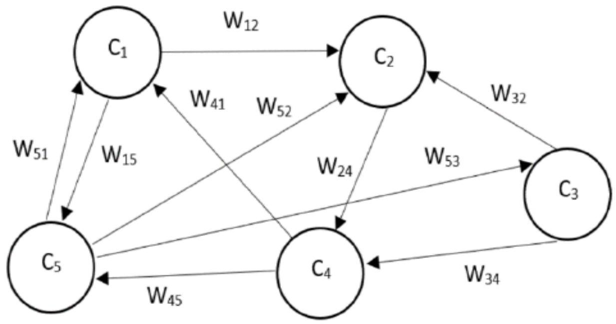

Figure 1. Example of an FCM.

Mathematically, an FCM is represented as a tuple consisting of four elements.

$$\Phi = \langle C, f(\cdot), r, W \rangle \qquad (1)$$

The set of concepts is denoted as $C \in \mathbb{R}m$, where each concept is represented by C1, C2, ..., Cm. Additionally, the activation function f (·) is applied to maintain concept values within a specified range r. Finally, the adjacency matrix $W \in \mathbb{R}m\times m$ is utilized to establish the relationships between the concepts such that Wij represents the weight assigned to the relationship from concept C_i to concept C_j. The total number of concepts in the FCM is m.

Various activation functions, as outlined in Table 1, can be employed to represent the activation of concepts in an FCM. The selection of an activation function depends on the problem under consideration. For example, in cases where evaluation criteria such as excellent, good, fair, poor, and insufficient are modeled, the sigmoid function is usually preferable to the hyperbolic tangent function. This preference is due to the fact that, in such cases, the range of values usually extends from 0 (indicating insufficient or poor) to 1 (indicating excellent), making negative values irrelevant.

Table 1. Activation functions used in FCMs.

Activation Function	Equation	Range
Bivalent	$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$	$f(x) \in \{0, 1\}$

Trivalent	f(x) = $\begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$	f(x) ϵ {-1, 0, 1}
Sigmoid	f(x) = $\frac{1}{1+e^{-iXx}}$	f(x) ϵ [0, 1]
Hyperbolic Tanget	f(x) = $\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	f(x) ϵ {-1, 1}

As we said before, in an FCM, the adjacency matrix W is employed to store the relationships among concepts. An example of such an adjacency matrix for the FCM of Figure 1 is shown in Figure 2.

$$\mathbf{W} = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \\ C_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{18} \\ w_{21} & 0 & w_{23} & 0 & w_{25} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{34} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 & w_{56} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & w_{67} & 0 \\ 0 & 0 & w_{73} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{82} & 0 & 0 & 0 & 0 & w_{87} & 0 \end{pmatrix} \end{array}$$

Figure 2. Example of Adjacency Matrix.

In certain methodologies, the assignment of weights in FCMs can involve domain experts, while in other approaches, data can also be utilized to deduce causal relationships. In this work, we will use both approaches. The assigned values determine the strength of the relationship and range from -1 to +1. A value of +1 for W_{ij} signifies that concept C_i activates concept C_j, -1 indicates an inhibiting effect between C_i and C_j, and a value of 0 denotes no causal relationship between the two concepts. Additional values within the range of -1 to +1 denote various degrees of causality. Also, additional approaches exist to denote the degrees of causality. For instance, linguistic terms can be used to define this causality. For example, if concept C_i has a significantly low incidence in concept C_j the causal relationship between the concepts could be classified as "Negative high”. Table 2 presents degrees of causality based on linguistic terms to assist domain experts in assigning relationships based on their knowledge.

Table 2. Type of causal relationships using generic logic rules based on concept values.

Linguistic term	Numerical value	Generic logic rule	
		Antecedent C_i	Consequent C_i
Positive complete	1	Very high	Very high
Positive high	0.75	High	High
Positive moderate	0.5	Medium	Medium
Positive low	0.25	Low	Low
Null	0	NA	NA

2.2.2. Reasoning of FCMs

The inference process in an FCM can be mathematically defined by incorporating three key components: the weight matrix W, the activation function, and the state vector a ∈ ℝm, which indicates the current activation levels of the concepts. The activation degree of a concept plays a crucial role in determining its activation status in a given iteration. For instance, in the range [0, 1], a concept is considered "activated" when its value is 0 at one iteration (t) and subsequently becomes greater than 0 in the next iteration (t + 1).

To derive the output in an FCM, a procedure is employed that entails iteratively computing the state vector through successive multiplications with the weights matrix until the system achieves a steady state. Equation 2 provides a formal representation of this process [18].

$$a_j(t+1) = f\left(\sum_{i=1, i\neg j}^{m} W_{ij} a_i(t)\right) \qquad (2)$$

At each iteration, the value of concept C_j at time t + 1, denoted by $a_j(t + 1)$, is calculated based on the values of the concepts at the previous iteration. The value of concept C_j at time t is $a_j(t)$. The system is considered to have reached a steady state when the values of the concepts in consecutive iterations satisfy the condition $a_j(t) = a_j(t - 1)$ or $a_j(t) – a_j(t - 1) \leq 0.001$ for all j.

In simulation scenarios, FCMs play a crucial role as they allow experts to analyze the performance of a system based on different initial conditions. The initial condition, represented as the iteration (0), encompasses a set of concepts that form the foundation of the reasoning mechanism. This initial condition is defined by a (0) and is denoted as:

$$a(0) = [a_1(0), a_2(0), \ldots\ldots.., a_m(0)] \qquad (3)$$

where $a_1(0)$ is the value of concept C_1 at iteration = 0. The FCM is able to infer from that initial condition, the final state of the variables (which is when the FCM converges).

3. FCM for the Evaluation of the level of Digital Transformation in organizations

3.1. Building of experts based FCM.

This section details the process of constructing our evaluation model about the level of digital transformation in organizations utilizing FCMs. This first FCM is based on experts in the field to assign the relationships. Thus, the construction of the FCM is based on a methodology used in [19], [20] that consists in six-steps (see Figure 3).

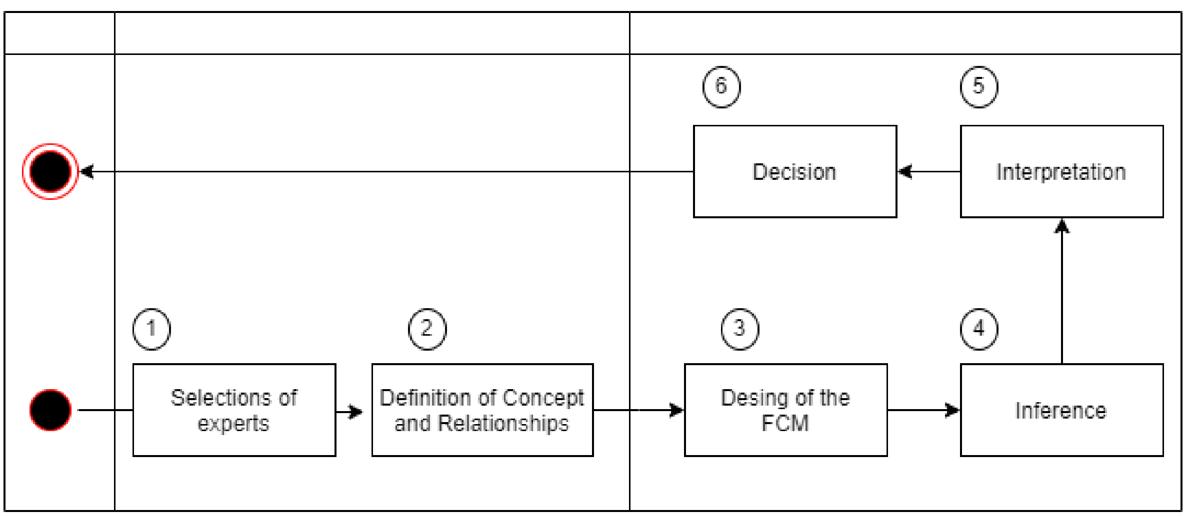

Figure 3. Flow diagram in digital transformation processes.

The rest of this section briefly describes these steps applied to our FCM.[3]

3.1.1. Selection of experts

This study involved the participation of five experts with considerable expertise in the area of digital transformation. In addition, three companies from the agribusiness sector were also selected, they have experience in digital transformation processes.

3.1.2. Concepts and relationships

The initial phase of this process is carried out by experts in digital transformation. Based on the established concepts and relationships of the COBIT model (see Figure 4), the experts constructed W with values ranging from 0 to 1 to indicate the relationships between the concepts. Table 3 shows an additional concept to those used in the COBIT model, which is the concept of the digital transformation level. This concept is the concept that will be inferred using the FCM.

Using Table 2 as a guide, the experts selected the linguistic term that best represented the relationship between each concept, and then assigned numerical values to these terms to create the weight matrix (W). For example, digital services of the customer category were related to strategies and ecosystems of the organization, and culture category with the product of the customer category, business model of the operations and internal processes category, among other relationships (see Figure 4 with the W matrix defined by the experts).

			Organization and Culture					Customer			Operations and Internal Processes			Information technology				
			C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
			Strategy	Culture and Management	Way of working	Leadership	Ecosystems	Digital Services	Products	Digital Channels	Suppliers and Supply Chain	Business model	Processes	Innovation	Digitization	Technological solutions	Data and Analytics	Target
Organization and Culture	C1	Strategy	0	0,312		0,625		0,375	0,25	0,25		0,625	0,125	0,375	0,25		0,562	0,5
	C2	Culture and Management	0,125	0	0,875	0,625	0,25			0,438	0,18	0,25	0,375	0,375		0,25	0,125	0,5
	C3	Way of working	0,25	0,75	0	0,375	0,125	0,375	0,25	0,125		0,25	0,375	0,25	0,4875	0,25		0,5
	C4	Leadership	0,5	0,5	0,25	0	0,125					0,25	0,1875	0,25			0,375	0,5
	C5	Ecosystems	0,25				0	0,375		0,625	0,375	0,125	0,25	0,375	0,125	0,125	0,125	0,5
Customer	C6	Digital Services	0,5		0,25		0,625	0	0,5	0,375	0,25	0,75		0,375	0,25	0,25	0,25	0,5
	C7	Products	0,375		0,25		0,25	0,375	0	0,437	0,375	0,125		0,25	0,437	0,25	0,25	0,5
	C8	Digital Channels	0,25	0,625	0,25		0,75	0,375	0,625	0		0,375		0,375	0,437	0,375	0,187	0,5
Operations and Internal Processes	C9	Suppliers and Supply Chain		0,375	0,25				0,25		0	0,125	0,75	0,125	0,375	0,437		0,5
	C10	Business model	0,625	0,25	0,25	0,25		0,625	0,375	0,25	0,5	0	0,625	0,625	0,125	0,25	0,375	0,5
	C11	Processes	0,25	0,25	0,25		0,25	0,625	0,25	0,25	0,875	0,25	0	0,25	0,187	0,375	0,375	0,5
Information technology	C12	Innovation	0,25	0,25			0,25	0,5		0,25	0,25	0,5	0,625	0		0,375	0,25	0,5
	C13	Digitization					0,25	0,625	0,375	0,375	0,375	0,75		0,25	0	0,375	0,687	0,5
	C14	Technological solutions			0,25		0,25	0,5	1	0,75	0,75	0,5		0,25	0,25	0	0,375	0,5
	C15	Data and Analytics	0,75			0,25	0,375		0,25	0,25		0,5	0,75	0,25	0,5	0,25	0	0,5
	CTD	Target	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	

Figure 4. Concepts and Relationships in Digital Transformation

3.1.3. FCM design

We have defined two FCMs, an experts-based FCM and a data-driven FCM. The *experts-based FCM* is built by integrating the knowledge of multiple experts into a unified global map. Equation (4) provided below [6], [21], [27] outlines the process of producing this consolidated map.

$$E_{ij}^{G} = \sum_{e=1}^{NE} \frac{E_{ij}^{e}}{NE} \tag{4}$$

The equation calculates the overall weight, denoted as EG_{ij}, for the FCM construction. It incorporates the expert opinions, represented by Ee_{ij}, on the causal relationship between concepts C_i and C_j. NE indicates the total number of experts. The FCMs are then developed, with our study utilizing the FCM Expert package (version 1.0.0) [2], while other options exist in the literature, such as [21], [26], [27]. Figure 5 displays the final FCMs. It is important to acknowledge that constructing FCMs with expert involvement introduces subjectivity, leading to potential bias in assigning causal relationships [21], [22], [26].

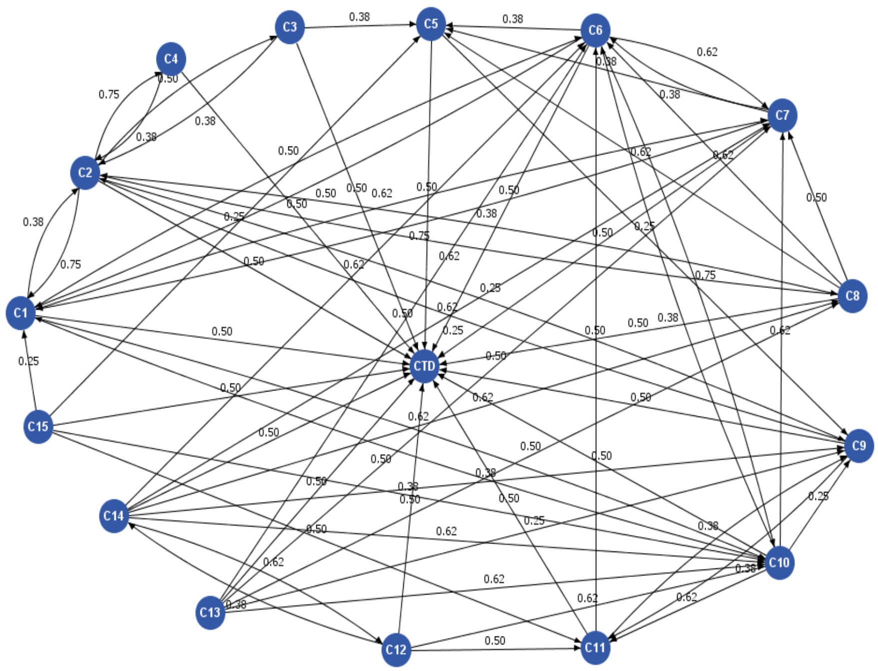

Figure 5. General FCM for evaluating the level of digital transformation in companies.

Also, for the definition of the matrix W, we have used a second approach based on data, called *data-driven FCM*. We used the particle swarm optimization method for the automatic construction of the FCM (PSO-FCM). This method uses datasets for the definition of the weight matrix (W). In the PSO-FCM algorithm, a particle i is defined as an FCM, while the particle position describes the quality of the FCM. The learning process of PSO-FCM is defined by the following procedure [31]: there are a set of particles (alternatives FCMs) with different qualities defined by equation (5):

$$v_i(t+1) = v_i(t) + s_1 r_1 \cdot (W_i best - W_i(t)) + s_2 r_2 \cdot (W_i gbest - W_i(t)) \qquad (5)$$

$$Wi(t+1) = (t) + (t) \qquad (6)$$

Equations 5 and 6 represent the update of the velocity and position of particle i (FCM), respectively; where v_i is the velocity of the particle, W_i^{best} is the best position obtained by particle i, W_i^{gbest} is the best position of a particle in the whole swarm; s_1 and s_2 are hyperparameters corresponding to the social and cognitive component.

Finally, r_1 and r_2 are random values between 0 and 1. This algorithm iterates until it converges and obtains the best particle (in our case, the best FCM).

3.1.4. Inference

In accordance with Figure 3, step 4 of the methodology is centered on executing the FCMs that were previously defined. Subsequently, the inference processes were conducted for each case study. The inference procedures is adhered to the reasoning process defined in section 2.2.2, with the sigmoid function serving as the designated activation function. Detailed information about the experimental configuration and the findings obtained in this phase are presented in section 4.2.

3.1.5. Interpretation

The interpretation of the FCM inference results for each case study was the primary objective of this step. Section 4.3 provides detailed information about the experimental setup, along with an extensive discussion of the results obtained during this phase, offering valuable insights and implications.

3.1.6. Decision

In the last stage of the systems for digital transformation, the expert practitioner is responsible for making informed decisions to drive improvements in the digital transformation journey.

4. Computational experiments

This section aims to evaluate the diagnostic efficacy of our FCM in different scenarios by employing it to categorize the levels of digital transformation in multiple organizations. Two organizations are selected as case studies for this purpose and two FCMs are defined, one based on experts and another on data. The COBIT concepts are subject to a thorough evaluation within each organization, forming the basis for FCM inference to ascertain the respective digital transformation levels.

4.1. Dataset

The dataset used in this section is from the National Administrative Department of Statistics -DANE-, a Colombian entity responsible for the planning, collection, processing, analysis, and dissemination of statistics of companies and official organizations, at the national level and international [23]. The original dataset has 635 variables, many of which are correlated with each other. We selected from the dataset the 15 variables that the experts defined as concepts in our FCMs (see Table 3). To make this selection, an analysis technique was used to determine the variables that are similar to the concepts of the COBIT model. These variables from the dataset are organizational methods, agile culture, collaborative work, work modality, work team, digital utility, products offered, distribution routes, supply chain, procedure, business pattern, analytics, novelty, digitalization and technologies. On the other hand, missing data is very common in these types of datasets; therefore, records with more than 80% missing values, for variables, are dropped [23]. Additionally, we used the Synthetic Minority Oversampling Technique (SMOTE) to balance it [32]. After this procedure, the dataset had 2556 records with 1215 for category 0 (has not a level of digital transformation) and 1341 for category 1 (has a level of digital transformation).

In the case of the learning approach, to find the best possible FCM using PSO, we use a hyperparameters optimization approach based on grid search corresponding to the combination of two activation functions (sigmoid and bivalent); two inference functions (standard Kosko and modified Kosko); and finally, initial population sizes between 50 and 100, 10 by 10. For each combination, the model is tested, and results are obtained in terms of accuracy, precision, recall and F1-Score. Finally, this dataset we divided into a part (80%) for the training of the FCM and the rest for the test phase.

4.2. Comparison of results for experts-based FCM and data-driven FCM.

In this section, we show the results of implementing our FCMs in some case studies.

4.2.1 Using the test dataset.

In the case of *data-driven FCM*, the results obtained by the application of the dataset described above correspond to a precision of 1, a precision of 1, a recall of 1 and an F1 score of 1, which indicates an excellent performance for the prediction of the concept that represents the digital transformation. The hyperparameter configuration that allowed us to achieve this result consists of an initial population size of 70 individuals, the sigmoid function as the activation function and the modified Kosko function as the reasoning inference function. The results obtained with the *FCM based on experts* correspond to an accuracy of 0.9835, precision of 0.98, Recall of 1 and F1-Score of 0.99, which indicates its performance is lower than data-driven FCM (see Table 3).

Table 3. Quality of the proposed FCM.

	accuracy	**precision**	**Recall**	**F1-Score**
Data-driven FCM	1	1	1	1
Expert-based FCM	0.9835	0.98	1.0	0.99

4.2.2 Case 1. For the Chamber of Commerce.

The chamber of commerce acts as a collective body, bringing entrepreneurs and business owners together. In an effort to streamline business operations, Chambers of Commerce are broadening their range of offerings to encompass virtual services, allowing users to conduct their tasks with enhanced speed and convenience.

Table 4 shows the input values for concepts in a range of (0.0, 0.25, 0.50, 0.75, 1.0), where 0 is the absence of value in that concept and 1 is the maximum presence of that concept. In this organization, after an evaluation of its current status in each of the concepts linked to the COBIT model of digital transformation, the values of each concept were determined as the average of the company's experts, the final values are mentioned below: the concept of Strategy (C1) with a value of 0.7, Culture and Management (C2) of 0.7, Way of Working (C3) of 0.5, Leadership (C4) of 0.3, Ecosystems (C5) of 0.5, Digital Services (C6) of 0.7, Products (C7) of 0.6, Digital

Channels (C8) of 0.8, Suppliers and Supply Chain (C9) of 0.4, Model of Business (C10) of 0.6, Processes (C11) of 0.4, Innovation (C12) of 0.8, Digitalization (C13) of 0.8, Technological Solutions (C14) of 0.9 and Data. and Analytics (C15) of 0.4. With the weights averaged by the experts, we proceeded to infer, resulting in Table 5.

Table 4. Input values for the case study 1.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0,7	0,7	0,5	0,3	0,3	0,7	0,6	0,8	0,4	0,6	0,4	0,8	0,8	0,9	0,4

Table 5 shows the results of the inference process. The initial vector of concepts is passed to the inference rule (Eq. 2). The final vector of this organization can be seen in Table 5, and each iteration is the dynamic evolution of the concepts. In the final vector, the digital transformation concept (CTD) reached a value of 0.9954 for the expert-based FCM. Based on this result, we can determine that variables related to digitization (0.638), data and analytics (0.659) are the concepts that most need to be enhanced in digital transformation in this organization. It is analyzed that the causal relationships affecting digitalization are customers and operations. The variables that affect data and analytics are affected by the organization, culture, and operations of the company. In the results of each of the concepts of digital transformation, the one that obtained the most performance was the FCM dataset. In the case of the data-driven FCM, the digital transformation concept (CTD) has a value of 0.9979, and in general, all the concepts have good results.

Table 5. Output vector for the case study 1.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0,9360	0,8878	0,7417	0,7754	0,889	0,9532	0,8539	0,8336	0,9502	0,9502	0,8832	0,741	0,6380	0,7129	0,6590	0,9954
Data-driven FCM	0,9854	0,9626	0,9514	0,9160	0,9691	0,9863	0,9793	0,9829	0,9761	0,9939	0,9778	0,9834	0,9666	0,9710	0,9770	0,9979

4.2.3. Case 2. Galavis Coffee Company.

It is a north Santander Company, founded in 1918, dedicated to the roasting and distribution of Colombian Coffee. Table 6 shows the input values for concepts in a range of (0.0, 0.25, 0.50, 0.75, 1.0), where 0 is the absence of value in that concept and 1 is the maximum presence of that concept. In this organization, after an evaluation of its current status in each of the concepts linked to the COBIT model of digital transformation, the values of each concept were determined as the average of the company's experts, the final values are mentioned below: the concept of Strategy (C1) has a value of 0.5, Culture and Management (C2) of 0.4, Way of Working (C3) of 0.3, Leadership (C4) of 0.5, Ecosystems (C5) of 0.3, Digital Services (C6) of 0.4, Products (C7) of 0.4, Digital Channels (C8) of 0.4, Suppliers and Supply Chain (C9) of 0.3, Business Model (C10) of 0.4, Processes (C11) of 0.5, Innovation (C12) of 0.3, Digitalization (C13) of 0.3, Technology Solutions (C14) of 0.3, and Data and Analytics (C15) of 0.3. With the weights averaged by the experts, we proceeded to infer, resulting in Table 7.

Table 6. Input table for the case study 2.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0,5	0,4	0,3	0,5	0,3	0,4	0,4	0,4	0,3	0,4	0,5	0,3	0,3	0,3	0,3

Table 7 shows the results of the inference process. The final vector of this organization can be seen in Table 7. In the final vector, the CTD decision concept reached a value of 0.9878 for the expert-based FCM. Based on this result, we can determine that variables related to innovation (C12), technological solutions (14) are the concepts that most need to be enhanced in digital transformation in this organization. In the results of each of the concepts of digital transformation, the one that obtained the most performance was the FCM dataset. In the case of the data-driven FCM, the digital transformation concept (CTD) has a value of 0.9980, and in general, all the concepts have again good results.

