Universidad de Alcalá
Escuela Politécnica Superior

Grado en Ingeniería Informática

Trabajo Fin de Grado

Modelo de predicción energético basado en métodos de aprendizaje automático (AA) explicables

Autor: Diego Javier Benito Gutiérrez

Tutor: David Fernández Barrero

Cotutor: Jose Lisandro Aguilar Castro

2023

UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería Informática

Trabajo Fin de Grado

Modelo de predicción energético basado en métodos de aprendizaje automático (AA) explicables

Autor: Diego Javier Benito Gutiérrez

Tutor: David Fernández Barrero

Cotutor: Jose Lisandro Aguilar Castro

Tribunal:

Presidente: Julia María Clemente Párraga

Vocal 1º: María del Mar Lendínez Chica

Vocal 2º: David Fernández Barrero

Fecha de depósito: 8 de Septiembre de 2023

Agradecimientos

Este trabajo representa el final de una etapa trascendental en mi vida, y quiero tomar este momento para expresar mi más profundo agradecimiento a todas las personas que han sido parte de este largo camino.

En primer lugar, quiero extender mi gratitud hacia mi familia, especialmente a mis padres, Gladys y Javier, quienes han sido mi principal apoyo y motivación, sin los que no hubiera sido posible superar esta etapa. También quiero agradecer a mi hermana su apoyo incondicional, siempre dispuesta a ayudar, sin esperar nada a cambio. Asimismo, quiero expresar mi gratitud hacia mis amigos, tanto dentro de la carrera como fuera, cuyo apoyo moral me ha dado el impulso para seguir adelante.

Por último, no puedo olvidar mencionar a mis profesores a lo largo de esta etapa que han contribuido a mi formación y aprendizaje. Agradecimientos especiales a mis tutores del Trabajo de Fin de Grado, David Fernández Barrero, por tener paciencia conmigo y responder a todas mis preguntas y al cotutor del proyecto, José Lisandro Aguilar Castro, que hizo posible el desarrollo de este proyecto.

Resumen

El presente proyecto se enfoca en la construcción de modelos explicables de predicción energética, los cuales se encargan de predecir las etiquetas del consumo de energía. A lo largo del desarrollo del proyecto, surgió la idea de ampliar el alcance de estudio y explorar el comportamiento de la técnica usada en otros dominios críticos, como el campo médico. La técnica utilizada para la construcción de los modelos explicables son los Mapas Cognitivos Difusos (MCD), su capacidad de autoexplicabilidad radica en mostrar de manera intuitiva cómo los conceptos se influyen mutuamente a través de relaciones causales, lo que ayuda a entender y visualizar de forma clara qué conceptos contribuyen a una predicción o resultado específico. El trabajo detalla la creación de modelos basados en MCD tanto en el ámbito de la energía como en el campo médico, utilizando conjuntos de datos reales. Una vez creados los modelos, se realiza un análisis de explicabilidad, analizando la explicación que proporciona el propio modelo, como también, usando métodos de análisis de explicabilidad reportados en la literatura. En resumen, este proyecto se centra en desarrollar modelos explicables para la predicción energética y médica, empleando conjuntos de datos reales, para después analizar detalladamente la explicabilidad de los modelos generados.

Palabras clave: Aprendizaje automático, clasificación, análisis de explicabilidad, dominios críticos.

Abstract

The current project focuses on the construction of interpretable models for energy prediction, which are responsible for forecasting energy consumption labels. Throughout the project's development, the idea of expanding the scope of study emerged, aiming to explore the behavior of the technique used in other critical domains, such as the medical field. The technique utilized for constructing these interpretable models is Fuzzy Cognitive Maps. Their self-explanatory capacity lies in intuitively showcasing how concepts mutually influence each other through causal relationships, facilitating clear understanding and visualization of which concepts contribute to specific predictions or outcomes. The project outlines the creation of models based on FCMs in both the energy and medical realms, employing real-world datasets. After the models are developed, an explanation analysis is conducted, assessing the inherent explanatory power of the model itself, as well as, using explanation analysis methods defined in the literature. In summary, this project is dedicated to the elaboration of self-explaining models for energy and medical prediction employing real datasets, to later analyze in detail the explainability of the models generated.