Table 7. Output vector for the case study 2

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0,8810	0,6178	0,6437	0,6651	0,7898	0,9344	0,6671	0,9636	0,9302	0,9302	0,7832	0,7110	0,4380	0,5129	0,6910	0,9411
Data driven FCM	0,9821	0,9589	0,9501	0,9002	0,9622	0,9791	0,9703	0,9754	0,9712	0,9905	0,9708	0,9788	0,9616	0,9699	0,9710	0,9980

4.3. Explicability analysis

An explainability analysis is carried in this section, rooted in the causality-driven FCM framework. FCM establishes influence connections between concepts through the application of causal relationship weights. These weights quantify the intensity or potency of one concept's influence on another, facilitating the determination of the most significant features guiding the model's decision-making. We carry out this analysis for the *FCM based on experts.*

Within this analysis, we gauge the sensibility of class concepts by excluding the least influential causal links with their input concepts. The primary objective here is to scrutinize the causal sway exerted by more influential concepts concerning their associated class concepts (outputs). This scrutiny aims to discern whether these influences hold substantial sway over the model's predictive capabilities.

4.3.1 Case 1. For the Chamber of Commerce.

According to Table 5, the concepts the most important are "C6-Digital services", with an influence of 0.9532 followed by "C10-Business model" and "C9-Suppliers and supply chain", with an influence of 0.9502, followed by "C1-Strategy", with an influence of 0.936, followed by "C5-Ecosystems", with 0.889).

Table 9 summarizes the filters selected for each data concept. Each filter is more restrictive than the previous one. In the context of the Chamber of Commerce, filter 1 is responsible for eliminating all causal relationships associated with the target concepts (classes) whose weights are equal to or less than 0.80. Filter 2 removes causal relationships with weights less than or equal to 0.75, while filter 3 removes all causal relationships with weights equal to or less than 0.66.

Table 9. Filters based on the weights of causal relationships with respect to the target variable

	Chamber of Commerce
Filter 1	<= 0.80
Filter 2	<= 0.75

Filter 3	<= 0.66

It can be seen that by applying filter 1, we eliminate 40% of the causal relationships (with C3, C4, C12, C13, C14 and C15), and that the level of digital transformation increased slightly (see Table 10).

Table 10. Using Filter 1.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	0.9625	0	0	0.969	0.9863	0.9793	0.9829	0.976	0.9939	0.9778	0	0	0	0	0.9979

It is observed that by applying filter 2, 33.33% of the causal relationships are eliminated (with C3, C12, C13, C14 and C15) and that the level of digital transformation is the same (see Table 11).

Table 11. Using Filter 2.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	,0.9625	0	0.916	0.969	0.9863	0.9793	0.9829	0.976	0.9939	0.9778	0	0	0	0	0.9979

By applying filter 3, 13.33% of the causal relationships are eliminated (with C13, C15) and the level of digital transformation is the same (see Table 12).

Table 12. Using Filter 3.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	0.9626	0.9514	0.916	0.970	0.9893	0.9729	0.9761	0.9839	0.9939	0.9778	0.9866	0	0,9709	0	0.9979

The previous tables show the results of the inference process. In the final vector, the concept of digital transformation (CTD) reached a value of 0.9979. Based on this result, we can determine that the variables C1, C2, and C5 to C11 are the concepts that most influence the digital transformation in this organization.

4.3.2. Case 2. Galavis Coffee Company.

According to Table 7, the set of concepts more important for the Galavis Coffee Company are "C8-Digital Channels" with an influence of 0.9636, followed by "C6-Digital Services" with an influence of 0.9344, followed by "C9-Suppliers and supply chain" and “C10-Business model” with an influence of 0.9302.

Table 14 summarizes the filters selected for each data concept. Each filter is more restrictive than the previous one. In the context of the Galavis Coffee Company concepts, filter 1 is responsible for eliminating all causal relationships associated with the target concepts (classes) whose weights are equal to or less than 0.85. Filter 2 removes causal relationships with weights less than or equal to 0.75, while filter 3 removes all causal relationships with weights equal to or less than 0.65.

Table 14. Filters based on the weights of causal relationships with respect to the target variable

	Chamber of Commerce
Filter 1	<= 0.85
Filter 2	<= 0.75
Filter 3	<= 0.65

It can be seen that by applying filter 1 eliminates 44% of the causal relationships (C2, C3, C4, C12, C13, C14 and C15) and that the level of digital transformation increased slightly (see Table 15).

Table 15. Using Filter 1.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	0	0	0	0.969	0.9863	0.9793	0.9829	0.976	0.9939	0.9778	0	0	0	0	0.9989

It is observed that by applying filter 2, 38.33% of the causal relationships are eliminated (C3, C4, C12, C13, C14 and C15) and that the level of digital transformation is the same (see Table 16).

Table 16. Using Filter 2.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	0.9625	0	0	0.969	0.9863	0.9793	0.9829	0.976	0.9939	0.9778	0	0	0	0	0.9979

By applying filter 3, 6.66% of the causal relationships of class are eliminated (C13) and the percentage of digital transformation is the same (see Table 17).

Table 17. Using Filter 3.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	CTD
Expert based FCM	0.9854	0.9625	0.9513	0.916	0.969	0.9863	0.9793	0.9829	0.976	0.9939	0.9778	0.9834	0	0.9709	0.977	0.9979

We can conclude that the concepts more influent are C1, C5 to C11. Upon closer examination, it becomes evident that strategy, Ecosystem, and the concepts of clients and internal processes have a high influence to mold the digital transformation in this organization.

5. Comparisons with other works

5.1. Quantitative Comparison with other ML techniques

In this section, we evaluate the predictive capacity of our *data-driven FCM* and compare it with other machine learning techniques proposed in the literature. For these tests, we have used the datasets of each paper, the same 10-fold cross-validation process, and for the evaluation model, the precision and F1 measures.

Eom et al.'s research [25] involved the utilization of machine learning methods to project the performance of innovation in ICT business processes, and extracted the critical variables influencing performance prediction. The central outcome emphasized the random forest (RF) model's precision in anticipating ICT business process innovation performance. It's essential to note that while the study predominantly addresses this element, it does not delve into other dimensions of digital transformation within MSMEs. The research effectively identifies the pivotal variables that hold substantial sway over performance prediction. The results of the comparative analysis of the two machine learning models are summarized in Table 18.

Table 18. Comparison of performance between [25] with our work.

	Learning methods.	Accuracy	F1-score
[25]	RF	0.9465	0.5808
This work	FCM	1.0000	1.0000

In particular, the FCM approach proved to be the most effective in terms of predictive performance, achieving an accuracy rate of 1 and an F1 score of 1. In contrast, the RF model recorded the least satisfactory predictive results, achieving an accuracy rate of 0.9465 and an F1 score of 0.5808.

The study presented in [29] leverages ergodic theory alongside advanced machine learning methods to analyze the intricate and multidimensional relationships between knowledge sources' breadth and depth and their impact on innovation performance The results of the comparative analysis of the two machine learning models are summarized in Table 19.

Table 19. Comparison of performance between [29] with our work.

	Learning methods.	Accuracy	F1-score
[29]	CNN	0.8365	0.7314
This work	FCM	0.99	0.98

In particular, the FCM approach proved to be the most effective in terms of predictive performance, achieving an accuracy rate of 0.99 and an F1 score of 0.98. In contrast, the CNN model recorded the least satisfactory predictive results, achieving an accuracy rate of 0.8365 and an F1 score of 0.7314.

5.2. Qualitative Comparison with other works.

To compare this study with other works, a set of three criteria were established, which are outlined below:

- Cri1: It uses fuzzy cognitive maps (FCMs).
- Cri1: It uses machine learning in the diagnostic model.
- Cri2: It explains the reason for the level of deduced digital transformation.

Table 20 provides a comprehensive comparison of our research with previous studies, focusing on digital transformation criteria.

Table 20. Comparison with other works.

	Cri1	Cri2	Cri3
[8]	√	√	X
[9]	X	X	√
[10]	X	X	√
[11]	X	X	√
[12]	X	X	√
[13]	X	X	√
[14]	X	X	√
[24]	X	√	√
[25]	X	√	X
[28]	X	X	√
[29]	X	√	√
[30]	√	X	X
This work	√	√	√

The first criterion is met by both our research, [8] and [30], [8] and [30] incorporate FCM for the diagnosis of dengue and the digital transformation in an organization. The second criterion is satisfied by [8], [24], [25], [29] and our study. These works use machine learning techniques for the construction of the diagnostic model. The third criterion is satisfied by [9], [10], [11], [12], [13], [14], [24], [28], [29] and our study. These works delve into the connection between digital transformation and MSME performance, conduct digital transformation maturity assessments, and provide comprehensive information on the implications of digital transformation. Our research is particularly focused on the digital transformation of agribusiness MSMEs, and also, performs a deep explainability analysis using techniques from the area of explainable artificial intelligence.

Our study successfully meets all the requirements of a diagnostic model, specifically designed for analyzing digital transformation within an organization. It employs FCMs, a non-intrusive machine learning technique, and effectively explains conclusions through causal relationships between concepts.

6. Conclusions

This study aimed to develop an FCM specifically designed to evaluate the digital transformation level in MSMEs. To achieve this, a computational tool was proposed to analyze and categorize variables into five types: Organization and Culture, Customer, Operations and Internal Processes, Information Technologies, and Objective. The values of these variables together determine the organization's level of digital transformation. Through the use of FCM, we were able to explore the dynamics of factors linked to digital transformation in MSMEs. Our approach prioritizes transparency and interpretability, providing a clear understanding of the outcomes based on the various concepts encompassed in the COBIT model of digital transformation.

According to the explainability analysis of our model, we were able to determine the essential variables in agro-industrial MSMEs that must exhibit high values to drive a significant digital transformation. The findings emphasized the importance of Strategy (C1) and Ecosystem (C5). Raising these variables to high values guarantees a notable level of digital transformation within organizations.

The FCM model advanced in this study comes with its own set of limitations. Primarily, it doesn't integrate the Digital Maturity Model, which furnishes information on the digitalization progress of processes, capabilities, and deficiencies in MSMEs. This model also quantifies the level of accomplishment in a company's digital transformation. The second point to consider is that the *based-expert FCMs m*ight introduce bias into the assessment of model accuracy, as experts were responsible for defining both the structure and weightings of the FCM. In contrast, *data-driven FCM* models generally provide increased precision by leveraging the collective insights of numerous MSME firms. However, it's vital to highlight that this dataset must undergo meticulous preparation, cleansing, and processing to ensure its fitness for purpose.

In subsequent research, it is essential to concentrate on reducing bias in the digital transformation evaluation model utilized within organizations. A learning process for the FCM topology (and its weights) should be explored using organizational historical data linked to the COBIT model variables. Also, in forthcoming research, the focus may revolve around understanding the means by which alignment is nurtured within the IT function and in its interactions with the business realm, leveraging the capabilities of artificial intelligence tools to propel digital transformation. Finally, another future work is the recommendation for action in the field of Business Process Management (BPM) and process optimization through machine learning and artificial intelligence [24].

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Information

Jose Aguilar was partially supported by grant 22-STIC-06 (HAMADI 4.0 project) funded by the STIC-AmSud regional program.

Author contribution

J*. Fuentes:* Conceptualization, Formal analysis, Simulation, Writing, *J. Aguilar:* Conceptualization, Methodology, Formal analysis, Writing, *W. Hoyos:* Methodology, Formal analysis, Writing, *E. Montoya:* Writing, Funding acquisition.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

[1] B. P. Examples, “Integrating Digital Transformation Strategies into Firms: Values, Routes and Best Practice Examples,” *Management and Technological Challenges in the Digital Age*, no. January, pp. 107–128, 2018, doi: 10.1201/9781351238922-5.

[2] G. Nápoles, M. L. Espinosa, I. Grau, and K. Vanhoof, “FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps,” *International Journal on Artificial Intelligence Tools*, vol. 27, no. 7, 2018, doi: 10.1142/S0218213018600102.

[3] W. Hoyos, J. Aguilar, and M. Toro, “A clinical decision-support system for dengue based on fuzzy cognitive maps,” *Health Care Manag Sci*, vol. 25, no. 4, pp. 666–681, 2022, doi: 10.1007/s10729-022-09611-6.

[4] I. C. Bolaños-Portilla, I. C. Hurtado-Sánchez, and I. Restrepo-Tarquino, “Diffuse cognitive maps for analysis of vulnerability to climate variability in andean rural micro-watersheds,” *DYNA (Colombia)*, vol. 87, no. 212, pp. 38–46, 2020, doi: 10.15446/dyna. v87n212.79943.

[5] A. Hosseini Dolatabad, J. Heidary Dahooie, J. Antucheviciene, M. Azari, and S. H. Razavi Hajiagha, “Supplier selection in the industry 4.0 era by using a fuzzy cognitive map and hesitant fuzzy linguistic VIKOR methodology,” *Environmental Science and Pollution Research*, no. 0123456789, 2023, doi: 10.1007/s11356-023-26004-6.

[6] N. Perozo, J. Aguilar, O. Terán, and H. Molina, “A verification method for MASOES,” *IEEE Trans Cybern*, vol. 43, no. 1, pp. 64–76, 2013, doi: 10.1109/TSMCB.2012.2199106.

[7] E. Gökalp and V. Martinez, “Digital transformation maturity assessment: development of the digital transformation capability maturity model,” *Int J Prod Res*, vol. 60, no. 20, pp. 6282–6302, 2022, doi: 10.1080/00207543.2021.1991020.

[8] H. Demirkan, J. C. Spohrer, and J. J. Welser, “Digital Innovation and Strategic Transformation,” *IT Prof*, vol. 18, no. 6, pp. 14–18, 2016, doi: 10.1109/MITP.2016.115.

[9] P. C. Verhoef *et al.*, “Digital transformation: A multidisciplinary reflection and research agenda,” *J Bus Res*, vol. 122, no. July 2018, pp. 889–901, 2021, doi: 10.1016/j.jbusres.2019.09.022.

[10] L. Heilig, S. Schwarze, and S. Voß, “An analysis of digital transformation in the history and future of modern ports,” *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2017-Janua, no. January, pp. 1341–1350, 2017, doi: 10.24251/hicss.2017.160.

[11] C. Legner *et al.*, “Digitalization: Opportunity and Challenge for the Business and Information Systems Engineering Community,” *Business and Information Systems Engineering*, vol. 59, no. 4, pp. 301–308, 2017, doi: 10.1007/s12599-017-0484-2.

[12] B. Horlach, P. Drews, I. Schirmer, and T. Böhmann, “Increasing the agility of it delivery: Five types of bimodal IT organization,” *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2017-Janua, pp. 5420–5429, 2017, doi: 10.24251/hicss.2017.656.

[13] A. Singh and T. Hess, “How chief digital officers promote the digital transformation of their companies,” *MIS Quarterly Executive*, vol. 16, no. 1, pp. 1–17, 2017, doi: 10.4324/9780429286797-9.

[14] G. Remane, A. Hanelt, R. C. Nickerson, and L. M. Kolbe, “Discovering digital business models in traditional industries,” *Journal of Business Strategy*, vol. 38, no. 2, pp. 41–51, 2017, doi: 10.1108/JBS-10-2016-0127.

[15] N. P. Y. Yuliantari and N. M. W. A. Pramuki, “The Role of Competitive Advantage in Mediating the Relationship Between Digital Transformation and MSME Performance in Bali,” *Jurnal Ekonomi & Bisnis JAGADITHA*, vol. 9, no. 1, pp. 66–75, 2022, doi: 10.22225/jj.9.1.2022.66-75.

[16] T. Aguiar, S. B. Gomes, P. R. Da Cunha, and M. M. Da Silva, “Digital transformation capability maturity model framework,” *Proceedings - 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference, EDOC 2019*, no. Section VI, pp. 51–57, 2019, doi: 10.1109/EDOC.2019.00016.

[17] Z. van Veldhoven and J. Vanthienen, “Designing a comprehensive understanding of digital transformation and its impact,” *32nd Bled eConference Humanizing Technology for a Sustainable Society, BLED 2019 - Conference Proceedings*, no. March 2022, pp. 745–763, 2020, doi: 10.18690/978-961-286-280-0.39.

[18] B. Kosko, “Fuzzy cognitive maps,” *International Journal of Man-Machine Studies*, vol. 24, no. 1. pp. 65–75, 1986. doi: 10.1016/S0020-7373(86)80040-2.

[19] L. Morales, C. Ouedraogo, J. Aguilar, J. . Chassot. S. Medjiah, K. Drira, “Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform. *Service Oriented Computing and Applications,* vol. **13**, pp. 199–219, 2019. doi: 10.1007/s11761-019-00266-w

[20] J. Aguilar, M. Cerrada, F. Hidrobo, Y. Tech, F. Narciso, “Una Metodología para el Modelado de Sistemas de Ingeniería Orientado a Agentes”. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 12, pp. 39-60, 2008.

[21] J. Aguilar, J. Contreras, “The FCM designer tool,” *Studies in Fuzziness and Soft Computing*, vol. 247, no. August, pp. 71–87, 2010, doi: 10.1007/978-3-642-03220-2_4.

[22] J. Aguilar, “A fuzzy cognitive map based on the random neural model,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 2070, pp. 333–338, 2001, doi: 10.1007/3-540-45517-5_37.

[23] Departamento Administrativo Nacional de Estadística, “Colombia - Encuesta de Desarrollo e Innovación Tecnológica - EDIT - Industria 2017 - 2018,” *vol. 8*, p. 413, 2020, [Online]. Available: http://microdatos.dane.gov.co/index.php

[24] D. Paschek, C. T. Luminosu, and A. Draghici, “Automated business process management-in times of digital transformation using machine learning or artificial intelligence,” *MATEC Web of Conferences*, vol. 121, 2017, doi: 10.1051/matecconf/201712104007.

[25] T. Eom, C. Woo, and D. Chun, “Predicting an ICT business process innovation as a digital transformation with machine learning techniques,” *Technol Anal Strateg Manag*, 2022, doi: 10.1080/09537325.2022.2132927.

[26] H. Sánchez, J. Aguilar, O. Terán, J. Gutiérrez de Mesa, “Modeling the process of shaping the public opinion through Multilevel Fuzzy Cognitive Maps”, Applied Soft Computing, vol. 85, 2019, doi.: 10.1016/j.asoc.2019.105756.

[27] J. Aguilar “Multilayer Cognitive Maps in the Resolution of Problems using the FCM Designer Tool”, Applied Artificial Intelligence, vol. 30, pp. 720-743, 2016, doi: 10.1080/08839514.2016.1214422

[28] L. Casi and L. Resmini, “Spatial complexity and interactions in the FDI attractiveness of regions,” *Papers in Regional Science*, vol. 93, no. S1, pp. S51–S78, 2014, doi: 10.1111/pirs.12100.

[29] P. Hajek and R. Henriques, “Modelling innovation performance of European regions using multi-output neural networks,” *PLoS One*, vol. 12, no. 10, pp. 1–21, 2017, doi: 10.1371/journal.pone.0185755.

[30] J. Fuentes, J. Aguilar, W. Hoyos, E. Montoya, "Evaluation of the Level of Digital Transformation in MSMEs using Fuzzy Cognitive Maps based on Experts”, Proceeding XLIX Latin American Computer Conference, 2023.

[31] W. Hoyos, J. Aguilar, and Toro M. Federated learning approaches for fuzzy cognitive maps to support clinical decision-making in dengue. Engineering Applications of Artificial Intelligence. Vol. 123, April. pp 1-106371, 2023, doi: 10.1016/j.engappai.2023.106371

[32] K. Bence, T. Flóra, N. Csaba, Ó. Péter. Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment. Ecological Applications. Vol. 30, no. 321-357. Doi: 10.1002/eap.2043.

Appendix G. Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs

Meta-learning Architecture for Autonomous Cycles of Data Analysis Tasks to Automate Coffee Production of MSMEs

Jairo Fuentes [1,2], Jose Aguilar [2,3,4,*]
[1] Universidad Francisco de Paula Santander, Cúcuta, Colombia
[2] GIDITIC, Universidad EAFIT, Medellín, Colombia
[3] CEMISID, Universidad de Los Andes, Mérida, Venezuela
[4] IMDEA Networks Institute, Madrid, Spain
emails: jafuentesc@eafit.edu.co; aguilar@ula.ve

Abstract

In the coffee industry, efficiency is a determining factor in achieving product quality objectives, as well as ensuring sustainable practices, both from a financial and environmental point of view. On the other hand, in the MSME agroindustrial sector, it is crucial to carry out constant monitoring and detailed analysis of the characteristics of the coffee bean and the roasting process to identify areas for improvement in production. In a previous study, three autonomous cycles of data analysis tasks are described to automate operations in the production chains of MSMEs. The proposal of this article consists of a meta-learning (MTL) architecture for autonomous cycles. The proposed MTL framework has a meta-models table that stores relevant information about pre-trained machine learning models for autonomous cycles. This meta-model table is used to identify the most appropriate machine learning model each time a new coffee dataset is introduced into an organization. Specifically, the framework is designed to automate coffee production in MSMEs using different transfer learning techniques such as model, parameter and data transfer. It uses one of the models in the meta-model table on the new dataset if this new dataset is similar enough to one previously used to make that model; performs a parameter transfer and builds a new model if the similarity to another dataset is lower; or performs a data transfer and builds a new model if that similarity is even lower. This article explores the use of this MTL framework in some case studies, using records of coffee roasting processes from different coffee companies. The results indicate the great effectiveness of our system in reusing/adjusting/creating machine learning models in various situations such as evaluating coffee quality, reducing waste, and selecting coffee processing processes, reaching an accuracy of 90.45% on average in the results.

Keywords: Meta-learning, Data Analysis, Machine Learning

1. Introduction

Coffee production is an industrial sector of great importance for some countries due to its economic and environmental impact. On an economic level, coffee is one of the most in-demand products globally, with a production network that includes millions of farmers and distributors. In many developing regions, the coffee industry is carried out by MSMEs, boosting employment and strengthening local economies. On an environmental level, coffee production requires large amounts of water and fertile soil, and its processing produces a great deal of pollution (the pulp and husk can cause environmental damage).

Given the above, the implementation of digital technologies and artificial intelligence in this industry allows for improved operational efficiency, increased profitability and the promotion of more sustainable practices. In particular, coffee production must operate at certain levels of efficiency to ensure that bean quality, the roasting process, and overall coffee processing methods meet financial and environmental sustainability goals. This will allow actors in the MSME agroindustrial sector to be aware of what is happening in coffee production and to implement strategies for its improvement.

On the other hand, in [23], the authors introduced the concept of autonomous cycles of data analysis tasks (ACODAT), and in [1], [2], [3] was applied to the management of smart classrooms, which were later used for the management of learning processes [8], [23], [24]. In [11], [13], [22], [25], the authors proposed using these cycles in collaborative industrial production processes within the framework of Industry 4.0. In turn, the authors of [18] used MTL for the selection and configuration of automated learning algorithms, and in [14] for the management of data transfer processes.

The previous works do not propose an MTL framework for ACODAT in the context of agro-industrial production processes of MSMEs. Specifically, for the management of machine learning models linked to coffee, such as models that detect and predict its quality, its loss, and determine coffee processing methods, among other tasks.

In this article, we present a novel MTL framework, aiming to automate the construction of machine learning models for quality prediction, determination of coffee processing methods, among others. In particular, the proposed solution seeks to adapt machine learning models previously created with coffee bean datasets to new coffee bean datasets. The main contributions of this work are:

- The definition of the first MTL architecture for ACODAT, which can be adapted to different contexts for the reuse/adaptation of machine learning models to new datasets.
- The first application of MTL architecture in the context of the coffee industry, for the management of machine learning models for the detection and prediction of quality, the prediction of waste, and the selection of processing methods for coffee beans in a roasting process, which can be successfully adapted to different scenarios in coffee production.

The rest of this article is structured as follows: Section 2 presents the theoretical framework. Section 3 describes our MTL_ACODAT framework, and section 4 its instantiation and some case studies. Furthermore, section 5 shows the general results and a comparison with other works. Finally, section 6 presents conclusions and future work.