Keywords: Machine Learning, Classification, explainability analysis, Critical domains.

Resumen extendido

El enfoque principal del proyecto es la construcción de modelos de Aprendizaje Automático explicables en el ámbito energético. Sin embargo, durante el desarrollo del proyecto, surgió el interés por ampliar el uso de la técnica empleada en otros dominios críticos como el sanitario; entendemos como dominios críticos aquellos en los que el modelo puede ser usado para tomar decisiones en donde las consecuencias pueden ser catastróficas. Al trabajar con dominios críticos, es necesario proporcionar al usuario un cierto grado de confianza de que el modelo funciona de la manera esperada. Nuestros modelos son explicables, eso quiero decir que el usuario pueda saber el motivo y las causas que han llevado al resultado obtenido (en nuestro caso, la predicción). En la construcción de los modelos se han empleado la técnica de Mapas Cognitivos Difusos (MCD). Un MCD permite analizar la explicabilidad en los modelos basándose en la relación de causalidad entre las variables que componen el modelo, lo cual permite determinar cómo el modelo ha llegado a una predicción dada.

En el estudio, se han empleado tres conjuntos de datos: uno en el dominio energético y dos en el campo médico. Los modelos de conocimiento desarrollados se encargan de predecir etiquetas, lo que los convierte en clasificadores. Con el conjunto de datos energético, se construye un modelo que predice el nivel de coste del consumo energético requerido. En cuanto a los conjuntos de datos médicos, con uno de ellos se crea un modelo que predice si una persona está o no enferma de COVID-19. Por otro lado, con el segundo conjunto de datos se construye un modelo que permiten predecir el grado de severidad del Dengue en un paciente. Finalmente, se define un enfoque de análisis para los MCD para analizar la explicación que aportan los modelos. Además de este enfoque, se estudia la explicabilidad de los modelos aplicando otros métodos de explicabilidad de la literatura.

Para el desarrollo correcto del proyecto, ha sido necesario realizar un estudio de otras investigaciones y técnicas empleadas en el ámbito del problema. El primer capítulo del proyecto se centra en exponer el contexto del problema y motivación para su resolución. El segundo capitulo presenta la base teórica del trabajo que engloba el estudio de trabajos relacionados, como también, los conocimientos teóricos requeridos para el desarrollo práctico. También se presenta de manera teórica las distintas técnicas de Aprendizaje Automatico (AA) y métodos de explicabilidad empleados en este proyecto. Una vez presentado el marco teórico, se procede a poner en práctica todo lo explicado anteriormente en el desarrollo experimental. En resumen, la sección de experimentación abarca el análisis y preparación de los datos, y la ingeniería de características. Finalmente, se presentan las conclusiones y líneas futuras de trabajo que emergieron a lo largo de la investigación.

Índice general

Índice de figuras

Índice de tablas

”

Lista de acrónimos

AA	Aprendizaje Automatico.
IA	Inteligencia Ariticial.
IAE	Inteligencia Artifical Explicable.
LIME	Local Interpretable Model Agnostic Explications.
MCD	Mapas Cognitivos Difusos.
PSO	Particle Swarm Optimization.

Capítulo 1

Introducción

En este capítulo se ofrece una visión general sobre la temática del estudio propuesto y cómo se llevará a cabo. El capítulo consta de 4 secciones:

- Contexto y justificación: Se detalla el contexto de aplicación del estudio y las razones que justifican su realización. Se hace una breve introducción a una serie de conceptos que serán relevantes a lo largo del proyecto. La definición y aplicación de dichos conceptos se encuentran detalladas en capítulos posteriores.
- Objetivos: Se establecen el objetivo general del trabajo propuesto, y una serie de objetivos específicos que se pretenden lograr en el transcurso del desarrollo del proyecto.
- Metodología: Se describe la metodología seguida en el desarrollo el proyecto. Se presentan las tareas llevadas a cabo en cada una de las fases que conforman la metodología aplicada. Por último, se establecen las relaciones entre estas actividades y el trabajo realizado.
- Estructura de la memoria: Se presenta la estructura y organización de la memoria, que incluye un breve resumen de cada capítulo, lo que permite al lector tener una idea general de la organización del documento.

En resumen, estas secciones ofrecen una visión integral del estudio propuesto, cubriendo desde el contexto y los objetivos, hasta la metodología y la estructura de la memoria. Esta información proporciona una base sólida para comprender el proyecto en su conjunto y seguir adecuadamente su desarrollo.

1.1. Contexto y Justificación

Imagina que recogemos un bolígrafo y un papel, y en ese papel apuntamos todas las tareas cotidianas que realizamos a lo largo del día, desde las más importantes hasta las más insignificantes. Si ejecutamos ese ejercicio y comenzamos a tachar las tareas en las que es necesaria la energía eléctrica, es probable que nos sorprendamos de la cantidad de actividades que dependen de la energía eléctrica; nuestra hoja se encontrará prácticamente llena de tachones. La dependencia de la energía eléctrica en nuestra vida cotidiana es evidente, y es necesaria para la realización de las tareas más básicas hasta las más complejas.

La vida sin energía eléctrica sería un desafío considerable, ya que afectaría a casi todos los aspectos de nuestra vida diaria. La comunicación, la comodidad, la alimentación, la limpieza, el entretenimiento y el trabajo se verían severamente limitados sin acceso a la energía eléctrica. En un escenario sin energía eléctrica, tendríamos dificultades para comunicarnos con otras personas debido a que la mayoría de los dispositivos de comunicación funcionan con electricidad. Sin la calefacción, el aire acondicionado y el ventilador tendríamos dificultades en las estaciones de invierno y verano. Sin electrodomésticos, muchas tareas domésticas serían más laboriosas. El entretenimiento también se vería afectado: la mayoría de los dispositivos utilizados para el ocio, como televisores, consolas de videojuegos y ordenadores, requieren de electricidad para funcionar. En el ámbito laboral, aquellos que dependen de la de tecnología verían sus trabajos seriamente afectados, lo que se traduce en pérdidas económicas tanto para las personas como las empresas.

¿Es malo uso de la energía eléctrica?, La respuesta es no. A todos nos gusta la comodidad que ofrece y el que diga que no te miente. La verdadera cuestión es, ¿es malo el uso que damos a la energía eléctrica? La respuesta es sí. El uso que le damos en la actualidad plantea numerosas cuestiones. Uno de los principales problemas es el uso no sostenible y excesivo de la energía eléctrica que tiene impactos negativos en el medio ambiente [1],[2], la salud humana [3],[4] y la economía [5],[6]. De acuerdo a [7], a medida que la población crece y se concentra en las ciudades, aumenta la demanda de energía de las ciudades ya que es necesario satisfacer nuevas necesidades de transporte, vivienda, industria y otros servicios. La eficiencia energética es un tema de gran interés para los individuos, empresas y gobiernos. En estos últimos años se han desarrollado numerosos trabajos que tienen como objetivo la creación de modelos de eficiencia energética [8], [9], [10].

Otro de los problemas actuales para los diferentes actores del sector energético, es la predicción de los precios de la energía, tanto a corto como a largo plazo. A corto plazo, la predicción de los precios es crucial para los operadores del sistema eléctrico, debido a la necesidad de gestionar la oferta y demanda en tiempo real. En particular, las predicciones permiten tomar decisiones informadas sobre la programación de la generación de energía, la compra y venta en los mercados energéticos, y la gestión de la red eléctrica para garantizar un suministro estable. Se han usado distintas técnicas en distintos trabajos para predecir los precios de la energía a corto plazo [11], [12], [13]. El estudio de la predicción de precios a largo plazo tiene en cuenta factores macroeconómicos, políticas energéticas, avances tecnológicos y cambios en la estructura del mercado. Los gobiernos y operadores del sistema eléctrico realizan un análisis de escenarios usando los modelos predictivos para evaluar posibles tendencias futuras y tomar decisiones estratégicas en consecuencia.