2. Theoretical Framework

2.1. Meta-Learning

The concept of meta-learning, also called learning to learn, involves exhaustively examining the performance of machine learning models on multiple tasks to try to improve them. The objective is to extract knowledge, or meta-data, to improve the performance of these models, in particular, when they want to be used in new tasks [15], [16], [17], [18].

Meta-learning seeks to understand and take advantage of previous experiences. In general, meta-learning requires detailed knowledge of the context to properly adjust machine learning algorithms to the requirements of the problems to be solved. To do this, in an initial phase, it is necessary to collect meta-data that details the behavior of pre-trained machine learning models. This meta-data describes the specific configurations of the machine learning algorithms used in training the models, including their hyperparameters, model quality metrics, training times, and the characteristics of the datasets used, among other things, all of them are known as meta-attributes. That meta-data is then used to identify and reuse that knowledge in new contexts, for new datasets, in emerging tasks [29],[30].

The above may involve performing several tasks. For example, eventually, if their data has similar behaviors, these models can be used on new datasets; or if the similarity of the datasets is lower, the hyperparameters found in the previous techniques can be reused; or, if they are not so similar and the new dataset is missing data, data can be transferred. In the first case, the meta-data of the machine learning model must be adjusted to the new use in the new dataset, in the other cases, the new machine learning model created is simply a new input in the meta-model's table. Thus, it is seen that the reuse of prior knowledge is made possible, or learning new uses of previously learned models, among other things.

In general, meta-learning makes it possible to acquire the ability to master new tasks significantly faster than traditional methods would allow. It does this by greatly accelerating and improving the creation of new learning models using machine learning techniques [1], [2],[3]. Also, meta-learning can be used to address the challenge of handling changes in machine learning models, known as "concept drift" [34]. Using meta-learning techniques, the reappearance of previous contexts can be detected and the corresponding models can be proactively activated. This approach is especially useful in situations where concepts may reappear, such as seasonal changes, allowing prior knowledge to be reused.

2.2. ACODAT

ACODAT has as its main objective to adopt a data-driven perspective to solve business challenges [8], [11], [13], [22], [23], [24], [25]. Different tasks related to data analysis must be performed simultaneously to achieve a specific goal within the process under observation. These tasks are interconnected with each other, and each plays a unique role in the cycle: (i) observing the process, (ii) analyzing and interpreting what is happening in it, and (iii) making decisions to guide the process toward achieving its objectives. This integration of tasks in a closed cycle makes it possible to efficiently address complex problems. The functions of each task are described below.

- **Observation task:** this is a set of tasks in charge of monitoring the process. They must capture data and information on the behavior of the environment. In addition, they are responsible for the preparation of such data.
- **Analysis tasks:** these are a set of tasks aimed at interpreting and understanding the data collected to diagnose what is happening in the context monitored by the cycle. These activities involve the creation of knowledge models (machine learning models) based on the observed dynamics.
- **Decision-making task:** these activities comprise the implementation of actions aimed at making decisions that promote the improvement of the process in which the cycle is implemented. These actions have an impact on the dynamics of the process to optimize it. The results of these actions are reviewed again during the observation and analysis stages of the cycle, which initiates a new iteration of the cycle.

ACODAT has been used in previous research [8] and is based on the principles introduced by IBM in 2001 [4], [5], [6], [7], [10], [12]. It has been applied in various domains such as smart classrooms [8], [13], [23], [24], telecommunications [11], Industry 4.0 [25] and smart cities [22]. It adopts the autonomous computing approach, with the goal of providing systems that use intelligent control cycles with autonomous capabilities. In general terms, an ACODAT requires:

- A multidimensional data model to hold the collected data that characterizes the behavior of the context, and the machine learning models built with this data.
- An infrastructure to merge the indispensable technological tools to carry out data analysis activities.

Although ACODAT has demonstrated its effectiveness in various fields, it has not yet been extended with MTL processes to improve its adaptive capabilities to the characteristics of different contexts autonomously.

3. MTL_ACODAT Framework.

In our research, we propose MTL architecture to automatically generate knowledge models for ACODAT. Our MTL mechanism allows ACODAT to learn how to quickly adapt to new environments. Specifically, our MTL architecture requires detailed knowledge of the various machine learning aspects, which include knowledge of data sources, data characteristics (meta-features), and knowledge models (meta-models) previously defined. Below, we present our architecture.

3.1. Conceptual architecture

Figure 1 outlines the general structure of the proposed framework, and Figure 2 shows the flowchart of the MTL_ACODAT framework. In this framework, the main objective is to make informed decisions about how to choose and/or exploit information about knowledge models previously developed using machine learning techniques. To do this, the framework needs a knowledge base that includes information about the previous knowledge models, the datasets used for their constructions, and the hyperparameters of the machine learning techniques used. In particular, each time a new dataset and the knowledge model to be built (predictor, prescriber, etc.) are introduced to the framework, the *selection module* provides recommendations on the most appropriate knowledge models and the corresponding hyperparameter configurations of machine learning techniques. Next, the *adaptation module* analyzes the similarity of the input dataset with the one most similar to it in the framework to determine what to do. If they are very similar, then the knowledge model is simply used; if they are somewhat similar, then only the parameters of the machine learning techniques will be used; and finally, if they are a little similar, then it is evaluated whether the new dataset has little data. If it has a lot of data, then a new knowledge model is simply built, otherwise, data transfer can be carried out or synthetic data can be generated, depending on the similarity of the new data with those in the framework. Following this, the *management module* incorporates this new knowledge into the knowledge base residing within the meta-models table. Thus, our approach can reuse old knowledge, can integrate new knowledge (knowledge models), among other things, based on MTL principles. Next, we go on to describe each of the modules of architecture.

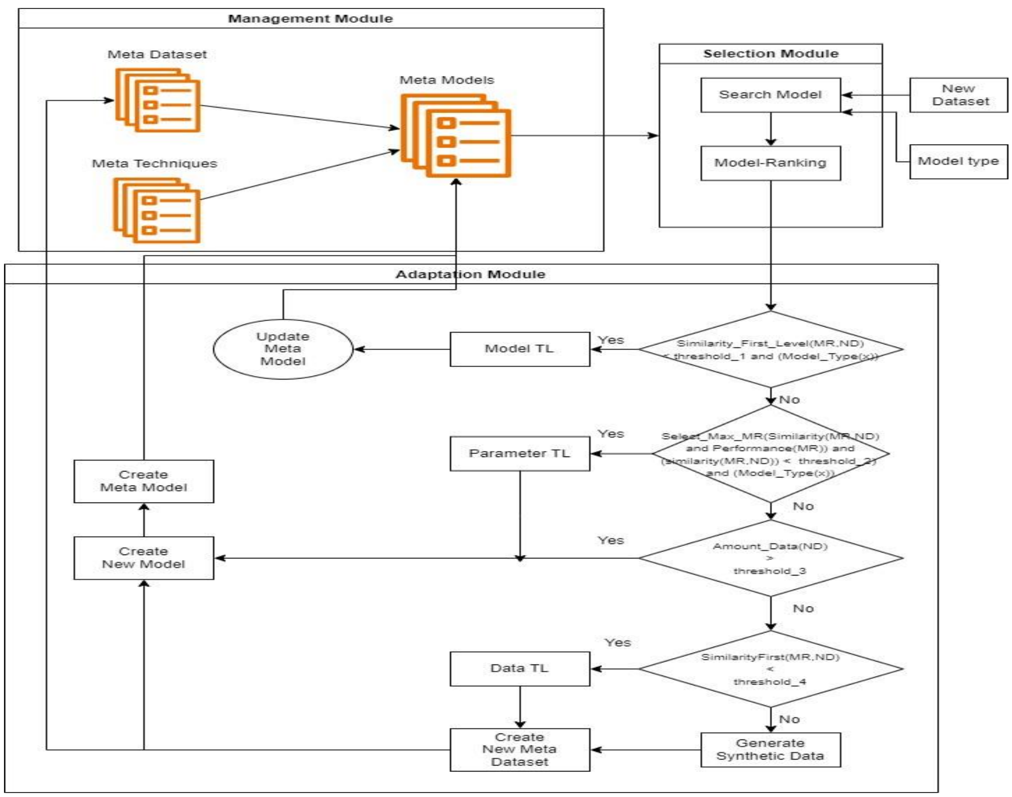

Figure 1. Conceptual Architecture of the MTL_ACODAT Framework for MSMEs

In Figure 2, we can see that the MTL_ACODAT framework flowchart naturally integrates synthetic data generation and transfer learning in its different modalities (data, models or parameters).

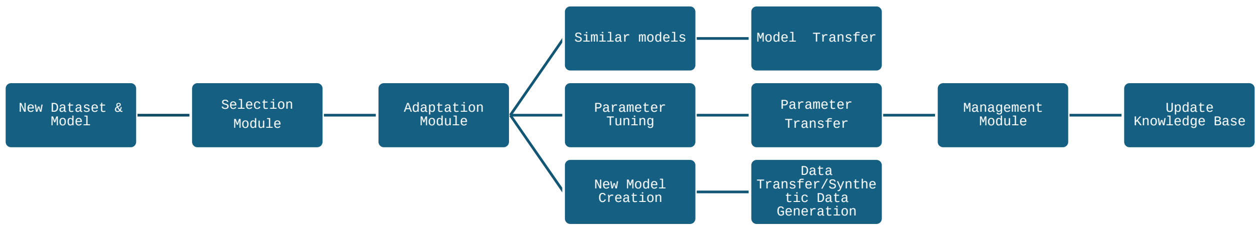

Figure 2. Flowchart of the MTL_ACODAT Framework

3.2. Management Module

This module manages the different components of the framework. In particular, it maintains the information of the knowledge base made up of a table with the meta-models, another with the meta-data, and finally, another with the meta-techniques. Thus, each record (knowledge model) in the meta-models table is related to the data (saved in the meta-data) and the machine learning technique (saved in the meta-techniques) that were used to create it. This record in the meta-models table also stores the values of the quality metrics of that knowledge model and any other information about it. On the other hand, the record in the meta-data contains information about the datasets, such as what its attributes are, where it is located, etc. Finally, in the meta-techniques, there is information on the machine learning techniques, such as the optimal values of their parameters, among other things. Thus, this module mainly records the various knowledge models previously obtained by different learning algorithms in previous datasets.

Algorithm 1 describes the macro algorithm of this module. Basically, it has two steps, if a new knowledge model is created, then this new information must be recorded in the tables in the knowledge base (see step 1.1); otherwise, the meta-model and meta-technique tables are updated, and the new dataset is recorded in MD (see step 1.2).

Algorithm 1. Management module.

Input: Action performed in the adaptation module
MD: Meta-Datasets

MT: Meta-Techniques MM: Meta-Models Procedure: 1. Update of information in the knowledge base depends on the action performed in the update module. 1.1 If a new knowledge model is created, then include this information in MD, MT and MM 1.2 If an old knowledge model is used, then update MM and MT and include this new dataset in the MD
Output: Updated Knowledge base

Particularly, the meta-datasets table is composed of its identifier (ID_MD), the description of its variables (VD1, ..., VDN), its location and creation date, among other things (see Table 1). The meta-techniques table (see Table 2) contains the identifier of the machine learning technique (ID_MT), the description of its behavior (description), its parameters (P1, P2, PN) and its name (name_ml). Finally, the meta-models table (see Table 3) is composed of the knowledge model identifier (MM_ID), its description (description) specifying its type of knowledge model (descriptive, predictive, diagnostic, prescriptive, among others), the ID_MT of the learning technique used to build it, the ID_MD of the dataset used to build it, and specifically, the variables used (VD1 ... VDN), and the values of the model in the quality metrics (QM).

Table 1. Meta-Datasets Table

Id_MD	Description	Vd1…Vdn	Location	Date Created

Table 2. Meta-Techniques Table

Id_MT	Description	Parameters (P1, P2, Pn)	Name_ML

Table 3. Meta-Models Table

Id_MM	Description	Id_MT	VD1 …VDn	Quality (QM)	Id_MD

3.3. Selection Module

The activation of the selection module occurs every time a new requirement arrives for the construction of a knowledge model, with the dataset to be used (see macro algorithm in Algorithm 2). The module proceeds to extract the characteristics of the incoming dataset and begins to determine the datasets that were used to build knowledge models of the same type as the one desired to be built. Then, it goes on to compare the input dataset with those datasets (compare their characteristics) to generate a ranking (classification) of similarity between them and the new dataset. Thus, using this ranking, this module offers a suggestion of the most similar knowledge models, and the hyperparameters of the machine learning techniques used to build them.

Algorithm 2. Selection Module.

Input: New Model (NM) and New dataset (ND)
Procedure: 1. For a given NM and ND do 1.1 Vector_Similar_Models= Sort models based on the similarity of NM and ND
Output: Vector_Similar_Models

3.4. Adaptation Module

The adaptation module defines the action to be carried out (see Algorithm 3). If the similarity of the dataset in the first position of Vector_Similar_Models and the new dataset is less than threshold1, then old knowledge is used (step 2, which is called Transfer Learning of Model, Model_TL). Otherwise, an old knowledge model is selected based on a new ranking that includes the similarity between the datasets and the performance (quality) of the old knowledge models. If the first model in this new ranking has a value greater than threshold2 (step 5), then a new knowledge model is created using its machine learning technique (its parameters, which is called Parameter_TL) (see step 6). Otherwise, if the new dataset has enough data and is greater than threshold3 (step 8), then a new Model is created (step 9) with the new dataset. Otherwise, if the new dataset has little data and the similarity of the dataset in the first position of Vector_Similar_Models is less than threshold4 (step 11), then it is used a data transfer (called Data_TL) to this new dataset of the model (see figure (step 12), otherwise, it is necessary the generation of synthetic data, (step 15). Finally, for these last steps, a new model is created with the new dataset (see Algorithm 3). When this comparison is finished, the management module is invoked.

Algorithm 3. Adaptation Module.

<u>Input:</u> Vector_Similar_Models

1. if (first position of Vector_Similar_Models > threshold1)
    1.1 Use knowledge model of the first position of Vector_Similar_Models
        % Model Transfer
    Else
    a. performance of the first position of Vector_Similar_Models = Perfomance_Models
    b. If Perfomance_Models > threshold2)
        1.3.1 Use the parameters of the machine learning technique of the knowledge model of the first position of Vector_Similar_Models to create a new Model
            % Parameter Transfer
        Else
        1.3.2 If (Amount (ND) > threshold3)
            1.3.2.1 Create a new Model.
        Else
            1.3.2.2 If (first position of Vector_Similar_Models > threshold4)
                1.3.2.2.1 Create a New_ Dataset with ND and dataset of the first position of Vector_Similar_Models
                1.3.2.2.2 Create a new Model.
                % Data Transfer
                Else
                1.3.2.2.3 Create a New_ Dataset =Synthetic_Data (ND)
                1.3.2.2.4 Create a new Model

Output: new module or new performance of the reused model

Step (1) of the algorithm in Algorithm 3 involves measuring the similarity between the different variables that make up the datasets, to determine those with the greatest similarity. This implies comparing variable by variable, to determine the similarity in the statistical behavior between them (if they follow the same distribution, the similarity in the centrality and dispersion metrics, which are highly correlated, among other aspects can be analyzed). Based on this, it can be determined to reuse an old knowledge model (Model_TL) (see Figure 3). Thus, this procedure involves comparing the similarity between the data from the source domain (with which a previous knowledge model was built) and that from the target domain (new dataset), to choose the one that is most similar to the latter. This allows selecting a model that has been trained in the source domain and then adjusting it to the target domain [14], [20]. This approach is based on the premise that there are similarities between the source and target domains that can be exploited to take advantage of already established knowledge and adapt it to a new context. This approach turns out to be an efficient tactic when the source domain is rich in training data and notable similarities with the target domain can be found.

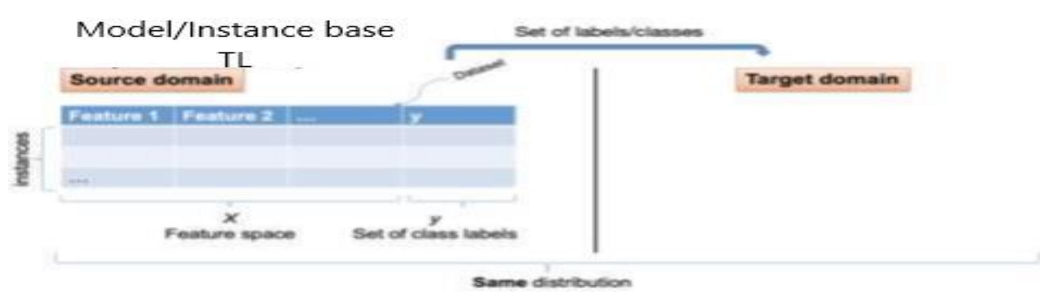

Figure. 3. Example of Model Transfer Learning

Parameter-based transfer learning (Parameter TL, see step (6) of the algorithm in Algorithm 3) refers to the transfer of hyperparameters from the source domain to the target domain (see Figure 4). This approach assumes that models for related tasks share the same hyperparameters of the machine learning techniques. Thus, parameter-based transfer learning offers the advantage of a quick construction and improving performance on the new task [14], [16], [20].

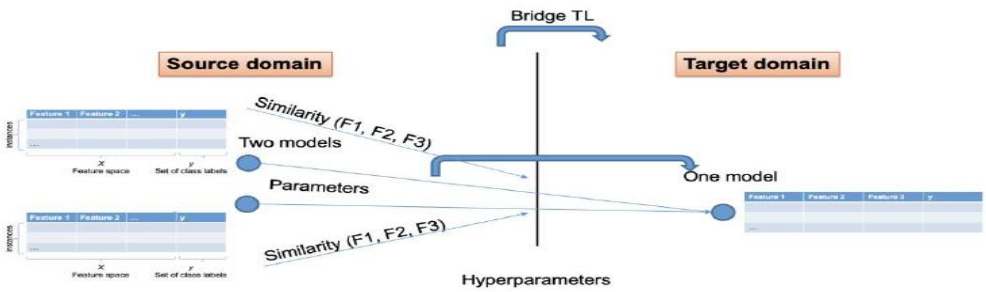

Figure 4. Example of Parameter Transfer Learning.

Data transfer learning involves the aggregation of data to the new dataset (Data TL, see step (12) of the algorithm in Algorithm 3), that is, new samples or instances. Therefore, the source domain transfers data due to the scarcity of data in the dataset in the destination domain. Figure 5 shows how the data transfer approach is used. The similarity between the instances in all features (F1, F2, F3) between the source domain and the target domain is used to compare the different sources. The source domain that has the greatest similarity to the target domain and a good metric quality is chosen. Therefore, this process generates new instances in the target domain based on the similarity between the datasets.

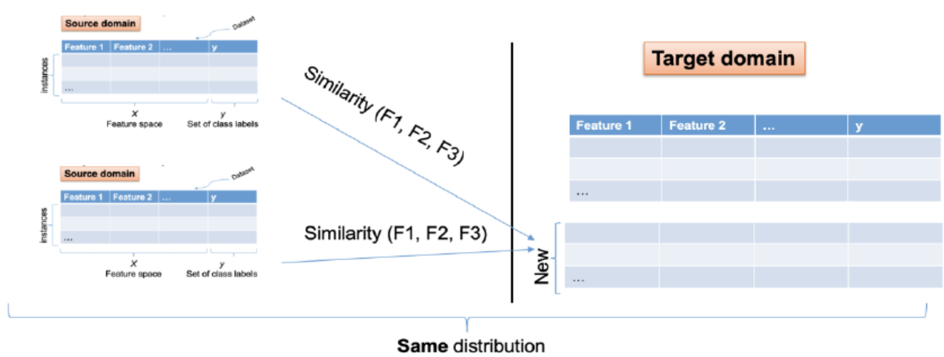

Figure 5. Example of Data Transfer Learning.

In summary, this module is responsible for choosing the best model, the most appropriate hyperparameters, or new data for a new dataset. Additionally, the knowledge base is regularly updated, allowing the framework to continually improve its performance over time.

Our architecture is an excellent solution to mitigate initial data limitations (cold starts), as it allows reusing models, parameters or data from previous machine learning processes. In addition, it explicitly invokes synthetic data generation processes when certain conditions are met. In particular, this ability allows it to respond to the lack of historical data in the agro-industrial sector. Our architecture can use any of the synthetic data generation methods, such as deep learning-based models, or data augmentation techniques based on interpolation or extrapolation methods.

In terms of data privacy and security, our architecture allows for several transfer learning mechanisms, each exploiting a different type of knowledge (models, parameters, and data). Collecting, processing, and storing that information in our architecture to enable transfer learning can create vulnerabilities in terms of privacy, security, and data protection. Now, our architecture can implement mechanisms to anonymize data, for ethical data use, and thus ensuring privacy. It can also include security mechanisms in its tables (encryption, access control, among others) to make the architecture robust.

4. Case study

4.1. Description of case study

To explain the functioning of the MTL framework, we will consider the ACODAT of Figure 6, which determines the type of input that will be transformed into an industrial process and, according to it, the production process to be implemented. This ACODAT was defined in [25]. Mainly, this cycle is made up of three tasks. Table 4 shows the overview of each task of this autonomous cycle.

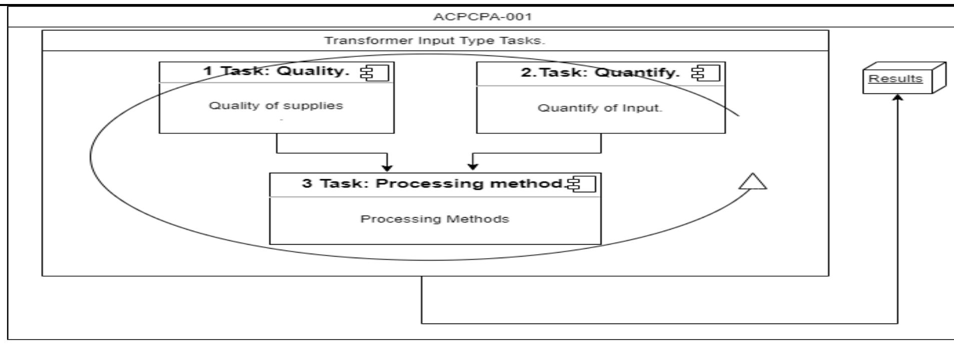

Figure 6. Autonomous Cycle for process automation in MSMEs (called ACPCPA – 001).

Table 4. Description of ACPCPA - 001 tasks.

Name of the task	Knowledge models
1. Determine the quantity of input to be transformed.	Diagnostic model
2. Determine the quality of the input to be processed.	Predictive model
3. Determine the method of processing.	Predictive model

Task 1. Quantity of input to transform: This task identifies the raw materials required to meet customer demands. It analyzes different aspects such as the historical evolution, uses of the product, etc. This task uses a diagnostic model to identify the product and the quantity to transform.

Task 2. Quality of inputs to process: This task determines the quality of the inputs, which depends on cultural practices, storage and transport services used, among other things. This task uses a predictive model to determine this quality.

Task 3. Method of processing to use: This task identifies factors related to the processing method. For example, natural washed, natural, or mechanical drying, automatic or manual selection, among others. A predictive model is used for this task.

4.2. Instantiation case study Coffee Galavis.

The cycle was instanced in the Café Galavis Company located at Cucuta, Colombia. This company, founded in 1918, is dedicated to the roasting and distribution of Colombian Coffee. They have sophisticated equipment for the process that takes place there, from the time the raw material is received at the company's facilities to its storage in optimal temperature and humidity conditions. In addition, they include the most rigorous selection of the bean, its purification prior to processing, and the most hygienic conditions for its internal transport to the packaging machines of the company. The datasets used in this case to build the knowledge models are described in Table 5.