Resulta interesante en este contexto el desarrollo de modelos de conocimiento autoexplicables que permitan al usuario comprender la razón detrás de los resultados. En general, en los dominios críticos son relevantes los modelos autoexplicables. Un dominio crítico es aquel en el que un fallo en una predicción de un modelo de Aprendizaje Automático puede tener consecuencias catastróficas. Veamos el caso del dominio médico. Imagine a un doctor que examina los datos del historial y una serie de pruebas realizadas recientemente a un paciente con la intención de diagnosticarle una enfermedad; el doctor inserta todos

los datos tomados del paciente en un modelo de apoyo a la toma de decisiones que tiene una precisión diagnóstico de forma correcta de un 99 %. Resulta que el resultado ofrecido por el modelo es incorrecto y el paciente es diagnosticado mal. En esta situación se presentan de forma simplificada dos escenarios, un primer escenario donde se le diagnostica al paciente con una enfermedad que no tiene y se le da un susto, y un segundo escenario donde se le dice al paciente que no tiene esa enfermedad, lo que provoca un diagnóstico tardío que puede provocar la muerte del paciente debido a ese mal diagnóstico. En este segundo escenario, un grupo de expertos en medicina quiere saber cómo es posible que este paciente haya muerto si sus datos son prácticamente iguales a los datos de un paciente diagnosticado la semana pasada de forma correcta. En esta situación se nos presenta un grave problema, resulta que el modelo con una precisión 99 % no es capaz de darnos una respuesta por una sencilla razón, es una caja negra, solo conocemos sus entradas y sus salidas, pero desconocemos totalmente el proceso realizado para tomar esa decisión de forma incorrecta. El campo de la energía eléctrica también es un dominio crítico, imaginemos un modelo cuyo resultado es una predicción del consumo de energía de un hogar, un edifico industrial o un dispositivo. Si el modelo predice de forma incorrecta a la baja el nivel de consumo de un dispositivo, el dispositivo en cuestión se verá afectado y paralizado. En caso contrario, si el modelo predice a la alta, desperdiciamos recursos y dinero. En ambos dominios descritos, es necesario que el usuario confíe en los resultados ofrecidos por el modelo. El modelo debe ser explicable y proporcionar una explicación de por qué se ha llegado a ese resultado.

El presente proyecto se centra en la creación de modelos de conocimientos explicables para la estimación de los costes del consumo de energía. Resultó interesante durante el desarrollo del proyecto explorar el comportamiento de los modelos de conocimiento en otros dominios críticos como el médico. El trabajo detalla la creación de modelos de AA en los dominios de energía y médicos empleando la técnica de Mapas Cognitivos Difusos. Una vez creados los modelos, se analiza la explicabilidad proporcionada por la propia técnica, pero también, se emplean otros métodos de explicabilidad de la literatura para analizar la explicabilidad obtenida del modelo.

Los MCD se han empleado de forma exitosa en la construcción de modelos en ambos dominios, dentro del dominio médico encontramos los siguientes trabajos [14], [15], [16], por otro lado, en el dominio energético tenemos los siguientes trabajos [17],[18],[19]. Los trabajos anteriores se caracterizan por la creación de modelos de MCD basados en expertos. Este trabajo se centra en el uso de datos para la creación de los modelos predictivos usando MCD.

1.2. Objetivos

El objetivo principal del proyecto se divide en dos sub-objetivos: El primero es proponer un modelo autoexplicable de predicción, el cual fue usado para predecir el costo de consumo energético, o si el paciente tiene COVID-19 o Dengue. El segundo consiste en definir un enfoque de análisis de explicabilidad y explorar su uso en los distintos casos de estudio, comparando el análisis realizado con otras técnicas de explicabilidad.