Table 5. Datasets used to build the tasks in ACODAT.

Name of the task	Data sources
1. Determine the quantity of input to be transformed.	Production demand, customers, temperature, and humidity of the environment.
2. Determine the quality of the input to be processed.	Inputs used on farms, cultural practices, storage and transport services, demand, size, density, acidity, Defect category.
3. Determine the method of processing.	aroma, flavour, acidity, body, uniformity, sweetness, and moisture.

These datasets used in each of the data analysis tasks previously indicated come from the companies Colcafe, Café Premium De Colombia S A S and Café Galavis, all from Colombia. The datasets were directly provided by their IT departments, reaching a prior confidentiality agreement with them. In addition, the collection process was complemented with extra information from their websites. In the preprocessing process, a feature engineering phase was carried out in order to clean and normalize the data, eliminate outliers and determine the most relevant variables for each of the data analysis tasks, according to the machine learning algorithms to be applied, both supervised and unsupervised.

In particular, the confidentiality agreement signed with the companies to use their datasets establishes the anonymization of the data, respect for the use of the data for the purpose of the work, and its storage in a secure place.

4.2.1. Coffee Bean Quality Model

The dataset utilized for this task consists of evaluations conducted by experts on various coffee blends, where each cup taste is assigned a score. Each entry contains detailed information regarding the coffee's origin, the recipe applied, the characteristics of the raw materials, and the resulting product. With a dataset size of 1,347 records and 44 variables, the primary objective is to estimate coffee quality using unsupervised learning algorithms [17], [28].

The process begins with feature engineering procedures aimed at removing features or variables with numerous missing or outlier values and minimizing redundancy among highly correlated variables, among other tasks. This process yields a dataset comprising 845 records and 34 columns. Subsequently, an exhaustive examination is conducted to identify the most significant variables for gauging coffee bean quality. Employing Pearson and Spearman correlation analyses, this assessment highlights flavor, aroma, consistency, and acidity as critical variables. These variables are then utilized to formulate the unsupervised model.

For the definition of the quality model, three unsupervised learning techniques were used: K-means, Agglomerative Clustering and DBScan. Thus, two distance-based clustering techniques and one Density-Based clustering technique are used. In this case of the machine learning techniques, we have used three clusters (K=3) to delineate coffee quality into low, medium, and high tiers. To determine the most appropriate model, we have utilized the Silhouette Index, Density-Based Clustering Validation, and the Davies–Bouldin index [9], [26] (see Table 6).

Table 6. Comparison of models.

Model	Silhouette Index	Density-Based Clustering Validation	Davies–Bouldin index
K-Means	0.85	-0.41	0.09
Clustering	0.64	-0.45	3.82
DBScan	0.41	-0.122	10.75

Therefore, in this three-category model, the K-means model exhibits optimal performance in describing coffee quality based on flavor, aroma, consistency, and acidity variables. In this case, K-means gave the best results because the data distribution is uniform, which fits with the K-means assumption that clusters have well-separated spherical shapes. The biggest problem with this technique is obtaining the value of K, which can limit obtaining good results. If the data distribution had been more complex, DBScan would have been better, but since this was not the case, its silhouette index and Density-Based Clustering Validation values are very low. It is evident that it fails to form clear dense clusters. It is possible to improve these results by exploring other more versatile alternatives such as GMM techniques that can adapt to more complex data distributions. In the case of K-means, the K optimization strategy can be improved.

4.2.2. Task for the prediction of the decrease of the bean in the roasting process.

In the realm of coffee roasting, it's a common phenomenon for coffee beans to undergo expansion while concurrently undergoing moisture reduction, culminating in a decline in the final batch weight [19],[26]. Within this procedural domain, an array of product and procedural variables exists, with some amenable to manipulation to attenuate coffee bean loss. The principal endeavor at hand involves devising a supervised model geared towards predicting the extent of bean reduction during the roasting process. Thus, both the input attributes of the product must be considered (including the previously ascertained grain quality) and the coffee roasting procedural variables.

As before, a feature engineering process is carried out. The Dataset has 2,632 records and 23 variables, from which the grain decrease will be estimated. Next, the grain decrease value is calculated for each record as: $a\ sted\ weight)/Green\ weight$.

It represents the percentage of weight loss during the roasting process. In the sampled data, this loss ranges between 7.6% and 19.87%, with 75% of the observations exceeding 14.5% and half of them between 14.5% and 15.8. %. To develop the predictive model, three classic supervised learning approaches were used: random forest, logistic regression, and support vector machine [27]. For evaluation purposes, the data set was split into training and testing sets, with 80% dedicated to training and 20% to testing. After this partitioning, R^2, MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) and MSE (Mean Square Error) were calculated to evaluate the performance of the model. The most successful model, Random Forest (as shown in Table 7), achieved a remarkable 0.9 in R^2, 0.81 in MAPE, 0.08 in MAE and 0.5 grams in MSE, showing high accuracy with respect to the other models tested.

Table 7. Training and testing accuracy for all fitted models.

Model	R^2	MAPE	MAE	MSE
Random Forest	0.9	0.81	0.08	0,5 grams
Logistic Regression	0.67	0.71	0.35	3.5 grams
SVM	0.66	0.68	0.42	4.1 grams

Several factors explain why, in this case, Random Forest outperformed logistic regression and SVM in estimating the weight reduction of the bean in the coffee roasting process:

- It has greater accuracy in non-linear models: the coffee roasting process does not follow a linear behavior, which makes models such as logistic regression less effective. Random Forest, based on multiple decision trees, manages to better capture these complex relationships.
- It has a great capacity to handle high-dimensional data: the existence of 2632 records and 23 variables requires a model that can process large volumes of

data without affecting its performance. Random Forest handles this information efficiently by randomly selecting features in each division.

- Greater Robustness In the face of Overfitting, SVM can tend to overfit if the hyperparameters are not correctly configured. Random Forest, on the other hand, minimizes this risk by combining multiple decision trees, guaranteeing stability in the results.
- Identification of Key Variables, one of the strengths of Random Forest is its ability to evaluate the importance of each variable in predicting bean weight loss. This facilitates the optimization process in the coffee industry.

4.2.3. Task for the determination of the "Coffee processing method".

Apart from the quality of the coffee, it is also possible to establish the "Processing method". For this task, a dataset with 1347 records and 44 variables is studied, from which the "Coffee processing method" will be determined. Within these variables are also included the variables of bean quality and bean decrease in the roasting process previously estimated.

To build the identification model of the "Production method", we will use two machine-learning techniques: Logistic Regression and K nearest neighbours. The techniques were trained with 883 registers and tested with 464. The best result was obtained with logistic regression (0.72, 0,69, and 0.32), with respect to K nearest neighbours (see Table 8).

Table 8. Training and testing accuracy for the models.

Model	R^2	MAPE	MAE
Logistic Regression	0.69	0.72	0.32
K-nearest neighbors	0.63	0.64	0.58

In this case, logistic regression gave better results because it allowed identifying the relationships between the input and output variables well. KNN had a lower performance than expected. Although, in the case of logistic regression, its assumption of linear relationships may prevent it from adequately capturing complexity, in this case, it was sufficient. Particularly, in this task, other techniques such as artificial neural networks, random forest or gradient boosting could be used to follow up on the relationships between the data, which seemed to have less linear behavior, which would surely allow us to improve the results.

4.3. Instantiation of the MTL architecture in coffee production.

The objective of the management module is to store the information of the known knowledge models. Thus, this module is responsible for storing Meta-Models, with its Meta-Datasets and Meta-Techniques. For example, for the predictive models of task 2, three traditional supervised learning techniques were used: Random Trees (RF), Logistic Regression (LR) and Support Vector Machine (SVM). The MM table has information on these knowledge models (see Table 11) and their quality Metrics (M1, M2, etc.). Also, it has information about the datasets used in the Meta-Datasets (see Table 9). For example, the variables from the Dataset 02 are total, ncf flame 1, greenweight, roasted weight, water, color, humidity, and density. Finally, the Meta-techniques table has information on the three machine learning techniques used to build this knowledge model and their parameters (see Table 10). One example of Logistic Regression is its ID (MT_03) and its parameters (P1, P3, P4 and P5), among other variables.

Using the previous case study, the knowledge base tables of the framework are filled. For example, we could have ACODAT instantiated for several types of coffee, so we would have a set of knowledge models for each type of task and type of coffee using various machine learning techniques and the characteristics of each of the datasets by type of coffee. An example of the information saved when instantiating the ACODAT in different cases is shown in the tables below.

Table 9. Meta-Datasets table in the case study

Id_MD	Description	Vd1…Vdn	Location	Date Created
Dat_01	Coffe Galavis	flavor, aroma, consistency, and acidity	www.cafegalavis.com	2015
Dat_02	Coffe Colcafe	Total, ncf flame 1, greenweight, roasted weight, cwater, color, humidity, density.	www.colcafe.com	2020

Table 10. Meta-Techniques table in the case study

Id_MT	Description	Parameters (P1, P2, Pn)	Name_ML
MT_01	SVM	P1.	Supervised
MT_02	Random Forest	P1, P2.	Supervised
MT_03	Logistic Regression	P1, P3, P4, P5.	Supervised
MT_04	K-means	P1, P2, P3.	Unsupervised
MT_05	Clustering	P1, P2, P4, P5, P6.	Unsupervised
MT_06	DBSCan	P1, P2, P4, P5.	Unsupervised

Table 11. Meta-Models table in the case study

Id_MM	Description	Id_MT	P1, P2, Pn	VD1 …VDn	Quality (QM)	Id_MD
MM_01	Predictive	MT_01	P1, P2, P3, P4.	Vd1, Vd2, Vd3, Vd4	M1, M2, M5	Dat_01

MM_02	Predictive	MT_02	P1, P3, P4.	Vd1, Vd2, Vd3, Vd4, Vd5	M1, M3, M4	Dat_02
MM_03	Diagnostic	MT_04	P1, P2, P3..	Vd1, Vd2, Vd3, Vd4, Vd5, Vd6.	M1, M2, M6	Dat_01
MM_04	Diagnostic	MT_04	P1, P2, P3.	Vd1, Vd2, Vd3.	M1, M2, M5	Dat_02

4.3.1. Coffee Bean Quality Model for the new case study

We suppose a new dataset 03 with the next characteristics (see Table 12): Vd1, Vd2, Vd3, Vd4, Vd5.

Table 12. New Dataset

Id_MD	Description	Vd1…Vdn	Location	Date Created
Dat_03	Café Premium De Colombia S A S	Vd1, Vd2, Vd3, Vd4, Vd5	www.cafepremium.com	2022

And for that new dataset of a new type of coffee, we need to instantiate the ACODAT. For example, for the second task, we need to develop a predictive model. The first step is to determine the similarity between the new dataset and those stored in the framework (selection module). This procedure must establish a similarity metric between the source domains (Xs) and the target domain (Xt). The name similarity method was used, which examines the affinity between words considering the substrings they share, after normalizing the name of the features under evaluation. An example of this ranking is shown in Table 13.

Table 13. Ranking (MT)

(0.7) Dat01	(0.4) Dat02

4.3.2 Meta-Learning to predict the behavior of the Coffee roasting process.

This example will focus on the design of a supervised model for predicting the level of shrinkage in the roasting process, considering the input characteristics of the product and the controllable variables of the process (task 2 of the previous ACODAT). In this case, we are going to assume several situations, to evaluate the behavior of the framework.

a) First case: In this case, we assume that threshold 1 is less than 0.7. Thus, we can use the information of Dat_01, so we choose the predictive model MM_01 from 14 and we simply use it.

b) Second case: in this case, we assume that threshold 1 is 0.8, so we must proceed to make a new ranking considering the similarity and quality of the predictive models made with the datasets. In this ranking, the MM_02 model of dataset Dat_02 comes first, and since we suppose its performance is lower than the threshold2, then we do a parameter transfer of the technique used with the best performance in that model and that dataset (in this case, MT_02). Table 14 shows an example of an update of information that would occur once the new prediction model has been created using our framework.

Table 14. New Meta-Model.

Id_MM	Description	Id_MT	P1, P2, Pn	VD1 …VDn	Quality (QM)	Id_MD
MM_05	Predictive	MT_02	depth=4, rank=Gini.	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity	R2=0.8, MAPE=0.9,	Dat_03

c) Third case: in this case, it is considered that the other two cases have not been met. If the new dataset has a lot of data, we simply build a new dataset; but suppose it has little data, then we go to see if the similarity of the dataset with other datasets in the framework is greater than threshold3. We assume yes with Dat_01, then a data transfer is made. If not, synthetic data would be constructed. Table 15 shows an example of an update of information that occurs once the new prediction model has been created using our framework.

Table 15. New Meta-Model.

Id_MM	**Description**	**Id_MT**	**P1, P2, Pn**	**VD1 …VDn**	**Quality (QM)**	**Id_MD**
MM_05	Predictive	MT_01	C=0.2,	Aroma, Flavor, Balance, Uniformity	R2=0.6, MAPE=0.7,	Dat_03

To determine the thresholds of the MTL_ACODAT framework in the case study that predicts size reduction in the roasting process, statistical analysis and empirical tests were considered as explained below:

- *Threshold 1 (Similarity level for reuse of existing models).* This threshold establishes the minimum degree of similarity necessary between a new dataset and previously recorded data to allow the reuse of a predictive model. It is calculated by performing correlation studies and distance analysis (particularly, Jensen-Shannon distance) between the data distributions of the data sets. It was set at 0.7 to ensure that the similarity of the new data set is sufficiently similar so that the model to be reused is compatible, based on the work [39]. A lower threshold would allow the use of models created with data with significant differences, while a very high value would excessively restrict reuse.
- *Threshold 2 (Quality limit for parameter reuse):* This threshold sets the level of accuracy needed to determine whether parameters from a previously trained

model can be reused. In the case study, the performance metrics R^2 and MAPE were used, and it was set to 0.8 to ensure that only good quality models are considered for parameter transfer [39. A lower threshold could compromise the performance of the new model by accepting parameters from low-quality models, while a higher value could exclude models that could be improved with slight adjustments.

- *Threshold 3 (Condition of sufficient data for modeling):* This threshold determines whether the dataset is of adequate size to train a new model without the need to transfer information from another source. This value was based on statistical learning studies that establish the minimum amount of data needed to avoid overfitting and improve the generalization capacity of the model [40]. A threshold was set considering that a dataset is small if its size is less than the 25th percentile of the previous sets in the system. In this way, it is ensured that the models are not trained with insufficient data, reducing the risk of overfitting. In the case of small volumes, the transfer of data from a similar set or the generation of synthetic data is chosen.

Now, these thresholds are flexible and can be dynamically adjusted according to new criteria or using new information.

As for the parameter adjustment methods, which only occurred in the cases of parameter transfer, data transfer, creation of synthetic data or creation of models from the original dataset, a classic grid-based hyperparameter optimization process was performed [41].

4.4 Importance of case study results

The impact that our MTL framework would have on MSMEs in the coffee sector would be reflected in their production processes. The transfer learning strategies underlying our MTL framework allow information to be shared between companies in the sector to build more robust and accurate knowledge models with fewer data requirements and a lower computational burden. This is done in a context where confidentiality is guaranteed, without incurring additional costs for them, something prohibitive for MSMEs. The most relevant specific impacts include:

- *Definition of more efficient knowledge models (for example, quality prediction):* By using pre-existing knowledge models trained with historical coffee data, small and medium-sized companies can improve the accuracy of their models.
- *Adaptability to different production scenarios:* by sharing information between MSMEs, they also share the different environmental conditions and cultivation techniques used by each one, which gives greater adaptability to the models built with this shared data, facilitating their operation in changing scenarios.
- *Decreased development times and computational costs:* Implementing transfer learning reduces both development costs and data collection and preparation costs, by avoiding building knowledge models from scratch.
- *Improvements in the production process:* the tasks of the autonomous cycle improve planning and quality in production, by optimizing grain selection, improving quality control, and managing processes more efficiently.

In conclusion, the incorporation of our MTL framework facilitates the access of MSMEs to the digitalization and automation of their processes, to increase their competitiveness and profitability.

5 Discussion

5.1. General results

The MTL framework has some distinctive features, such as the ability to reuse models generated with historical coffee data. In addition, it could adjust to the introduction of new variables, which gives it notable flexibility and the ability to adapt to different contexts. MTL improves learning capabilities according to the new data presented to it. In other words, our MLT framework is capable of learning how to learn.

Additionally, to test the results of our different transfer learning approaches (Model TL, Parameters_T and Data_T), we defined the following hypotheses [22], [28]:

$$H0: \bar{\gamma}_{TL} = \gamma_{no-TL}$$

$$H1: \bar{\gamma}_{TL} > \gamma_{no-TL}$$

where $\bar{\gamma}_{TL}$ is the R2 mean of 1000 runs on the testing set using the different transfer learnings for the previous case study and γ_{no-TL} corresponds to the R2 mean of the model without the application of transfer learning. To test the hypotheses, the student's t test with superior tail alternative and 95% confidence was used.

The average values and standard deviation of the evaluation metrics of the models are shown in Table 16. Furthermore, the results of the statistical test are detailed to evaluate whether the application of TL contributes to an improvement in the performance of the model compared to the model that does not use TL. According to the results obtained, the use of the TL approach leads to a significant improvement in model performance, in contrast to the model that does not incorporate TL (p value = <0.001).

Table 16. Descriptive statistics and results of Student's t test (p value) to determine performance improvement using TL.

Metrics	$\bar{\gamma}_{TL}$	σ_{TL}	γ_{no-TL}	**Valor p**
Accuracy	0,89	0,007	0,770	< 0.001
F1-Score	0,87	0,007	0,810	< 0.001

In summary, the results showed an increase in quality from 0.77 to 0.89.

5.2. Qualitative Comparison with previous work

In this section, we propose several criteria to contrast our proposal with similar previous studies.

- Criterion 1: The work is for ACODAT.
- Criterion 2: The work implements modeling processes that adjust adaptively using a Meta-Learning process.

- Criterion 3: The work proposes a distributed learning approach.
- Criterion 4: The work focuses on data-driven optimization in agriculture.
- Criterion 5: The work presents an application of AI in MSMEs.

Table 17 shows a comparative qualitative evaluation with related studies, based on the criteria mentioned above.

Table 17: Qualitative Comparison with previous work.

	Cri1	Cri2	Cri3	Cri4	Cri5
[11]	√	x	x	x	x
[15]	x	√	x	x	x
[17]	x	√	x	x	x
[18]	x	√	x	x	x
[28]	√	x	x	√	√
[31]	x	√	√	x	x
[32]	x	x	x	√	x
[33]	x	x	x	x	√
[35]	x	√	√	x	x
[37]	x	x	x	√	x
[38]	x	x	x	√	√
This work	√	√	x	√	√

For criterion 1, work [28] specified ACODAT for a smart classroom. Furthermore, in the work [11], autonomous cycles of collaborative processes are used for Integration Based on Industry 4.0. Now, none of them propose meta-learning processes. Our research implements ACODAT, and the transferability of the model is shown in its ability to adapt and generalize to new tasks with precision. Our approach considers the possibility of retraining models with new data that may be associated with cycle tasks, among other meta-learning aspects, with the aim of integrating knowledge and experience into the original model.

For criterion 2, our work, [15], [17 and [18] are the only one that meets. This criterion becomes important because the inclusion of adaptive modeling methods constitutes an essential element in any project that involves the application of machine learning. Adaptation processes in modeling allow the model to continually adjust and improve based on new data or changes in the conditions of the system being modeled. The work [15] is presented as a viable option for implementation in real-time facility management systems within smart buildings, with applications ranging from early detection of equipment failures to automated monitoring and optimization of energy use. Tavares et al. [17] used a meta-learning strategy to improve the detection of anomalous traces in event logs in order to achieve the best discriminatory ability between common and irregular traces. The method extracts meta-features from an event log and recommends the most suitable encoding technique to increase anomaly detection. Garouani et al. [18] presented a meta-learning-based approach to automate predictive ML models built from industrial big data. In our approach, the introduction of an autonomous cycle gives the models the ability to adapt through the possibility of monitoring and adjusting the models.

For criterion 3, Lin et al. [35] explored the fusion of federated learning and meta-learning principles and defined future research trajectories. They specifically discuss federated meta-learning viewed as collaborative learning between edge nodes, or between the edge and the cloud, with the goal of taking full advantage of knowledge transfer. Liu et al. [31] addressed challenges associated with model customization in distributed learning systems, such as the absence of centralized data and device diversity. Their proposal enables machine learning models to be efficiently fine-tuned to new data on local devices while preserving privacy by using federated learning. This process is facilitated by collaboratively creating a global model, which is customized for each device using meta-learning methods. Our approach in particular does not allow for a distributed learning process.

For criterion 4, Muchen et al. [32] investigated how data analytics and technology impact the optimization of agricultural practices. Their study focuses on the application of advanced tools such as machine learning models, and big data to improve the sustainability, efficiency, and productivity of the agricultural sector. Huo et al.'s work analysed research work on IoT-based smart agriculture. They present architectures, devices and communication technologies used in IoTn smart agriculture. The work of Mishra et al. [37] described the advances in smart agriculture, its main components linked to sensing and monitoring, data collection, communication technologies and decision support systems. They describe the use of artificial intelligence to monitor crops, identify pests and diseases and forecast yields. The authors of [28] presented an application in the coffee context of the use of prediction models based on machine learning techniques. The work [38] presented several projects about artificial intelligence in agriculture. Our work is an example of the application of AI in agricultural environments to automate the management of the production process by exploiting their data.

Finally, for criterion 5, Linaza et al. [38] described a summary of research projects in the use of artificial intelligence in agriculture. In particular, they emphasize the European case, where there are quite a few MSMEs. They conclude that in that sector they are in the early stages in the use of AI technologies, mainly in monitoring and production optimization tasks. Met et al. [33] presented an intelligent system that improves financial product recommendations for MSMEs. This system makes use of advanced machine learning algorithms to process and analyze massive data related to the financial behavior and business needs of MSMEs. On the other hand, the prediction models proposed in [28] can be used by coffee MSMEs. Finally, our proposal is specifically for coffee MSMEs, but our proposal can be extended to any industrial sector, not only agricultural.

From the perspective of MSMEs, an approach like ours would give them competitive advantages by giving them: i) greater flexibility and adaptability to market changes, ii) closer relationships with customers by being able to translate their needs into personalized products or services based on the data collected, iii) reduced operating costs by optimizing their production, and iv) constant innovation guided by their data and that of their context. The most important possible cons or limitations that SMEs could have in using an architecture like this are i) the shortage of specialized human resources to manage the technological platform, ii) the initial investment required to install the infrastructure. The basic difference between large companies and SMEs is in their scale and the resources available, which allows them to overcome the limitations, but eventually due to their organizational structure some of the advantages would be lost.

Our research is the only one that proposes an autonomous process that incorporates adaptive machine learning modeling processes, which makes it easier for the model to transform and improve over time by including new data or changes in the underlying system. In addition, in our proposal, the knowledge acquired when building the models can be transferred to new tasks or new autonomous cycles. These characteristics make it useful for a variety of applications. In particular, in the coffee area to share knowledge between MSMEs that have different coffee beans, to exploit quality models, loss detection (quantity), among others.

5.3 Quantitative Comparison with previous work

In order to make a quantitative comparison of our architecture with previous works, we will contextualize our architecture in the same contexts of previous works, and we calculate the quality of the machine learning models produced with our approach and the one reported in those works using their same metrics (accuracy and silhouette index (SSW)). In addition, we add two features, the generality of the approach to be able to be applied in different contexts, and the scalability to be able to be used on large-scale data, or the opposite, with little data. Table 18 shows the quantitative comparison.

Table 18: Quantitative Comparison with previous work

Work	Generalization	Scalability	Accuracy	SSW	Our approach
[11]	Mediu	Medium	0.87	0.68	0.90 (0.78)
[15]	Medium	Medium	0.85	-	0.83
[17]	Medium	Medium	0.77	-	0.79
[18]	Medium	High	0.80	-	0.80
[28]	Low	Medium	0.61	-	0.78
This work	**High**	**High**	**-**	**-**	**-**

In terms of generalization, approaches without meta-learning show limited generalization, as they build static models [28]. Works that feature the use of meta-learning [15], [17], [18], same as approaches that use autonomic computing loops [11], show moderate generalization thanks to their adaptive capacity. Our model, by integrating meta-learning with ACODAT, achieves more robust generalization.

Regarding scalability, rigid approaches such as [11], [15], [17], [28] have moderate scalability, for very specific datasets (or domains). On the other hand, [18] is highly scalable thanks to its massive data analysis strategies. Our approach maximizes scalability by combining ACODAT and meta-learning, allowing dynamic updating of models without the need for complete retraining.

As for the quality of the results, our approach maintains the results obtained with the meta-learning architectures. Particularly, the work [15] achieves a high value of 85% due to the use of advanced autoencoders, and our architecture follows it quite well. In our case, we manage to maintain its results thanks to the combination of incremental learning and knowledge transfer. Now, in the case of those without adaptation mechanisms based on meta-learning, our architecture clearly improves them. Even in [11] that develops a clustering model (0.68) and a classification model (0.87), our architecture manages to perform better for both cases (0.90 and 0.78, respectively).

6. Conclusions

Meta learning is a huge need in the world of machine learning, and it can be exploited in a wide variety of ways. In any learning process, whether successful or not, useful knowledge accumulates that can be used in new learning processes. It is not necessary to start from zero every time; Instead, previous experiences should be taken advantage of. The framework proposed in this work allows this, giving different options for using knowledge, from already built knowledge models to the reuse of parameters of techniques or data.

Our MTL architecture for ACODAT allows self-control and self-management processes to reuse autonomous cycles of automation of an agro-industrial production chain. Our framework allows reusing previous machine learning models, or data sources or parameters, for new data sources that define new autonomous cycles. Thus, the framework allows the instantiation of an ACODAT in new data sets (productive processes). In turn, the MTL architecture is enriched with new knowledge models, with the best parameters of machine learning techniques, and/or data, as new data is added. All that knowledge (models, techniques and data) is exploited when new data arrives (for example, a new ACODAT instantiation). Consequently, one of the most relevant characteristics of our MTL architecture is its ability to adapt to new data, without forgetting the previous ones. In summary, our framework provides greater adaptability to systems based on machine learning, improves the quality of the models by reusing previously optimized data or parameters, reduces training time, and being based on ACODA allows dynamic adjustments.

Our MTL architecture is the first to be used to determine coffee quality, roasting and processing methods in the coffee industry. The effectiveness of our approach lies in its ability to build new models associated with various types of coffee, thus being able to instantiate the ACODATs for automating the coffee production processes to different qualities and coffee seeds in an efficient and flexible way. Although our MTL architecture has been successfully applied in coffee production, it has a high degree of scalability to other agro-industrial domains. In particular, it can be used to automate the productive processes of the food industry taking into account the following aspects: i) the processes of the production chain to be automated must be clearly identified ii) the availability and quality of the data must be identified iii) the appropriate ACODAT must be specified and implemented for said organizational environment. From there, the framework can be instantiated to begin to autonomously manage the ACODAT instantiations made by companies in the industrial sector under study.

Despite its capabilities, our framework has certain limitations that need to be addressed to improve its effectiveness. The most relevant ones are: The first is the need to collect a significant amount of data to train the initial models in the knowledge base, which can be difficult in certain situations. Secondly, it requires having a significant number of models, techniques and data to make it interesting, otherwise it will have the cold start problem of many recommendation/adaptation systems. On the other hand, it requires a person with expertise in the topics related to ACODAT to be able to implement it in a real environment properly. Despite the limitations, for MSMEs, it is very beneficial because of all the knowledge that they will be able to reuse over time.

In future research, it would be interesting to explore the effectiveness and behavior of various synthetic data generation algorithms within our architecture. In addition, the possibility of implementing a Federated Meta-Learning approach with ACODAT will be considered to distribute the Meta-Knowledge generated in the architecture, which could contribute to reducing both the complexity and the necessary computational resources. Federated learning allows AI models to be trained in a decentralized manner, without the need to transfer information outside each company. This facilitates cooperation between MSMEs without compromising data security. The combination of synthetic data and federated learning helps overcome data scarcity in the agro-industrial industry, ensuring more accurate decision-making without compromising data privacy. Particularly, they address the privacy and cold start issues that are so important in this agro-industrial sector. Finally, future work should analyze the behavior of the autonomous cycle in real-time constrained contexts, which involves testing its functionality in real-world scenarios.

Conflict-of-Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of Data

The data used in this paper will be available upon reasonable request.

Funding Declaration

There was no funding.

Ethics Approval Declaration

This article does not work with humans or animals. All information used had the informed consent of its owners for inclusion in the study. The study was conducted in accordance with the Declaration of Helsinki.

References

[1] Knight S., Gibson A., Shibani A. (2020) Implementing learning analytics for learning impact: Taking tools to task, The Internet and Higher Education, 45.,

[2] Law, N., Liang, L. (2020). A Multilevel Framework and Method for Learning Analytics Integrated Learning Design. Journal of Learning Analytics, 7(3), 98-117. https://doi.org/10.18608/jla.2020.73.8

[3] Amarasinghe, I., Michos, K., Crespi, F., Hernández-Leo, D. (2022). Learning analytics support to teachers' design and orchestrating tasks. Journal of Computer Assisted Learning,

[4] Lanus E.*et al*., 2021 "Test and Evaluation Framework for Multi-Agent Systems of Autonomous Intelligent Agents," *Proc. 16th International Conference of System of Systems Engineering (SoSE)*, pp. 203-209,

[5] Kephart J., Chess D. (2003) “The vision of autonomic computing,” Computer, 36, 41–50, doi: 10.1109/MC.2003.1160055.

[6] IBM, “An architectural blueprint for autonomic computing,” IBM White Paper, vol. 36, no. June, p. 34, 2006, [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0e99837d9b1e70bb35d516e32ecfc345cd30e795.

[7] Jiang X., Yu F., Song T., Leung V. (2022) "Intelligent Resource Allocation for Video Analytics in Blockchain-Enabled Internet of Autonomous Vehicles With Edge Computing," *IEEE Internet of Things Journal*, 9 (16), 14260-14272.

[8] Monsalve-Pulido, J., Aguilar, J., Montoya, E. and Salazar, C. (2024), "Autonomous recommender system architecture for virtual learning environments", Applied Computing and Informatics, 20(1/2), 69-88. https://doi.org/10.1016/j.aci.2020.03.001

[9] Aguilar J., Salazar C., Velasco H, Monsalve-Pulido J, Montoya E. (2020). "Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents" *Computation* 8, ttps://doi.org/10.3390/computation8020030).

[10] Athavale J., Baldovin A., Graefe R., Paulitsch M. Rosales r. (2020) "AI and Reliability Trends in Safety-Critical Autonomous Systems on Ground and Air," *Proc. 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pp. 74-77,

[11] Morales L., Ouedraogo C., Aguilar J., Chassot C., Medjiah S., Drira K. (2019) "Experimental Comparison of the Diagnostic Capabilities of Classification and Clustering Algorithms for the QoS Management in an Autonomic IoT Platform", Service Oriented Computing and Applications, 13 (3), 199–219.

[12] Braten A. Kraemer F., Palma D, (2021) "Autonomous IoT Device Management Systems: Structured Review and Generalized Cognitive Model," *IEEE Internet of Things Journal*, 8 (6), 4275-4290.

[13] Sánchez M., Exposito E., Aguilar, J. (2020) Implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context, Computers in Industry, 121,doi: 10.1016/j.compind.2020.103247.

[14] Pan, J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22 (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[15] Moon J., Noh Y., Jung S., Lee J., Hwang E. (2023) “Anomaly detection using a model-agnostic meta learning-based variational auto-encoder for facility management,” Journal of Building Engineering, 68.

[16] Moulavi D., Jaskowiak P., Campello R. Zimek A. Sander J. (2014) Density-Based Clustering Validation Proc. 14th SIAM International Conference on Data Mining.

[17] Tavares G, Junior S (2021), “Process mining encoding via meta-learning for an enhanced anomaly detection,” Communications in Computer and Information Science, 1450, 157 – 168, Springer.

[18] Garouani M., Ahmad A., Bouneffa M, Hamlich M, Bourguin G. Lewandowski A. et al. (2022). “Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data”. Journal of Big Data, https://doi.org/10.1186/s40537-022-00612-4.

[19] Porras-Zúñiga M., Vargas-Elías G., Araúz- Madrid L., Abarca-Alpízar N. (2019). Effect of temperature on the speed of coffee roasting. Technology in March Magazine, 32, 20–27. https://doi.org/10.18845/tm.v32i7.4255.

[20] Toscano-Miranda R., Aguilar J., Hoyos W., Caro M., Trebilcok A., Toro M. (2024) Different transfer learning approaches for insect pest classification in cotton, Applied Soft Computing, 153.

[21] Pacheco, F., Rangel, C., Aguilar, J., Cerrada, M., & Altamiranda, J. “Methodological Framework for Data Processing based on the Data Science

Paradigm”. Proc. XL Latin American Computing Conference. 2014, https://doi.org/10.1109/CLEI.2014.6965184.

[22] Aguilar, J., Cerrada, M., Mousalli, G., Rivas, F., Hidrobo, F. (2005). A Multiagent Model for Intelligent Distributed Control Systems. Lecture Notes in Computer Science, 3681, 191–197, doi: 10.1007/11552413_28

[23] Aguilar, J., Cerrada, M., Hidrobo, F. (2007). A Methodology to Specify Multiagent Systems. Lecture Notes in Computer Science, 4496. d 92–101, doi: 10.1007/978-3-540-72830-6_10

[24] Aguilar, J. (2001). A general ant colony model to solve combinatorial optimization problems. Revista Colombiana De Computación, 2(1), 7–18.

[25] Fuentes J, Aguilar J, Montoya E. (2025) “Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs”. Journal of Industrial Information Integration, 44.

[26] Majid A., Abdullah E, Diana, A. (2015). “A Review of Quality Coffee Roasting Degree Evaluation. Journal of Applied Science and Agriculture”, 10(7), 18–23.

[27] Zhang, Y., Wei, S., Zhang, L., & Liu, C. (2019). Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features. Journal of Medical and Biological Engineering, 39(3), 381–392. https://doi.org/10.1007/s40846-018-0411-0.

[28] Kim, J. (2022). “Coffee Beans Quality Prediction Using Machine Learning”. SSRN Electronic Journal, https://doi.org/10.2139/ssrn.4024785.

[29] Lake B., Ullman T., Tenenbaum J., Gershman S. (2017) “Building machines that learn and think like people”, Behavioral and brain sciences, 40.

[30] Yao H., Wu X., Tao Z., Li Y., Ding B., Li R., Li Z. (2020) Automated relational meta-learning”, Proc. International Conference on Learning Representations.

[31] X. Liu, Y. Deng, A. Nallanathan, and M. Bennis, (2023) Federated Learning and Meta Learning: Approaches, Applications, and Directions, IEEE Communications Surveys and Tutorials, 26(1), 571–618, doi: 10.1109/COMST.2023.3330910.

[32] L. Muchen, R. Hamdan, and R. Ab-Rahim (2022) “Data-Driven Evaluation and Optimization of Agricultural Environmental Efficiency with Carbon Emission Constraints,” Sustainability (Switzerland), 14(19), doi: 10.3390/su141911849.

[33] I. Met, A. Erkoc, S. E. Seker, M. A. Erturk, B. Ulug (2024) “Algorithms for SME Banking”, International Journal of Intelligent Systems, doi: 10.1155/2024/5585575.

[34] J. Gama, P. Kosina, (2009) “Tracking recurring concepts with meta-learners,” Lecture Notes in in Artificial Intelligence, 5816, 423–434, doi: 10.1007/978-3-642-04686-5_35.

[35] S. Lin, Z. Zhou, Z. Zhang, X. Chen, J. Zhang, J. (2021). “Edge Intelligence via Federated Meta-Learning”. In: Edge Intelligence in the Making. Synthesis Lectures on Learning, Networks, and Algorithms. pp 53–79, Springer, doi: 10.1007/978-3-031-02380-4_3

[36] Huo D, Malik A, Ravana S., Rahman A, Ahmedy (2024), “Mapping smart farming: Addressing agricultural challenges in data-driven era“, Renewable and Sustainable Energy Reviews, 189, doi: 10.1016/j.rser.2023.113858.

[37] Mishra, H., Mishra, D. (2024). AI for Data-Driven Decision-Making in Smart Agriculture: From Field to Farm Management. In: Chouhan, S., Saxena, A., Singh, U.P., Jain, S. (eds) Artificial Intelligence Techniques in Smart Agriculture. Springer, pp 173–193. Doi: 10.1007/978-981-97-5878-4_11

[38] Linaza, M. T., Posada, J., Bund, J., Eisert, P., Quartulli, M., Döllner, J., Pagani, A., G. Olaizola, I., Barriguinha, A., Moysiadis, T., Lucat, L. (2021). Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture. *Agronomy*, *11*(6), doi: 10.3390/agronomy11061227

[40] Al Dallal J. Alkhazi B.(2024) Exploring the Impact of Alternatives of Object-Oriented Cohesion Measures on Machine Learning-Based Predictions of Inheritance Reusability, *IEEE Access*, 12, 159252-159266, doi: 10.1109/ACCESS.2024.3484007.

[40] Ogundimu E., Altman D, Collins G. (2016) Adequate sample size for developing prediction models is not simply related to events per variable, Journal of Clinical Epidemiology, 76, 175-182, doi: 10.1016/j.jclinepi.2016.02.031.

[41] Yang, Li (2020). "On hyperparameter optimization of machine learning algorithms: Theory and practice". Neurocomputing. 415, 295–316.

.

Appendix H. Meta-learning Architecture for ACODAT in the Context of Agro-Industrial Production Chains of MSMEs.

Meta-learning Architecture for ACODAT in the Context of Agro-Industrial Production Chains of MSMEs

Jairo Fuentes [1,2], Ricardo Dos Santos [3], Jose Aguilar [2,3,4,*] Donghui Shi [5]
[1] Universidad Francisco de Paula Santander, Cúcuta, Colombia
[2] GCOMP, Universidad EAFIT, Medellín, Colombia
[3] CEMISID, Universidad de Los Andes, Mérida, Venezuela
[4] IMDEA Networks Institute, Madrid, Spain
[5] Dept of Comp. Eng. School of Electronics and Information Eng., Anhui Jianzhu Univ., Hefei, China

Abstract. In the MSME agribusiness sector, constant monitoring and detailed analysis of product characteristics are essential to identify areas where production can be improved. The proposal of this work is a meta-learning (MTL) architecture integrated into autonomous cycles of data analysis tasks (ACODATs) that automate the production process of MSMEs to allow their self-adjustment. The proposed MTL framework includes an ACODAT to guide the entire meta-learning process and a meta-model knowledge base that stores information about existing machine learning models, the machine learning techniques used to create them, and with which dataset they were made. Thus, the architecture has an ACODAT module to guide the behavior of the MTL framework, and meta-data and meta-techniques that describe previous datasets and techniques used to build the models described in the metamodel. These are used to select the most appropriate machine learning model for each ACODAT task of the automation process when it is instantiated on a new dataset. Particularly, from the information contained in the MTL framework, the system is capable of performing model transfer, data transfer or parameter transfer processes, or invoking procedures for the generation of systematic data, in order to define the new machine learning models for the tasks of the new ACODAT of the automation process. The framework has been tested in an MSME in the coffee area. In the context of the case study, an ACODAT for a coffee company composed of three tasks to control its production process is instantiated. The proposed architecture demonstrated its ability to adapt the ACODAT to a new scenario, obtaining very satisfactory results for each task.

Keywords: meta-learning, autonomous cycle of data analysis tasks, machine learning, transfer learning, automation process.

1. Introduction

Meta learning, also called “learning to learn,” is a subcategory of machine learning that defines artificial intelligence (AI) architectures to understand and adapt machine learning models to new tasks on their own [1, 2, 3]. Meta learning’s primary aim is to provide machines with the skill to learn how to learn [4, 5, 19]. Meta-learning is useful in an agro-industrial production chain in contexts where adaptability, efficiency and optimization are needed at different stages of the process [30]. The use of meta-learning in a production chain allows responding to the unpredictable dynamics of the environment and improves both productivity and sustainability [31, 32].

On the other hand, in [6. 7], the concept of autonomous cycles of data analysis tasks (ACODAT) was introduced to automate processes in an environment, and applied to the management of smart classrooms, particularly, for the management of learning processes [8, 11]. In [13, 23, 29], the authors proposed using these cycles in collaborative industrial production processes within the framework of Industry 4.0. In turn, the authors of [19] used MTL Architecture for the selection and configuration of automated learning algorithms, and in [15], for the management of data transfer processes. Despite the widespread use of ACODAT in different domains, there are no applications that allow models that make up an ACODAT to learn how to learn from context, to follow the dynamics of the environment. Thus, the application of MTL as a layer that allows automating the process of automatic adaptation of the models that make up an ACODAT is a great challenge.

In turn, typically, in many countries, coffee producers are MSMEs. Specifically, coffee production must operate at certain levels of efficiency to ensure that the quality of the beans, the roasting process, and the general methods of coffee processing meet financial and environmental sustainability objectives. This will allow actors in the MSME agro-industrial sector to be aware of what is happening in coffee production to implement strategies for its improvement. Previous works have proposed ACODATs to automate the agro-industrial production processes of MSMEs in the coffee sector [26, 28]. However, these previous works have not proposed an MTL framework for ACODAT.

In this paper, we present a novel MTL framework, which aims to automate the management of machine learning models in an ACODAT. In particular, the framework allows selecting pre-existing machine learning models to reuse them (which we will call model transfer), or to reuse their parameters (which we will call parameter transfer), but also, eventually, to reuse the data with which these models were built (which we will call data transfer), or simply to invoke synthetic data generation algorithms. All the above is guided by the framework according to the similarity between the input dataset with respect to the datasets with which the previous models were built in previous ACODATs instantiated in specific contexts. Thus, the MTL framework seeks to reuse machine learning models built with previous datasets in other ACODATs in new datasets in which a given ACODAT is instantiated. The main contributions of this work are:

- An MTL architecture for ACODAT, which allows it to adapt to different contexts automatically by reusing/adapting the machine learning models that define its tasks to new datasets.
- An ACODAT module to guide the behavior of the MTL framework.
- An MTL architecture in the context of the coffee industry, for the automation of the production process of MSMEs.

The organization of the work is as follows: Section 2 introduces the theoretical framework, with related works. Subsequently, Section 3 describes the proposed MTL architecture. Section 4 presents the case study for the coffee sector. Then, Section 5 performs a comparative Analysis with other works (qualitative and quantitative); and finally, Section 6 presents the conclusions and future works.

2. Theoretical Framework

2.1. Meta Learning

The concept of meta-learning in artificial intelligence refers to the ability of a system to continuously improve its learning process. Instead of focusing only on a specific task, the model seeks to adapt and learn new tasks more quickly and efficiently [1, 2, 4, 5, 22]. The goal is to capture knowledge or metadata that helps optimize the performance of these models in new applications [2, 3, 19]. MTL aims to understand and use past experiences to improve its ability to adapt to new tasks. By learning from previous problems, the model can identify useful patterns or strategies, allowing it to optimize its future learning process and solve new tasks more quickly and efficiently. In general, detailed knowledge of the context is needed to properly tune machine learning algorithms to specific problems. This starts with collecting metadata describing the behavior of pre-trained models, including details of algorithm configurations such as hyperparameters, quality metrics, training times, and characteristics of datasets, known as meta-attributes. This metadata is then used to reuse this knowledge in new contexts and emerging tasks [1, 2, 25].

This allows for different types of reuses of that acquired knowledge. For example, these models can be used on new datasets if they exhibit similar behaviors; hyperparameters found in previous studies with these techniques can be reused to build new models if the similarity between datasets is lower; or data can be transferred if they are not so similar and the new set lacks sufficient information. In the first case, the model's metadata must be adjusted to the new dataset; in the last case, the new machine learning model becomes a new input for the metamodel. In this way, the reuse of previous knowledge and the learning of new uses for already learned models are facilitated, among other benefits [26, 27]. Thus, MTL allows setting up new data analysis tasks at a much faster rate. This is possible because it greatly accelerates the creation of new learning models reusing knowledge [16, 18].

Particularly, in [16] an unsupervised anomaly detection method, called MAVAE, is proposed for time series sensor data. The proposed method uses a variational autoencoder (VAE) as an anomaly detection model and adapts the model to a new target task with few unlabeled anomaly data using an MTL approach. In [18], Tavares et al. use an MTL strategy to improve the detection of anomaly traces in event logs to fine-tune the best discriminative ability between common and irregular traces. In [19], Garouani et al. present an MTL-based approach to automate the construction of predictive machine learning models in the context of industrial big data. The approach takes advantage of the AMLBID framework, an automated machine learning tool for industrial big data analysis with meta-features. Furthermore, in [27], Xu et al. present an adaptive MTL model based on user relevance for improving user cold-start recommendations. Yao et al. [25] propose an automated relational MTL framework that automatically extracts the relationships between tasks and builds a meta-knowledge graph. When a new task arrives, it can quickly find the most relevant structure and adapt the knowledge of the learned structure. In turn, Garcia et al. [20] defined an MTL architecture to adapt predictive models. Thus, all these works propose a set of characteristics to address the MTL. However, none of these works presents proposals aimed at carrying out processes of adaptation of learning models in contexts where ACODATs are used as process automation mechanisms, the subject of this work.

2.2. ACODAT

Given the value of innovation in MSMEs and the current opportunities to leverage organizational data and its environment, it is possible to develop data-driven strategies that guide the automation process. One such strategy includes the implementation of autonomous cycles of data analysis tasks (ACODATs), proposed in previous research [26, 28], which allow the generation of machine learning models using various sources of information to facilitate the automated management of their production processes. Multiple data analysis tasks need to be performed at the same time to achieve a specific goal in the observed process. These tasks are interrelated and each performs an essential function in the cycle: (i) observing the process, (ii) analyzing and interpreting events, and (iii) making decisions that guide the productive process toward its objectives. The integration of these tasks in a closed loop enables continuous and autonomous resolution of complex problems. The functions of each task are detailed below [6, 7]:

- **Observation task:** involves a series of tasks that are responsible for monitoring the process. These tasks must capture data and information about the behavior of the environment and prepare that data for analysis.
- **Analysis tasks:** are tasks to interpret and understand the collected data to diagnose what is happening in the monitored context. This includes the creation of knowledge models (machine learning models) based on the observed dynamics.
- **Decision-making task:** These tasks involve the implementation of actions aimed at decision-making to improve the process in which the cycle is implemented. The actions influence the dynamics of the process to optimize it, and the results are reviewed again in the observation and analysis stages, beginning a new iteration of the cycle.

ACODAT has been implemented in past research, following the principles that IBM introduced in 2001 [9, 10, 14]. Its applications include areas such as smart classrooms [8, 11], telecommunications [12], Industry 4.0 [6, 13, 24] and smart cities [7]. It adopts an autonomic computing approach with the aim of delivering systems that operate with intelligent control loops and autonomous capabilities. In general, an ACODAT requires a multidimensional data model to store the collected data, which describes the behavior of the context, and the machine learning models generated with said data.

Although ACODAT has proven its effectiveness in various fields, it has not yet been expanded with MTL processes to improve its capabilities to adapt autonomously to the characteristics of different contexts.

3. ACODAT MTL Framework.

In this work, we introduce a new MTL architecture for the automatic creation of knowledge models for ACODAT. The MTL mechanism helps ACODAT to learn to adapt quickly to new scenarios. The MTL architecture requires extensive and detailed knowledge of several aspects of machine learning, including data sources, meta-features, and already defined meta-models. A description of the architecture is presented in this section.

3.1. Conceptual Architecture

Figure 1 shows the general structure of the proposed framework. This architecture uses a knowledge base that includes information about the knowledge models previously stored in it, and the datasets and the hyperparameters of the techniques used to build them. When a new model needs to be built for a new dataset, the system compares the similarity of the new dataset with the existing datasets in the framework to decide the procedure to follow. In particular, the options are to use an existing model if the new dataset is very similar to the one used to build the model, to use only the parameters if they are somewhat similar, and to build a new model or generate synthetic data if they are not very similar. Thus, this architecture allows reusing previous knowledge and integrating new knowledge. The modules of the architecture are briefly described below:

- **Management Module:** This module manages the knowledge base, which consists of a metamodel table, a metadata table, and a meta technique table. Each knowledge model in the metamodel is linked to a dataset (in the metadata) and to the machine learning technique (in the Meta Techniques) used to create it. In addition, the metamodel stores the model's quality metrics and other relevant data. The metadata contains details about the datasets, such as their attributes and location. Meta Techniques store information about machine learning techniques, including the optimal values of their parameters.

- **Selection Module:** When a new requirement to build a knowledge model on a dataset is received, then the selection module is activated. It first extracts features from the incoming dataset. It then compares the new dataset with the previous datasets that have been previously used to build previous knowledge models, to generate a similarity ranking.
- **Adaptation Module:** The adaptation module makes decisions based on the similarity ranking. Based on this ranking, the module can decide several things: perform a model transfer (reuse it) if they are very similar, perform a parameter transfer (reuse the parameters of the techniques) if they are somewhat similar, perform a data transfer (if they are not very similar), or generate synthetic data (in these last two cases, if the new dataset is missing data). In all cases, a new model is built with the new dataset, and the management module is invoked.

The selection module uses an ACODAT to guide the process of optimizing the creation of the ACODAT models being instantiated, ensuring the selection of the resources (models, techniques or data, according to the type of transfer to be performed) to adapt the ACODAT. In the following section, the operation of the ACODAT module and its impact on the general architecture will be detailed.

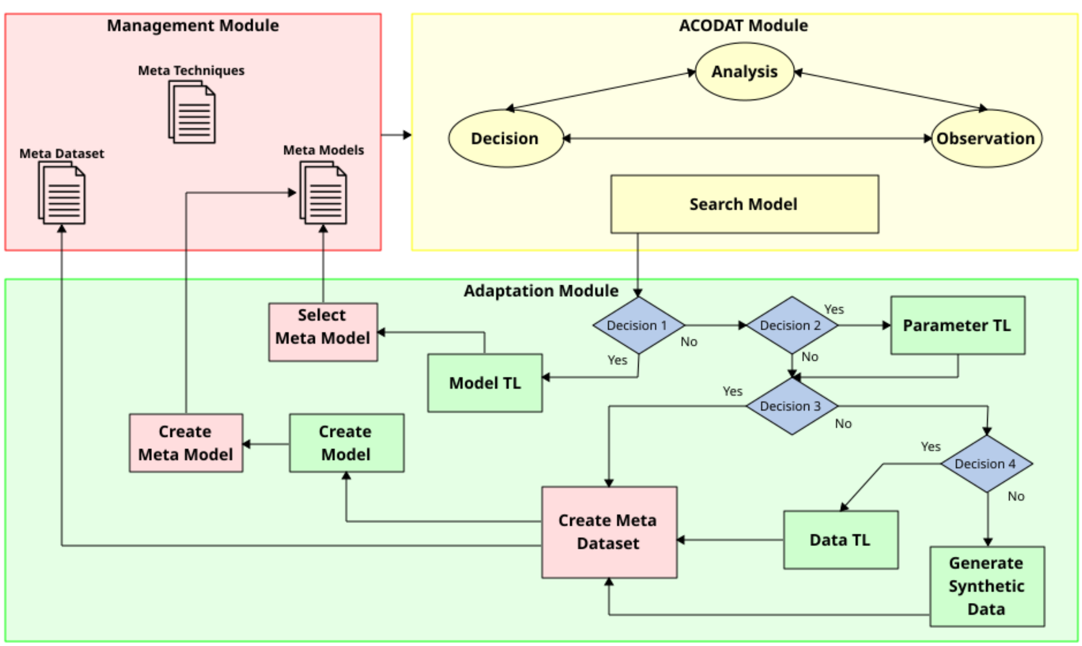

Figure 1. New MTL architecture ACODAT.

3.2. ACODAT Module

The ACODAT module is responsible for supervising the execution of the other modules of the framework, but also, the ACODAT that is being adapted. When this module is activated, a macro algorithm is executed that starts with the *Observation task*, which monitors and collects data and information from the system or environment being monitored (information about the meta-datasets, meta-techniques and meta-models), but also about the ACODAT being supervised (tasks being executed, techniques and datasets used, etc.). Then, the *Analysis task* carries out the processes aimed at interpreting and understanding the collected data to diagnose what is happening in the monitored context, especially, on the adaptation needs at the level of each task of the supervised ACODAT. Finally, the *Decision task* is activated, which in this case involves actions aimed at building the learning models for the ACODAT being instantiated. In this case, it invokes the *adaptation module* to determine what type of transfer (of models, data, among others) to perform for each task of the supervised ACODAT. Finally, this module, being an autonomous cycle, observes the behavior of the ACODAT being instantiated to optimize it (the results of the learning models are reviewed again in the Observation and Analysis tasks, starting a new iteration of the cycle).

4. Case Study

This section presents a real case study showing how the architecture works

4.1. Experimental Context

ACODAT MTL architecture was implemented in the Café Galavis company, located in Cúcuta, Colombia. This company, founded in 1918, is dedicated to the roasting and distribution of Colombian coffee, and has sophisticated equipment for the entire process, from receiving the raw material to its storage under optimal temperature and humidity conditions. In addition, they carry out a rigorous selection of the grain, purify it before processing, and ensure the strictest hygienic conditions for its internal transport to the packaging machines. Specifically, the ACODAT that was designed to automate the production process of this company is composed of the following tasks (see Figure 2):

- Task 1. Quantity of input to transform: It analyses various aspects such as the historical evolution and uses of the raw material to generate the product (the seed, in this case). It requires a diagnostic model to determine the raw material and the quantity to be transformed.
- Task 2. Quality of inputs to process: The quality of these inputs is established through this task, depending on factors such as cultural practices and storage and transportation services. In this task, a predictive model is used to determine this quality.
- Task 3. Method of processing to use: The identification of factors related to the processing method of the coffee, such as natural drying, natural washing, mechanical drying, and automatic or manual selection, is the main function of this task. In this task, a classification model is used for this purpose.

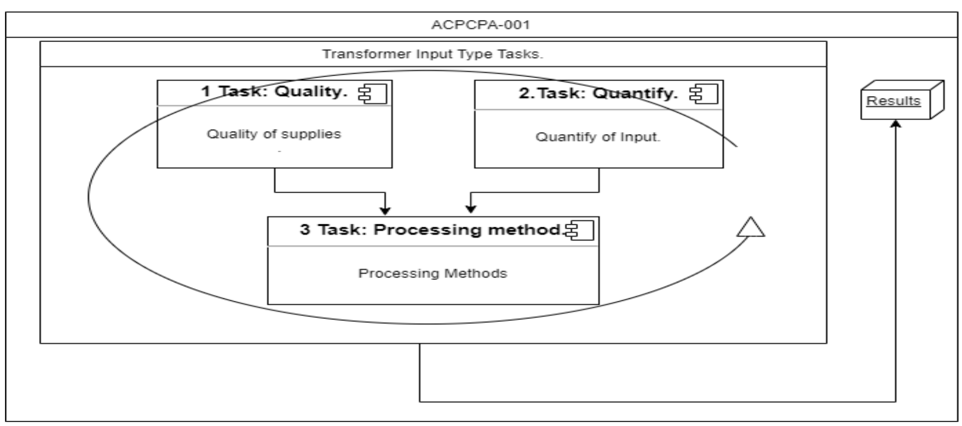

Figure 2. ACPCPA - 001 tasks.

The datasets used to solve the different tasks are as follows:

- **File1.csv:** This dataset is composed of 1,286 records and 56 characteristics (processing method, seed variety, Aroma, flavour, acidity, body, uniformity, sweetness, and moisture, and other variables) [23]. The data used includes expert evaluations of different coffee blends, assigning a rating to each cup tested. Each entry provides detailed information about the origin of the coffee, the recipe applied, the characteristics of the raw materials, and the final product [23].

- **File2.csv:** This dataset is composed of 2,032 records and 22 characteristics (green weight, roasted weight, shrinkage, environment humidity, density, color, and other variables) [21]. From this set, the weight reduction during roasting (target variable) will be calculated using the formula: reduction = Roast weight / Green weight. The analysis of the data shows a weight reduction ranging from 7.6% to 19.87%, with 75% of the records showing a reduction of more than 14.5%. Half of the data is concentrated in a range between 14.5% and 15.8%.

4.2. Current state of the MTL Architecture

This section describes the information that the MTL framework currently has, particularly its tables containing the MetaModels, Metadata and Meta_Techniques, because this information is used by the MTL modules in their tasks to adapt the ACODAT to the new datasets of the Café Galavis company. Table 1 has the following datasets in the Meta-Datasets Table.

Table 1. Current Datasets in the Meta-Dataset Table of the MTL framework

Id_MD	Description	Vd1…Vdn	Location	Date Created
Dat_01	Coffe Colcafe	aroma, acidity, bitterness, overall impression and body,	https://colcafe.com/	2015
Dat_02	Café Premium De Colombia S A S	flavor, aroma, consistency, and acidity	www.cafepremium.com	2022

Table 2 has the following models in the Meta-Models Table.

Table 2. Current Models in the Meta-Models Table of the MTL framework

Id_MM	Description	Id_MT	P1, P2, Pn	VD1 …VDn	Quality (QM)	Id_MD
MM_01	Predictive	MT_04	**learning rate=0.3**	flavor, aroma, consistency, and acidity	R2=0.4, RMSE=2.1	Dat_01
MM_02	Predictive	MT_02	number of hidden layers=3, number of nodes in each layer=4, convergence error=0.2	flavor, aroma, consistency, and acidity	R2=0.5, RMSE=31.5	Dat_01
MM_03	Predictive	MT_03	**learning rate=0.2,** depth=5, rank=Gini,	flavor, aroma, consistency, and acidity	R2=0.1, RMSE=24,1	Dat_01
MM_04	Predictive	MT_01	depth=4, rank=Gini.	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity	R2=0.8, MAPE=0.9,	Dat_02
MM_05	Diagnostic	MT_05	K=2	flavor, aroma, consistency, and acidity.	silhouette index=0.6	Dat_02
MM_06	Diagnostic	MT_07	K=4, Distance technique= Euclidean, distance calculation=max,.	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity.	silhouette index=0.5	Dat_01
MM_07	Diagnostic	MT_06	Epsilon (eps) = 5 and Minimum Points (minPts): 5.	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity.	DBCV =0.9	Dat_01
MM_08	Classification	MT_01	depth=4, rank=Gini.	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity	accuracy=0.9, F1-score=0.7,	Dat_01
MM_09	Classification	MT_03	**learning rate=0.**	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity	accuracy=0.8, presicion=0.6,	Dat_02
MM_10	Classification	MT_08	**C=4 and gamma=0.2**	Aroma, Flavor, Aftertaste,	accuracy=0.9, presicion=0.9	Dat_01

				Acidity, Body, Balance, Uniformity		
MM_11	Classification	MT_02	number of hidden layers=3, number of nodes in each layer=4, convergence error=0.2	Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity	accuracy=0.6, f1-score=0.5,	Dat_02

Table 3 has the following techniques in the Meta-techniques Table. Specifically, it currently has three unsupervised learning techniques registered: K-means Clustering, Agglomerative Clustering and DBScan (Density-Based Spatial Clustering). In this way, two distance-based clustering techniques and one Density-Based Clustering technique can be used. Additionally, the architecture has registered four regression learning techniques: Gradient Boosting Regressor, Random Forest Regressor, backpropagation neural network, and Linear Regression [22]. Finally, the architecture has four classification techniques registered in its meta-techniques Table: Random Forest Classifier, Logistic Regression, SVM - Linear Kernel, and backpropagation neural network.

Table 3. Current Techniques in the Meta-Techniques Table of the MTL framework

Id_MT	**Description**	**P1, P2, Pn**
MT_01	Random Forest	depth, rank=Gini.
MT_02	Backpropagation	number of hidden layers, number of nodes in each layer, convergence error
MT_03	Gradient Boosting Regressor	**learning rate,** depth, rank
MT_04	Linear Regression	**learning rate**
MT_05	K-means	K.
MT_06	DBSCan	Epsilon and Minimum Points
MT_07	Agglomerative Clustering	K, Distance technique, distance calculation,
MT_08	SVM	C and gamma

The information contained in the MTL framework knowledge base (the 3 tables above) will be used to adapt the ACODAT *ACPCPA - 001* for Café Galavis company based on its two datasets File1 and File2.

4.3. ACODAT Module of the MTL Architecture in the Case Study

This module starts by activating its Observation and Analysis tasks. The Observation task collected data and information from the architecture and supervised ACODAT. The Analysis task then analyses the ACODAT to be adapted, specifying the types of data analysis tasks that comprise it, and data sources to be used to adapt the ACODAT, among other things (see Table 4, with the results of this task).

Table 4. Overview of the tasks in the supervised autonomous cycle.

Task	Model Type	Technique Type	Dataset to use
1. Coffee Bean Quality	Diagnostic	Unsupervised	File1.csv: aroma, flavour, acidity, body, uniformity, and other. variables
2. Decrease of the bean in the roasting process.	Predictive	Supervised	File2.csv: green weight, roasted weight, shrinkage, environment humidity, density, color, and other variables.
3. Coffee processing method	Classification	Supervised	File1.csv: processing method, seed variety, Aroma, flavour, acidity, body, uniformity, sweetness, and moisture, and other. variables

A. Task 1. Coffee Bean Quality Model

The objective is to identify the most relevant descriptors to evaluate the quality of coffee. As a result, a dataset is obtained composed of the key attributes: aroma, flavor, aftertaste, acidity, body, balance, and uniformity. A search is conducted for models that meet the specifications listed below (see Figure 3):

- Model Type: Diagnostic
- Technique Type: Unsupervised/Clustering
- Model Description: Coffee Bean Quality
- Dataset Description: The data used includes expert evaluations of different coffee blends, assigning a rating to each cup tested.
- Dataset Location: Cúcuta, Colombia
- Dataset: File1.csv

The MTL framework realizes that the task is specified in Table 2 (of Meta-models) and proceeds to determine if it can use these models. To do this, it verifies that the datasets with which they were built are similar to the new data set (using the data in Table 1). When it realizes that the dataset _01 is very similar, then it proceeds to choose the model according to the technique that gave the best results (using Table 2), in such a way as to perform a *Model Transfer*.

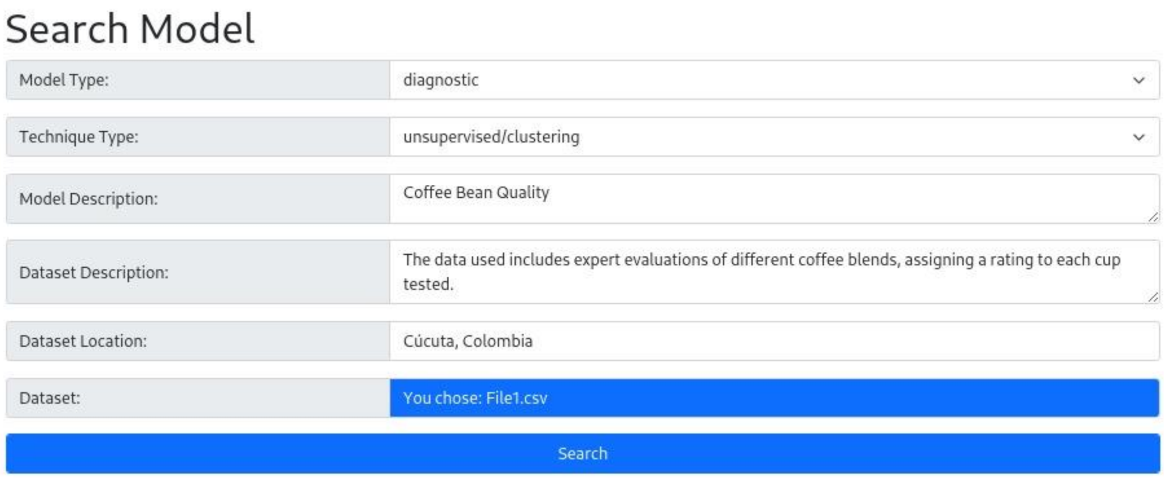

Figure 3. Task 1 on Coffee Bean Quality.

In this way, the adaptation module of the MTL framework verifies the type of transfer to be performed, and it is determined to use the *Transfer model* since the models can be reused (see Fig. 4). Since the architecture currently has two unsupervised learning techniques registered in its meta-techniques (Agglomerative Clustering and DBScan) associated with this dataset (see Table 2), it chooses the one with the best performance, DBScan (with DBCV=0.9), very close to one.

Finally, the ACODAT of the MTL framework verifies the quality of the new model as a result of the adaptation module. It verifies that the DBScan model shows remarkable effectiveness in coffee quality diagnosis tasks based on flavor, aroma, acidity, and other variables.

Result

Traceability

ID	1
DESCRIPTION	Coffee Bean Quality
TYPE	diagnostic
ID_MT	6
TYPE_TECHNIQUE	unsupervised/clustering
PARAMETERS	{"algorithm": "auto", "eps": 0.5, "leaf_size": 30, "metric": "euclidean", "metric_params": null, "min_samples": 5, "n_jobs": -1, "p": null}
METRICS	{"DBCV": 0.9, "Calinski-Harabasz": 2014.7773, "Davies-Bouldin": 0.0723, "Homogeneity": 0, "Rand Index": 0, "Completeness": 0}
ID_MD	1
FEATURES	[["Aroma", "Sabor", "Regusto", "Acidez", "Cuerpo", "Balance", "Uniformidad"], ["float64", "float64", "float64", "float64", "float64", "float64", "float64"]]
TARGETS	

Match

Figure 4. Task 1 result using DBScan.

B. Task 2. Decrease of the bean in the roasting process

In the coffee roasting process, it is common for the beans to expand and lose moisture, which causes a decrease in the final weight of the batch, a phenomenon that has been confirmed by [25]. In this procedure, there are a variety of both product and process variables, some of which can be adjusted to minimize grain loss. The goal of this task of the supervised ACODAT is to create a supervised model that can predict bean reduction during roasting, considering both product attributes (including previously evaluated bean quality) and process variables that can be adjusted. The following are the specifications for the search of the models for this task:

- Model Type: Predictive
- Technique Type: Supervised/Regression
- Model Description: Decrease of the bean in the roasting process
- Dataset Description: The data used includes the coffee roasting process.
- Dataset Location: Cúcuta, Colombia
- Dataset: File2.csv

Again, the MTL framework realizes that the task is specified in Table 2 (of Meta-models) and determines whether it can use those models verifying that the datasets with which they were built are like the new dataset (using the data in Table 1). When it does not find a similar model (using Table 1), then it proceeds to build a new model by checking if the dataset has enough data. If there is not enough data, then it checks if there is a similar dataset in Table 1. In this case, no similar dataset was found, so it proceeds to *Synthetic Data Generation*. Finally, the model is built with synthetic data and the techniques available in the architecture (Table 3). The Gradient Boosting Regressor algorithm turns out to be the most efficient for this specific task, obtaining an R2 of 0.9547 (see Figure 5).

Result

Traceability

ID	1
DESCRIPTION	Decrease of the bean in the roasting process
TYPE	predictive
ID_MT	8
TYPE_TECHNIQUE	supervised/regression
PARAMETERS	{"alpha": 0.9, "ccp_alpha": 0.0, "criterion": "friedman_mse", "init": null, "learning_rate": 0.1, "loss": "squared_error", "max_depth": 3, "max_features": null, "max_leaf_nodes": null, "min_impurity_decrease": 0.0, "min_samples_leaf": 1, "min_samples_split": 2, "min_weight_fraction_leaf": 0.0, "n_estimators": 100, "n_iter_no_change": null, "random_state": 123, "subsample": 1.0, "tol": 0.0001, "validation_fraction": 0.1, "verbose": 0, "warm_start": false}
METRICS	{"MAE": 0.0004, "MSE": 0.0, "RMSE": 0.0019, "R2": 0.9547, "RMSLE": 0.001, "MAPE": 0.0005, "TT (Sec)": 0.331}
ID_MD	1
FEATURES	[["ORDEN_x", "SC_NUMBACH", "TTOTAL", "NCF_LLAMA1", "PESOVERDE", "CAGUA", "fecha2", "D\u00eda", "Mes", "A\u00f1o", "Semana", "Fecha_3", "BATCH_SC", "FECHA_HORA_BACHE", "ORDEN_y", "MATERIAL", "COLOR", "HUMEDAD", "DENSIDAD", "Merma", "Cat_Merma", "reduction"], ["int64", "int64", "int64", "float64", "float64", "int64", "float64", "int64", "int64", "int64", "int64", "int64", "int64", "float64", "int64", "int64", "int64", "float64", "float64", "float64", "int64", "float64"]]
TARGETS	[["reduction"], ["float64"]]

Figure 5. Task 2 result using Gradient Boosting Regressor.

Finally, the ACODAT of the MTL framework verifies the quality of the new model, finding that the Gradient Boosting Regressor has a high degree of fit to the new dataset, as the R2 is very close to 1, with an MAE of 0.04 and an RMSE of 0.19. Therefore, the model can capture the underlying relationships between variables effectively for the new dataset.

C. Task 3. Coffee processing method.

For this task, we will use the original dataset used in task 1 with the purpose of performing a supervised classification of coffee processing methods. The variable to be classified will be the column 'Processing method', which presents five different classes, representing the different methods used in the coffee industry. The following are the specifications for the search of the models for this task:

- Model Type: Classification
- Technique Type: Supervised/Classification
- Model Description: Coffee processing method
- Dataset Description: The data used includes expert evaluations of different coffee blends, assigning a rating to each cup tested.
- Dataset Location: Cúcuta, Colombia
- Dataset: File1.csv

Again, the MTL framework realizes that the task is specified in Table 2 (of Meta-models) but it only sees moderately similar datasets with these models in Table 1, so it proceeds to do *Parameter Transfer* from the model of the most similar dataset built previously. In this case, it is the Random Forest Classifier. In addition, it realizes that the dataset is small and that there is a similar dataset (using Table 1), so it proceeds to do *Data Transfer* from that dataset. Finally, the model is built using the parameters obtained from the Random Forest and the augmented dataset.

Result

Traceability

ID	1
DESCRIPTION	Coffee processing method
TYPE	classification
ID_MT	1
TYPE_TECHNIQUE	supervised/classification
PARAMETERS	{"bootstrap": true, "ccp_alpha": 0.0, "class_weight": null, "criterion": "gini", "max_depth": null, "max_features": "sqrt", "max_leaf_nodes": null, "max_samples": null, "min_impurity_decrease": 0.0, "min_samples_leaf": 1, "min_samples_split": 2, "min_weight_fraction_leaf": 0.0, "monotonic_cst": null, "n_estimators": 100, "n_jobs": -1, "oob_score": false, "random_state": 123, "verbose": 0, "warm_start": false}
METRICS	{"Accuracy": 0.9217, "AUC": 0.0, "Recall": 0.9217, "Prec.": 0.9264, "F1": 0.9224, "Kappa": 0.8546, "MCC": 0.8562, "TT (Sec)": 8.919}
ID_MD	3
FEATURES	[["Numero_de_Bolsas", "Peso_Bolsa", "Aroma", "Sabor", "Regusto", "Acidez", "Cuerpo", "Balance", "Uniformidad", "Taza_limpia", "Dulzura", "Puntos_de_Copa", "Total_Puntos_Copa", "Humedad", "Categor\u00eda_Uno_Defectos", "Categor\u00eda.Dos.Defectos", "altitude_low_meters", "altitude_high_meters", "altitude_mean_meters", "Variedad_Arusha", "Variedad_Blue_Mountain", "Variedad_Bourbon", "Variedad_Catimor", "Variedad_Catuai", "Variedad_Caturra", "Variedad_Ethiopian_Heirlooms", "Variedad_Ethiopian_Yirgacheffe", "Variedad_Gesha", "Variedad_Hawaiian_Kona", "Variedad_Java", "Variedad_Mandheling", "Variedad_Marigojipe", "Variedad_Moka_Peaberry", "Variedad_Mundo_Novo", "Variedad_Other", "Variedad_Pacamara", "Variedad_Pacas", "Variedad_Pache_Comun", "Variedad_Peaberry", "Variedad_Ruiru_11", "Variedad_SL14", "Variedad_SL28", "Variedad_SL34", "Variedad_Sulawesi", "Variedad_Sumatra", "Variedad_Sumatra_Lintong", "Variedad_Typica", "Variedad_Yellow_Bourbon", "Color_Blue-Green", "Color_Bluish-Green", "Color_Green", "Metodo_de_procesamiento"], ["int64", "int64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "float64", "int64", "int64", "float64", "float64", "float64", "int64"]]
TARGETS	[["Metodo_de_procesamiento"], ["int64"]]

Figure 6. Task 3 result using Random Forest Classifier

Finally, the ACODAT of the MTL framework verifies the quality of the model found on the new dataset, confirming the excellent performance of the model, reflected in metrics such as Accuracy, Recall, Precision and F1, close to 1.

5. Comparative analysis

This section carries out a comparison with the works in literature. For that, two types of comparisons are carried out.

5.1. Qualitative comparison

In this section, we establish a series of criteria to compare our proposal with similar research carried out previously.

- Criterion 1: The work adapts ACODATs.
- Criterion 2: The MTL framework considers an autonomic adaptive process.
- Criterion 3: The MTL framework uses transfer learning techniques for the generation of new knowledge models.

Table 5 provides a qualitative comparative assessment based on the above criteria, of our approach with similar studies.

Table 5. Comparison with previous works.

	Cri1	Cri2	Cri3
[16]	x	√	x
[17]	x	x	x
[18]	X	X	x
[20]	x	√	x
[25]	x	√	x
[26]	√	x	x
[27]	X	√	x
This work	√	√	√

For criterion 1, our work is the only MTL approach specified for the adaptation of ACODATs. For criterion 2, in addition to our work that uses an ACODAT module to adapt an ACODAT, other researchers propose an adaptive process of supervised systems. Particularly, the work [16] presented an adaptive approach for real-time facility management systems, with applications ranging from early detection of equipment failures to automated monitoring and optimization of energy use. The work [20] proposes an MTL framework for identification models used in the fattening process of animals to detect anomalies in cattle weight. Furthermore, [25] defined automated relational meta-learning. Finally, [27] presented an adaptive MTL model based on user relevance. For criterion 3, our work is the only one that implements mechanisms for the generation of knowledge models using transfer learning techniques. Particularly, our approach implemented an ACODAT module to guide the behavior of the MTL framework, to determine the transfer learning technique to apply to adapt the supervised ACODATs.

5.2. Quantitative comparison

This section focuses on quantitatively comparing the quality of the learning models obtained by our MTL framework for adapting supervised ACODAT with respect to previous works. In the quantitative comparison, papers that present metrics of the models built using their MTL architectures were selected. Specifically, we compare the quality of the transfer decision taken by our architecture to build new knowledge models. Table 6 presents the metrics in various papers against those obtained in this work. The paper [25] achieved better metrics in the classification model compared to the papers [16] and [26], with an F1 score of 0.70 and an Accuracy of 0.80, but our paper was better with a 0.92 in F1 score and 0.92 in Accuracy. The article [17] achieved a Silhouette metric of 0.86 in the clustering model, but our work was better with a metric of 0.93 for DBCV. Finally, the work [27] used the regression technique with MAE metrics of 0.78 and RMSE of 0.95 our work improved with an MAE of 0.04 and RMSE of 0.19.

Table 6. Quantitative analysis with previous works (NA: Not apply)

Work	Classification		Prediction		Clustering
	F1-score	**Accuracy**	**MAE**	**RMSE**	**Silhouette**
[16]	0.70	0.80	NA	NA	NA
[17]	NA	NA	NA	NA	0.86
[25]	0.80	**0.95**	NA	Na	NA
[26]	0.72	0.88	NA	NA	NA
[27]	NA	NA	0.78	0.95	NA
Our approach	**0.92**	**0.92**	**0.04**	**0.19**	**0.93**

6. Conclusions

In this research, an MTL architecture has been introduced that uses a specific autonomous cycle to enable self-management and autonomous self-control processes in an ACODAT. A notable advantage of our proposal is the ability to maintain a continuous learning process with the arrival of new data or information for a supervised ACODAT. It can also adjust the models found in its knowledge base and their respective parameters. Thus, one of the most relevant features of our MTL architecture is its ability to adapt.

The MTL architecture we have developed is the first to be used in a real case, in this case, for coffee quality, spoilage, and processing method detection in the coffee industry. The effectiveness of our approach lies in its ability to adapt an ACODAT in an autonomic way using an ACODAT module. In the case study, an autonomous cycle based on three tasks was adapted for a coffee company.

On the other hand, qualitative and quantitative comparisons with other architectures implementing similar mechanisms were carried out, demonstrating that our proposal is capable of automatically generating high accuracy models for a wide range of tasks, including classification, regression and clustering. In addition, the architecture shows a remarkable ability to adapt to reduced datasets, thanks to the exploitation of synthetic data generation and data transfer techniques.

Despite the advantages it offers, our framework has certain drawbacks that need to be addressed to improve its effectiveness. The first step is to gather a significant amount of data to train the initial models in the knowledge base, which can be challenging in some situations. Moreover, they perform an autonomous real-time supervision process, which is time-consuming. This work presents a first case study. It would be interesting, in future research, to evaluate the sensitivity of our approach in different ACODATs, for different situations of process automation in MSMEs.

Particularly, to explore its use in different autonomous cycles in other agro-industrial chains.

References

[1] Ma Y., Zhao S., Wang W., Li Y., King I. (2022) Multimodality in meta-learning: A comprehensive survey," Knowledge-Based Systems, vol. 250.

[2] Liu Y, Gu F., Wu Y., Gu X., Guo J. (2022) A metrics-based meta-learning model with meta-pretraining for industrial knowledge graph construction, *Computers in Industry*, Vol 143.,

[3] Van Nguyen T., Huang P., Chien C., Chang C. (2023) Digital transformation for cost estimation system via meta-learning and an empirical study in aerospace industry, *Computers & Industrial Engineering*, Vol. 184.

[4] Chen C., Wang T., Liu C., Liu Y. Cheng L. (2023) Lightweight Convolutional Transformers Enhanced Meta-Learning for Compound Fault Diagnosis of Industrial Robot, *IEEE Transactions on Instrumentation and Measurement*, vol. 72.

[5] Tamascelli N., Paltrinieri N., Cozzani V. (2023) Learning From Major Accidents: A Meta-Learning Perspective, *Safety Science*, Vol. 158.

[6] Aguilar, J.; Ardila, D.; Avendaño, A.; Macias, F.; White, C.; Gomez-Pulido, J.; Gutierrez de Mesa, J.; Garces-Jimenez (2020) A. An Autonomic Cycle of Data Analysis Tasks for the Supervision of HVAC Systems of Smart Building. *Energies* vol. *13*.

[7] Sánchez M., Exposito E., J. Aguilar (2020) implementing self-* autonomic properties in self-coordinated manufacturing processes for the Industry 4.0 context, *Computers in Industry,* Vo. 121.

[8] Aguilar, J.; Salazar, C.; Velasco, H.; Monsalve-Pulido, J.; Montoya, E. (2020) Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents. *Computation*, vol *8*.

[9] Kephart. J., Chess D. 2003, "The vision of autonomic computing," Computer (Long Beach Calif), vol. 36, 41–50.

[10] IBM (2006) An architectural blueprint for autonomic computing," IBM White Paper, vol. 36.

[11] Riofrio G., Encalada E., Guamán D. Aguilar J. (2015) “Business intelligence applied to learning analytics in student-centered Learning processes”, Latin American Computing Conference (CLEI).

[12] Morales L., Ouedraogo C., Aguilar J., Chassot C., Medjiah S., Drira K. (2019) Experimental Comparison of the Diagnostic Capabilities of Classification and Clustering Algorithms for the QoS Management in an Autonomic IoT Platform, Service Oriented Computing and Applications, vol. 13, 199–219.

[13] López C., Santorum M., Aguilar J. (2019) Autonomous Cycles of Collaborative Processes for Integration Based on Industry 4.0", In Information Technologies and Systems (A. Rocha, C. Ferrás, M. Paredes, Eds.), Advances in Intelligent Systems and Computing, Springer, 177-186.

[14] Vizcarrondo J., Aguilar J., Exposito E., Subias A. (2017) MAPE-K as a Service-oriented Architecture, IEEE Latin-American Transactions, vol 15, 1163-1175.

[15] Pan, J., Yang, Q. (2010). A survey on transfer learning. In IEEE Transactions on Knowledge and Data Engineering vol. 22, 1345–1359.

[16] Moon. J, Noh, S. Jung, Lee J., Hwang E., (2023). Anomaly detection using a model-agnostic meta-learning-based variational auto-encoder for facility management, Journal of Building Engineering.

[17] Moulavi D., Jaskowiak P., Campello R. Zimek A. Sander J. (2014) Density-Based Clustering Validation, 14th SIAM International Conference on Data Mining.

[18] Tavares G. Junior S., (2021) Process mining encoding via meta-learning for enhanced anomaly detection, Communications in Computer and Information Science, vol. 1450, 157 – 168.

[19] Garouani M., Ahmad A., Bouneffa M, Hamlich M., Bourguin G. Lewandowski A. (2022). Using meta-learning for automated algorithms selection and configuration: an experimental framework for industrial big data”. Journal of Big Data.

[20] García R., Aguilar J. (2024) A meta-learning approach in a cattle weight identification system for anomaly detection, Computers and Electronics in Agriculture, Vol. 217,

[21] Abdul Majid, N., Abdullah E., Diana, A. (2015). A Review of Quality Coffee Roasting Degree Evaluation. Journal of Applied Science and Agriculture, vol. 10, 18–23.

[22] Zhang, Y., Wei, S., Zhang, L., Liu, C. (2019) Comparing the Performance of Random Forest, SVM and Their Variants for ECG Quality Assessment Combined with Nonlinear Features. Journal of Medical and Biological Engineering, vol. 39, 381–392.

[23] Kim, J. (2022) Coffee Beans Quality Prediction Using Machine Learning. SSRN Electronic Journal.

[24] Lake B., Ullman T., Tenenbaum J., Gershman S. (2017) Building machines that learn and think like people, Behavioral and brain sciences, vol. 40.

[25] Yao H., Wu X., Tao Z., Li Y., Ding B., Li R., Li Z., (2020) Automated relational meta-learning, International Conference on Learning Representations.

[26] Fuentes J, Aguilar J, Montoya E. (2024) Autonomous Cycle of Data Analysis Tasks for the determination of the Coffee Productive Process for MSMEs. Article in revision.

[27] Xu J., Zhang H., Wang X., Pin L. (2023) An Adaptive Meta-Learning model based on user relevance for user cold-start recommendation.” Knowledge-Based Systems. Vol. 18.

[28] Fuentes, J.; Aguilar, J.; Montoya, E.; Pinto, Á. (2024) Autonomous Cycles of Data Analysis Tasks for the Automation of the Production Chain of MSMEs for the Agroindustrial Sector. Information, vol. 15.

[29] Aguilar J., Jerez M., Exposito E. Villemur T. (2015) CARMiCLOC: Context Awareness Middleware in Cloud Computing, Latin American Computing Conference (CLEI).

[30] Radjaloa, V. H. (2024) Examining the linkages between technology adoption and components of corn agribusiness. Procuratio: Jurnal Manajemen & Bisnis, vol. 3.

[31] Riaño-Herrera D, Varela-Martínez D, Chenet J, García-García D, Díaz-Verus S, Rodríguez-Urrego L, (2023) Driving sustainable urban development: Exploring the role of small-scale organic urban agriculture in Bogotá, Colombia: A case study, Sustainable Cities and Society, Vol. 99.

[32] Paul, P. (2021). Agricultural Informatics and practices—The Concerns in Developing and Developed Countries. In: Choudhury, A., Biswas, A., Singh, T.P., Ghosh, S.K. (eds) Smart Agriculture Automation Using Advanced Technologies. Transactions on Computer Systems and Networks.

Appendix I. Analysis of Quantum Support Vector Machines in Classification Problems.

Analysis of Quantum Support Vector Machines in Classification Problems

Jairo Alberto Fuentes Camargo
GIDIS, Universidad Francisco de Paula Santander, Cúcuta, Colombia.
GIDITIC, Universidad EAFIT,
Medellín, Colombia
fuentesjairo@gmail.com

Jose Aguilar
GIDITIC, Universidad EAFIT,
Medellín, Colombia
CEMISID, Universidad de Los Andes,
Mérida, Venezuela
IMDEA Networks Institute, Spain
aguilar@ula.ve

Ángel Pinto
Universidad del Sinú
Monteria, Colombia
angelpinto@unisinu.edu.co

***Abstract*— A typical problem that is solved by supervised machine learning techniques is classification. There are many types of this problem, such as binary classification, multi-classification, and multi-labels. The development of techniques that perform well in different types of classification problems, and for datasets with many or few variables, or many or few data, remains a challenge. On the other hand, recent work indicates that machine learning can benefit from quantum computing in terms of reducing computational complexity and improving its performance. The purpose of this research is to analyze the use of Quantum Support Vector Machines (QSVM) for different types of classification problems. For this, several kernels (linear and Gaussian) are considered. In turn, a comparison is made with Random Forest. Also, a detailed analysis is carried out for the binary classification of coffee beans, obtaining a precision of 95% and an accuracy of 79%.**

***Index Terms*- Support Vector Machines, Quantum Computing, Classification Model, Coffee Bean Quality Classification.**

INTRODUCTION

A classic problem solved by supervised machine learning techniques is classification, which consists of determining/predicting an output (response). Classically, these response variables are categorical variables, and there are many ways to characterize those variables, such as binary, multi-class, or multi-label variables. In general, machine learning techniques perform well for very precise cases of classification, or for very specific datasets (with many or few variables, or many or few data).

On the other hand, Support Vector Machines (SVM) is a machine learning technique that has been successfully used in binary classification tasks, due to its ability to construct an optimal hyperplane in an n-dimensional space from a set of training data, which is used to classify new data. Currently, there is a variety of SVM, such as Gaussian SVM [2], multi-core SVM [3], twin SVM [1] and others. Furthermore, SVMs have been used in numerous fields such as facial recognition [5], medical diagnosis [6-8], handwritten character recognition [4], and supervision processes [9], among others.

In turn, the capabilities of quantum computing have inspired some researchers to integrate them with machine learning algorithms, spawning a new domain of research known as quantum machine learning in recent years [10-14]. Several quantum machine learning algorithms have been proposed in the literature, such as quantum support vector machines [20], quantum neural networks [15-17], quantum K-nearest neighbors [19], quantum clustering [21], and quantum principal component analysis [18], among others, demonstrating their superiority in terms of efficiency and speed over traditional machine learning [22].

In [26], classical SVMs were implemented using quantum cores for the analysis of data from satellites, and the design and implementation of hybrid SVMs with quantum cores were presented. Six SVM models were evaluated quantitatively, qualitatively and statistically: classical linear SVMs, SVMs based on RBF kernel and hybrid SVMs with kernels generated by simulated quantum circuits. Although these SVMs offer good results, they are still slightly less effective than recent deep learning models. In [27], the performance of the QSVM algorithm was compared with classical methods on an eye-state EEG dataset, considering accuracy and execution time. The results showed that the QSVM achieved an accuracy of 99.47% with different qubit values. The precisions obtained were very close to those of classical methods, demonstrating the suitability of quantum algorithms to analyze EEG signals.

The performance of traditional machine learning models was evaluated against a QKSVM classifier on a dataset with limited audio samples in [28]. The quantum kernel method performed better than classical models, in spite of the small sample size. This is because it can efficiently extract high-dimensional features. In [30], Sridevi et al. analyzed the learning capabilities of an SVM classifier with a kernel based on instantaneous quantum polynomial embedding and compared them with the classical SVM, particularly in situations with smaller sample sizes and feature vectors. The results show that the quantum core strategy outperforms its traditional counterpart, even with smaller datasets.

The purpose of this paper is to analyze the Quantum Support Vector Machines (QSVM) capacity in different types of classification problems (binary and multi-class), or datasets. In addition, we compare QSVM with different versions of SVM (with linear and Gaussian kernels), and with Random Forest. Finally, we carry out a detailed study of the utilization of QSVM in coffee bean classification tasks and compare it with other works. In our study, the results obtained on the test sets for various classification problems allowed us to reach more general conclusions about the capability of the QSVM. We find that QSVM performs better on multiple classification problems than on binary classification, regardless of the characteristics of the dataset, such as the amount of data and variables/dimensions. Furthermore, the quality difference compared to the random forest is significantly reduced.

This article is organized as follows. Section 2 presents quantum computing in SVM. Section 3 presents a detailed description of the implementation of the classification model using QSVM. Then, in section 4 is presented in detail the experiments carried out for a binary classification problem, the classification of coffee beans according to their quality, and the results obtained for other types of classification problems. In all of them, it is compared with Random Forest and with two versions of SVM, Linear and Gaussian. In addition, a qualitative analysis is done to compare it with other works. Finally, the last section presents the conclusions and future work.

QUANTUM SVM

Quantum computing takes advantage of the unique features of quantum physics, such as superposition, entanglement, and interference, applying them to computing. On the other hand, the SVM is a supervised learning algorithm used in a wide variety of classification and regression problems. Its use extends to various areas, such as medical signal processing, natural language processing, and image and voice recognition. The goal of the SVM algorithm is to determine a hyperplane that can effectively split different data classes. This optimal division involves finding the hyperplane that maximizes the margin between these classes, that is, the distance between the closest points of each class to the hyperplane.

Foundations of Quantum Support Vector Machines (QSVM).

Quantum computing is based on the properties of quantum systems, such as entanglement, interference and superposition, with the aim of accelerating some computational operations [17]. Therefore, compared with conventional computing devices, quantum computers can solve complex problems more efficiently. In general, there are several quantum models to operationalize these properties, from quantum Turing machines, through adiabatic quantum computing and measurement-based quantum computing. On the other hand, every calculation has three (3) elements: operations, data, and results. In the quantum context, these agree with the concepts of qubits, quantum gates and measurements.

In the context of SVMs, the concept of the kernel is fundamental. The kernel concept involves projecting the data into a feature space whose dimensions are higher [20]. To do this, all that is needed from the space is the kernel function that defines it. In this space, the input data is partitioned by a hyperplane. Typically, the feature space is the standard Euclidean space, just with an extra dimension.

A QSVM works similarly to an SVM but incorporates a quantum state space as its feature space [20]. Considering the above, a QSVM for classification performs all operations in a completely classical way, except in the kernel function calculation process. In this case, the kernel function employs a quantum computer to perform the following tasks: receive two vectors from the original data space, map them to quantum states using a feature map, compute the inner product of these quantum states, and return the result. In this way, when implementing quantum feature maps, we refer to circuits parameterized by the original data, which implies that they generate a quantum state completely conditioned by said information.

Thus, we have a feature map φ. The circuit configuration Φ, which will be used for the implementation, will depend on the classical data in the original space. Each time an input $\vec{x}$ is introduced, a circuit $\Phi(\vec{x})$ is generated so that the resulting quantum state is, $\varphi(\vec{x}) = \Phi(\vec{x})\,|0\rangle$. With the feature map already developed, we can formulate the kernel function as follows [20]:

$$(\vec{a},\vec{b}) = |\langle(a)|\varphi(b)\rangle|^2 = ||\langle 0|\, \Phi^\dagger(a)\Phi(b)\, |0\rangle||^2. \quad (1)$$

The possibility of obtaining an all-zero measurement, after having created the state $\Phi^\dagger(\vec{a})\ \Phi(\vec{b})\ |0\rangle$, is based on the orthogonality of the computational basic.

The circuit calculation for $\Phi^\dagger$ simply follows the inverse process of Φ. Considering that Φ is made up of a series of quantum gates, the procedure involves applying these gates in the circuit from right to left and inverting each of them. This is the way to implement a quantum kernel function: using a feature map that generates a circuit $\Phi(\vec{x})$ for any input $\vec{x}$, then preparing the state $\Phi^\dagger(\vec{a})\ \Phi(\vec{b})\ |0\rangle$ for the pair of vectors in which one wishes to compute the kernel, and finally, obtaining the probability of measuring zero on all qubits.

Feature maps

As mentioned before, a feature map is usually represented by a parameterized circuit $\Phi(\vec{x})$, which is based on the original data and can therefore produce a state that varies as a function of this data [20].

ZZ Features Map is a form of feature map that can process inputs $na_1, ..., a_n$ inputs into n qubits, like angle embedding. The creation of its parameterized circuit follows these steps:

1. Use a Hadamard gate (a way to represent a qubit in the quantum Fourier transform) on all qubits.
2. Apply a rotation Rz $(2x_j)$ on each qubit $_j$.
3. For each pair of elements $\{j, k\} \subseteq \{1, ..., n\}$ with $j < k$, do the following:

(a) Apply a CNOT gate directed to qubit k, which is controlled by qubit j.

(b) Apply a rotation Rz $(2(\pi - x_j)\ (\pi - x_k))$ on qubit k.

(c) Repeat step 3a.

A representation of the feature map of ZZ on three qubits can be found in Figure 1, which represents the circuit of the feature map. As in angle encoding, normalization is essential in the context of the ZZ feature map [20]. Normalizing variables to intervals such as [0, 1] is useful for striking an appropriate balance between separating the extremes of the dataset and using a wide region in the feature space.

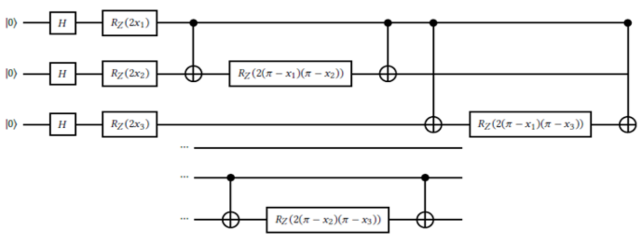

Figure 1. ZZ Feature Map in three qubits with x_1, x_2, x_3 inputs.

In a n-qubit circuit, this feature map can accept up to n numerical inputs $x_1, ..., x_n$. In this circuit, a rotation gate is applied on each qubit j, with the parameterization based on the value xj. In this feature map, the xj values are used as angles for the rotations, justifying the encoding term.

In angle encoding, it is possible to employ Rz gates with the initial state $|0\rangle$ selected, ensuring that the feature map operation is negligible, which can be easily confirmed by the definition of Rz. Therefore, when using Rz gates, it is typical to apply Hadamard gates that affect each qubit. This whole process is presented in detail in Figure 2.

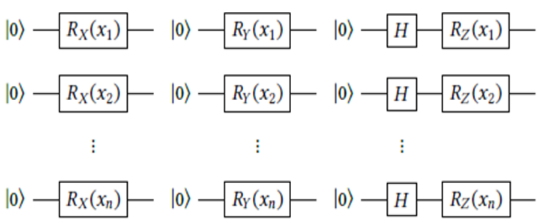

Figure 2. Angle encoding for one input ($x_1 ... x_n$) using different rotation gates.

For angle encoding in the feature map, it is essential to normalize the variables to a specific interval [25]. If normalized between 0 and 4π, the data will occupy a larger region in the feature space compared to normalizing between 0 and 1. However, this has the drawback that the extremes of the dataset will be mapped together due to the feature map effect. This occurs because, according to our definition of rotation gates, angles 0 and 2π are considered identical when dividing the entrance angle by 2. Therefore, the normalization decision involves finding a middle ground between separating the extremes in the feature space and making the most of the available area in the feature space.

OUR CLASSIFICATION MODEL BASED ON QSVM

In this section, we describe the implementation of the machine learning model based on QSVM. Our QSVM is a variant of SVM that incorporates a quantum core. We use Qiskit, created by IBM to define a quantum computing environment, as the software development kit [16]. Qiskit is a toolkit designed to develop quantum circuits and run them in a local simulator, or to submit them to IBMQ, a real quantum computer. So, we start by creating the first quantum circuit with a classical register. We define an empty quantum circuit with 2 qubits (q0_0 and q0_1) and 2 classical registers (c0_0 and c0_1) (see Figure 3).

```
circuit = qk.QuantumCircuit(q, c)
print(circuit)
```

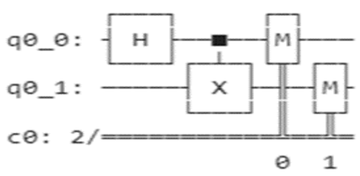

Figure 3. Our initial circuit.

We execute the circuit in the quantum simulator. To run the circuit, we employ Qiskit's Qasm quantum simulator and configure the circuit for the feature map using the qiskit.utils algorithm_globals and QuantumInstance libraries. Algorithm 1 describes the algorithm to define the initial quantic context.

Algorithm 1: Definition of the quantic context

```
# Utilization of Qiskit Aer's Qasm Simulator: Define where
   To execute the simulation.
simulator = qk.BasicAer.get_backend('qasm_simulator')
# Simulate the circuit in the simulator to get the result.
job = qk.execute(circuit, simulator, shots=100)
result = job.result()
# We obtain the aggregated binary results of the circuit.
counts = result.get_counts(circuit)
print (counts)
```

Next, we create the quantum core by importing the "ZZFeatureMap" package, It is one of the common feature maps in Qiskit's circuit collection. For training, the SVM dataset Cafe/ Data_Aprendizaje_Supervisado_Merma.xlsx' was used), and two labels were determined class_labels = ['Cat_Merma 1', 'Cat_Merma 0']. In this case, we use a Gaussian kernel as the kernel. We configure the SVM classifier using QSVC from qiskit_machine_learning.algorithms. Next, we define a new quantum kernel based on the fidelity exempla of qiskit_machine_learning.kernels (fidelityQuantumKernel). Algorithm 2 describes the algorithm for defining the feature map.

Algorithm 2: Definition of the feature map

```
from qiskit_machine_learning.algorithms import QSVC
qsvc = QSVC (quantum_kernel=new_kernel)
qsvc.fit(features, labels)
qsvc.score(features, labels)
```

Finally, we configure the QSVM classifier using qiskit_machine_learning.algorithms (we import QSVC). To generate the feature map loop, we specify an arbitrary vector x that we want to encode and create the loop for this data point. To achieve this, we install the pylatexenc library (see Figure 4).

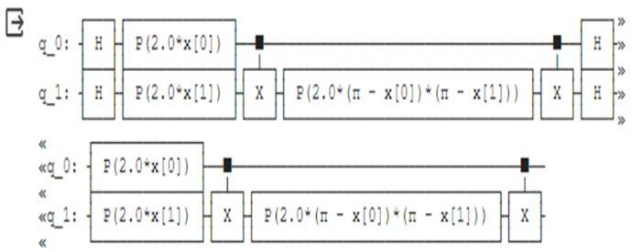

Figure 4. SVM classifier feature map

RESULTS

For the experiments in this section, we have used Python 3.10 programming language and Ubuntu operating systems, with the Qiskit library, version 0.23.4.

A Preliminary Analysis of the Results of a Binary Classification Problem

In this section, we test our QSVM classifier in an initial classification problem, a binary classification problem, the classification of coffee beans [29].

There are a total of 2068 items in the coffee bean dataset, each with 30 different attributes such as water amount, humidity, flame temperature, and quality category. The categories can be Cat_Merma_1 or Cat_Merma_2. Our dataset about coffee beans is previously prepared (cleaned, etc.). To find out if the classifier is accurate, a common strategy is to split the dataset into a training set and a test set. Here, the data is divided into 70% training and 30% testing. To facilitate the analysis, we performed a dimensional reduction of our data from 30 attributes to 5 using principal component analysis (PCA). Dimensionality reduction allows removing additional and non-essential features for the classification problem.

Before going into the quantum part, we will define two SVM for the classification, one for linearly separable datasets and another for nonlinearly separable datasets. Figure 5 shows the coffee quality dataset vs linearly separable blobs. Our dataset is clearly not linearly separable.

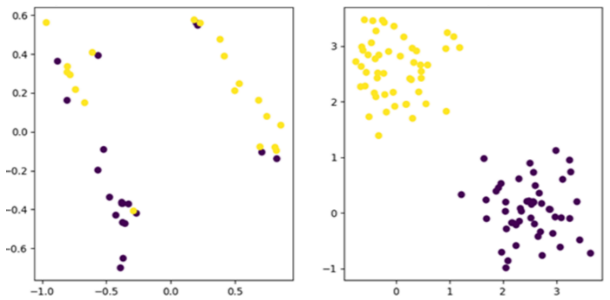

Figure 5. Coffee dataset vs linearly separable blobs.

We define a linear SVM classifier and train it on the dataset. Figure 6 shows the results, with a precision of 0.77 and an accuracy of 0.55.

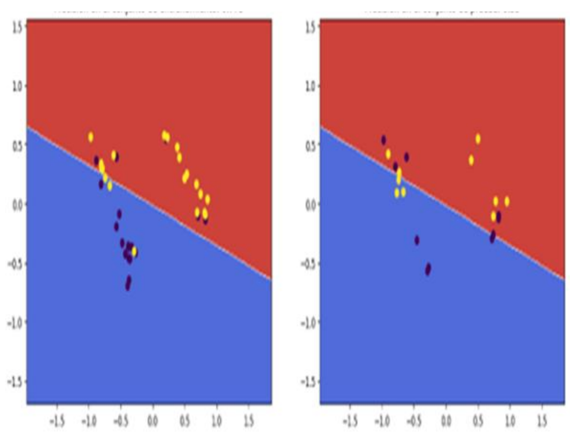

Figure 6. Results for the linear SVM Classifier (for the training and test sets).

Now, we implement an SVM with a Gaussian kernel. Figure 7 shows the results, with a precision of 0.9 and an accuracy of 0.75.

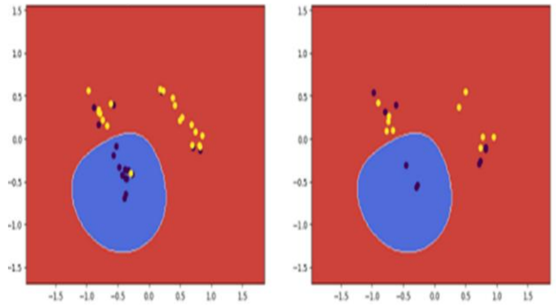

Figure 7. Results for the SVM with a Gaussian kernel (with the training set and test set).

Figure 8 shows the original used to train the QSVM and Figure 9 the classification using our QSVM classifier.

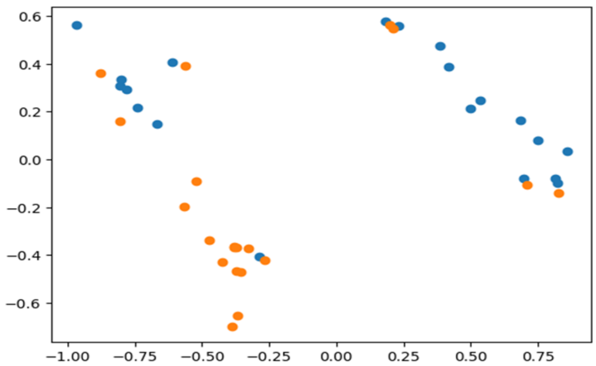

Figure 8. Coffee Dataset for training the QSVM.

In general, it obtained a precision of 95% and an accuracy of 79%.

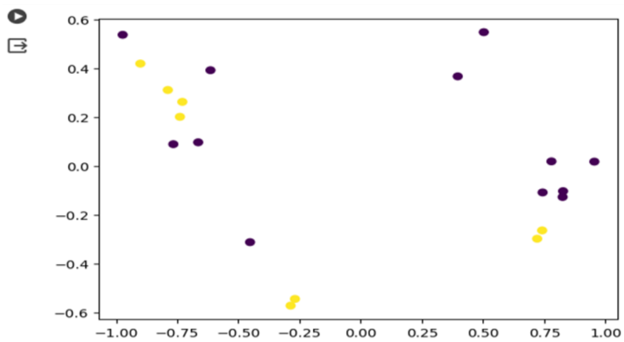

Figure 9. QSVM tag predictions.

Finally, we compare our QSVM classifier with Random Forest. In this case, the results of precision are 100% and of accuracy of 82%, which is the best result (see Figure 10).

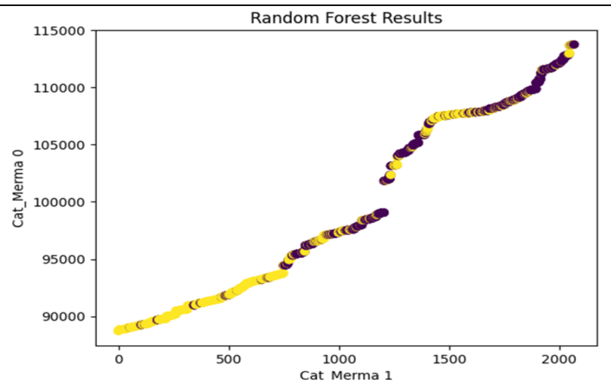

Figure 10. Random Forest results.

In summary, the most outstanding result was achieved with Random Forest (see Table 1), obtaining a precision of 100% and an accuracy of 82%. The relatively high level of accuracy suggests that there may be overtraining in this technique. The QSVM obtains the second best result, with an accuracy of 95% during training and 79% in testing, which reflects the interest in this technique.

Table 1. General Results for classification of coffee beans.

Model	Precision	Accuracy
Random Forest	100%	80%
Linear SVM	77%	55%
Gaussian SVM	90%	75%
QSVM	95%	79%

Quantitative Comparison in Different Classification Problems

The classification of coffee beans is a binary classification problem. In this section, we consider different types of classification problems (binary or multi-classes), and different characteristics of the datasets. The datasets used are described below.

The criterion for selecting the datasets is that they were binary or multi-classes classification problems, and that they had many variables (high dimension) that could not be reduced. Three datasets were selected: two from the medical field and one from the energy sector. The medical datasets pertain to COVID-19 and dengue, while the energy dataset covers energy demand, consumption, and price. The metadata for these datasets is detailed in Table 2.

Table 2. A brief introduction to the datasets

Dataset	Instances	No. of attributes	Type of classification problem
Dengue	32559	13	Multiclass
Energy	35145	62	Multiclass
Covid-19	5853480	53	Binary

The Energy dataset was obtained from the paper [23], and includes information about prices, energy consumption, and weather conditions, among other variables, in Spain, over a 4-year period. The original energy dataset comprises was collected hourly from January 1, 2015, at 00:00 hours to December 31, 2020, at 23:00 hours. Finally, a discretization was performed to determine the price levels (label) using the energy consumption.

The Dengue dataset corresponds to the data of patients with Dengue collected by the Colombia Ministry of Health in 2020 (https://medata.gov.co/dataset/dengue). The dataset preprocessing was carried out by Hoyos et al. in [24]. They selected 22 variables from the original dataset, considered by experts as the most relevant in Dengue. The selected characteristics include laboratory tests, symptoms, and the target variable used to classify the severity of Dengue. Except for the target variable, all characteristics are binary. The target variable differentiates between three (3) categories of Dengue: severe dengue, dengue with warning signs, and dengue without warning signs.

The COVID-19 dataset is from Israel (https://ourworldindata.org/coronavirus), and comprises tests conducted by the government between March and November 2020. The target feature classifies tests as positive or negative.

The following procedure was carried out on all datasets with all the techniques: a 5 cross-validation strategy was carried out to learn and test the classification models, using 80% of the data to train and 20% to test. To optimize the parameters of the techniques, a grid search type hyperparameter optimization process was used. Table 3 shows the results obtained for the metrics of Precision (P) and Accuracy (A).

Table 3. Performance Results for the Datasets

Model	Dengue		Covid-19		Energy	
	P	A	P	A	P	A
Random Forest	96%	81%	98%	89%	95%	92%
Linear SVM	91%	66%	81%	75%	88%	85%
Gaussian SVM	90%	70%	85%	81%	90%	85%
QSVM	94%	80%	88%	85%	94%	90%

According to the results, we confirmed two aspects: that Random Forest always gives the best results and that QSVM always gives the second-best results. We also found that the differences between Random Forest and QSVM are greatly reduced in multiclass problems, which indicates that QSVM improves significantly in those cases with respect to the binary classification cases. Finally, we see that the number of variables/dimensions does not have a great influence on the final quality of the models obtained (since there are datasets with many and few variables), and the order of quality between the techniques is maintained.

Qualitative comparison with other works

This work is compared with previous works using a set of three criteria, which are set out below.

- Cri1: Uses quantum computing (QC).
- Cri2: Utilization of machine learning algorithms in the classification model.

- Cri3: Uses quantum machine learning algorithms with datasets in the coffee agroindustrial sector.

Table 4 offers a comparison of our research with previous studies. The idea is to analyze previous works that already use quantum algorithms in classification tasks, and to do data-based analysis in the coffee agroindustrial sector.

Table 4. Comparison with previous works

	Cri1	Cri2	Cri3
[16]	√	√	x
[17]	√	√	x
[18]	√	x	x
[19]	√	√	x
[20]	√	x	x
[21]	√	x	x
[22]	√	√	x
This work	√	√	√

Because of the computing power it adds, quantum computing is being used in machine learning algorithms. In this way, we notice quantum versions of some machine learning techniques, such as quantum deep neural networks used in image recognition [16, 17], quantum SVM [20], quantum K-nearest neighbors [19], quantum clustering [21], and quantum principal component analysis [18], demonstrating in all cases its superiority in terms of quality and speed over traditional machine learning [22] (first criteria).

Now, some of that work is on tasks to build classification models [16, 17, 19, 22] (first criteria), but none of them are on analyzing the quality of the coffee bean (third criteria). We propose a viable approach for Micro, Small & Medium Enterprises (MSMEs), since we have used a simulator to implement the quantum solution. Thus, this is a useful, low-cost, and efficient tool for detailed analysis of the quality of the coffee bean to improve its impact on coffee production, which helps MSMEs meet their financial and environmental sustainability objectives.

CONCLUSIONS

This article has made an exhaustive analysis of the behavior of QSVM in different classification problems (binary and multiclass), showing its versatility in a simulated quantum computer. Furthermore, it has described in detail its implementation and behavior for the classification of coffee beans. For the comparison of the model, it was proposed to use several SVM models (using different kernels) and Random Forest.

The results obtained on the test sets for the different classification problems allowed us to reach more general conclusions about the capacity of the QSVM technique. We see that QSVM performs much better on multi-classification problems compared to binary classification problems, regardless of the characteristics of the datasets (lots of data and/or variables/dimensions). Furthermore, the quality difference with respect to random forest is drastically reduced.

Based on the results obtained, two significant conclusions can be drawn: first, that Random Forest consistently stands out as the most successful approach; second, that QSVM tends to deliver solid results, usually ranking second in performance. We also observe that the differences between Random Forest and QSVM are considerably reduced in multiclass classification problems, suggesting a significant improvement of QSVM in those scenarios relative to binary classification cases. Finally, as sample characteristics grow, the qubit wave function representation would experience exponential growth in simulation requirements, which would make it unfeasible when studying large classifier datasets. This would make it difficult for MSMEs to participate in the use of QSVMs and their computational costs.

In general, the results are quite encouraging, because they are close to those of Random Forest, even though QSVM was executed in the IBM simulator of the quantum computer that the Qiskit library has. One of the limitations was the test environment (IBM simulator), and another limitation is that it wasn't tested on multiple-label problems, but future works should extend both things. Also, future work should explore other ways to quantum implement SVMs.

ACKNOWLEDGMENTS

The authors would like to thank the HAMADI 4.0 project, code 22-STIC-06, of the STIC-AmSud regional program, for supporting this work.

REFERENCES

A. Qing, Y. Kang, and A. Wang, “A novel semi-supervised multi-label twin support vector machine,” Intelligent Automation and Soft Computing, vol. 27, no. 1, pp. 205–220, 2021.

K. S. Bhuvaneshwari, J. Uma, K. Venkatachalam, M. Masud, M. Abouhawwash et al., “Gaussian support vector machine algorithm-based air pollution prediction,” Computers, Materials & Continua, vol. 71, no. 1, pp. 683–695, 2022.

R. Sujatha, J. M. Chatterjee, N. Jhanjhi, T. A. Tabbakh and Z. A. Almusaylim, “Heart failure patient survival analysis with multi kernel support vector machine,” Intelligent Automation & Soft Computing, vol. 32, no. 1, pp. 115–129, 2022.

S. Dhankhar, M. K. Gupta, F. H. Memon, S. Bhatia, P. Dadheech et al., “Support vector machine based handwritten hindi character recognition and summarization,” Computer Systems Science and Engineering, vol. 43, no. 1, pp. 397–412, 2022.

S. Guo, S. Chen, and Y. Li, “Face recognition based on convolutional neural network and support vector machine,” in IEEE Int. Conf. on Information and Automation (ICIA), Ningbo, China, IEEE, pp. 1787–1792, 2016.

G. Jayandhi, J. Leena Jasmine and S. Mary Joans, “Mammogram learning system for breast cancer diagnosis using deep learning svm,” Computer Systems Science and Engineering, vol. 40, no. 2, pp. 491–503, 2022.

Y. Quintero, D. Ardila, E. Camargo, F. Rivas, J. Aguilar, “Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables”, Computers in Biology and Medicine, vol. 134, 2021.

J. Aguilar, M. Jerez, E. Exposito and T. Villemur, "CARMiCLOC: Context Awareness Middleware in Cloud Computing," Proceeding *Latin American Computing Conference (CLEI)*, 2015

M. Araujo, J. Aguilar, H. Aponte, "Fault detection system in gas lift well based on artificial immune system," Proceedings International Joint Conference on Neural Networks, pp. 1673-1677 2003.

S. Lloyd, M. Mohseni and P. Rebentrost, “Quantum algorithms for supervised and unsupervised machine learning,” arXiv preprint arXiv, 1307.0411, 2013.

S. Lloyd, M. Mohseni and P. Rebentrost, “Quantum principal component analysis,” Nature Physics, vol. 10, no. 9, pp. 631–633, 2014.

M. Schuld, I. Sinayskiy and F. Petruccione, “Prediction by linear regression on a quantum computer,” Physical Review A, vol. 94, no. 2, pp. 022342, 2016.

J. Biamonte, P.Wittek, N. Pancotti, P. Rebentrost, N.Wiebe et al., “Quantum machine learning,” Nature, vol. 549, no. 7671, pp. 195–202, 2017.

W. Jiang, J., Xiong, Y., Shi, “When machine learning meets quantum computers: a case study. In Proc. 26th Asia and South Pacific Design Automation Conference (ASP-DAC), pp.593–598. 2021.

I. Cong, S. Choi, M. Lukin, “Quantum convolutional neural networks”. Nat. Phys. Vol. 15, pp. 1273–1278, 2019.

Li, Y.C., Zhou, R.G., Xu, R.Q., et al.: A quantum deep convolutional neural network for image recognition. Quantum Sci. Technol. 5(4), 044003 (2020)

X. Niu, W. Ma, “Design of a novel quantum neural network”. Laser Phys. Lett. Vol. 17, no. 10, 105208, 2020.

S. Lloyd, M. Mohseni, P. Rebentrost, P. “Quantum principal component analysis”. Nat. Phys. Vol. 10, no. 9), pp. 631–633, 2014.

Y. Ruan, X. Xue, H., Liu, H., J. Tan, X. Li, “Quantum algorithm for k-nearest neighbors classification based on the metric of hamming distance. Int. J. Theor. Phys. Vol. 56, no 11, pp. 3496–3507, 2017.

P. Rebentrost, M. Mohseni, S., Lloyd, “Quantum support vector machine for big data analysis”. Phys. Rev. Lett. Vol. 113 no. 13, 130503 2014

A. Bishwas, A. Mani, V. Palade, “An investigation on support vector clustering for big data in quantum paradigm”. Quantum Inf. Process vol. 19, no. 4, 2020

[22] N. Mishra, M. Kapil, H. Rakesh, H., et al.: “Quantum machine learning: a review and current status”. In: Sharma N., Chakrabarti A., Balas V.E.,Martinovic J. (eds) DataManagement, Analytics and Innovation. Advances in Intelligent Systems and Computing, pp. 101–145, 2021.

[23] L. Morales, C. Ouedraogo, J. Aguilar, C. Chassot, S. Medjiah, K. Drira, “Experimental comparison of the diagnostic capabilities of classification and clustering algorithms for the QoS management in an autonomic IoT platform”. Service Oriented Computing and Applications, vol. 13, pp. 199–219, 2019.

[24] W. Hoyos, J. Aguilar, M., Toro, “A clinical decision-support system for dengue based on fuzzy cognitive maps”. Health Care Manag Sci, vol. 25, pp. 666–681, 2022.

[25] J. Aguilar, C. Salazar, H. Velasco, J. Monsalve-Pulido, E. Montoya, “Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents”. Computation vol. 8, 2020.

[26] M. Artur, M. Jakub, C. Grzegorz, S. Filip, G. Bartosz. “Detecting Clouds in Multispectral Satellite Images Using Quantum-Kernel Support Vector Machines”. Ieee journal of selected topics in applied earth observations and remote sensing, vol. 16, 2023.

[27] A. Gamzepelin, K. Murat. “Comparison of QSVM with Other Machine Learning Algorithms on EEG Signals”. 2023 11th International Symposium on Digital Forensics and Security (ISDFS). DOI: 10.1109/ISDFS58141.2023.10131123.

[28] K. Hasmitha N, S. Sridevi, K. Prasanna. “Quantum Kernel-aided Support Vector Machine Classifier for Improved Speech Classification”. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | DOI: 10.1109/ICCCNT56998.2023.10307618.

[29] J. Aguilar, “A General Ant Colony Model to solve Combinatorial Optimization Problem”, Revista Colombiana de Computación, Vol. 2, pp. 7-18, 2001.

[30] I. Sridevi, B Indira, S. Shivanya, S. Sahil, S. Aishwarya. “Quantum Enhanced Support Vector Machine with Instantaneous Quantum Polynomial Encoding for Improved Cyclone Classification”. 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC). DOI: 10.1109/ICRTAC59277.2023.10480791.