Se plantean una serie de objetivos específicos, enfocados en las tareas que son necesarias para lograr los objetivos descritos. Los objetivos específicos son presentados a continuación:

- Realizar un análisis del estado de arte a nivel de explicabilidad en los MCD y en otras técnicas basadas en causalidad.
- Crear y evaluar los modelos predictivos basados en MCD en diferentes conjunto de datos.
- Desarrollar un método de análisis de explicabilidad basado en causalidad
- Realizar un análisis de explicabilidad para los modelos de MCD usando nuestro enfoque de causalidad y otros métodos de explicabilidad empleados en la literatura.

1.3. Metodología

Para la realización del proyecto se usará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [20], que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, siendo la más utilizada en este tipo de proyectos. Basada en CRISP-DM, las fases de trabajo son las siguientes (Ver figura 1.1):

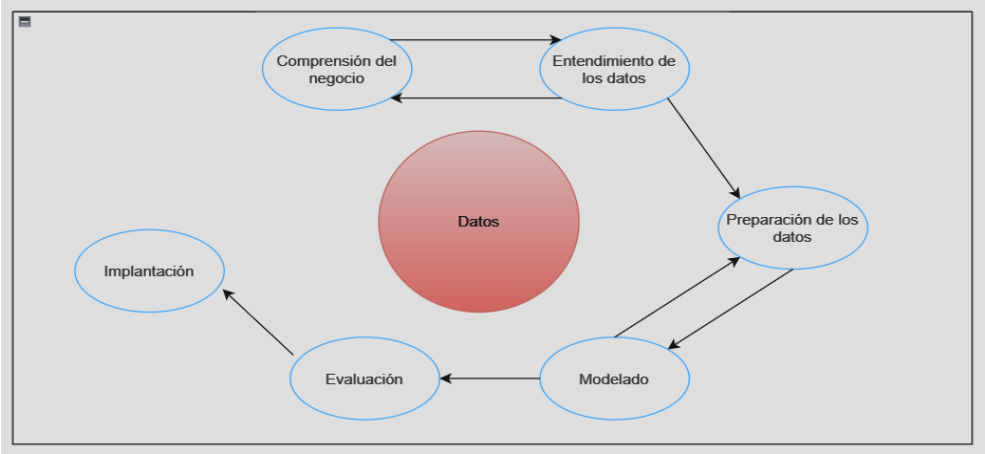

Figura 1.1: Metodología de desarrollo de CRISP-DM

1. Comprensión del negocio: Esta fase busca entender los objetivos del proyecto desde la perspectiva del negocio, es decir, cuáles son las necesidades que nuestro proyecto debe satisfacer. Se corresponde con la elaboración del anteproyecto, la introducción y estado del arte.
2. Entendimiento de los datos: En esta fase se realiza la recopilación de los datos iniciales, se describen, se realiza una exploración inicial y verificación de la calidad de los mismos. Se corresponde con el proceso de elección de los conjuntos de datos usados en el proyecto.

3. Preprocesamiento de los datos: Esta fase implica varias subtareas como el análisis exploratorio de datos, selección de características, transformación de características, creación de características nuevas y manejo de valores faltantes y atípicos. El resultado final de esta fase son los conjuntos de datos usados en la construcción de los modelos.

4. Modelado: Implica la construcción de los modelos de conocimiento. El trabajo [15] describe una metodología de desarrollo de MCD basada en expertos. Se propone una adaptación de la metodología a nuestro contexto de construcción de modelos a partir de conjuntos de datos. Al enmarcar el proceso de desarrollo de MCD dentro de la metodología de CRISP-DM, distintas fases del desarrollo de MCD se encuentran encapsuladas en las fases de trabajo de CRISP-DM. Las fases de desarrollo de un MCD basado en datos se describen a continuación (Ver figura 1.2):

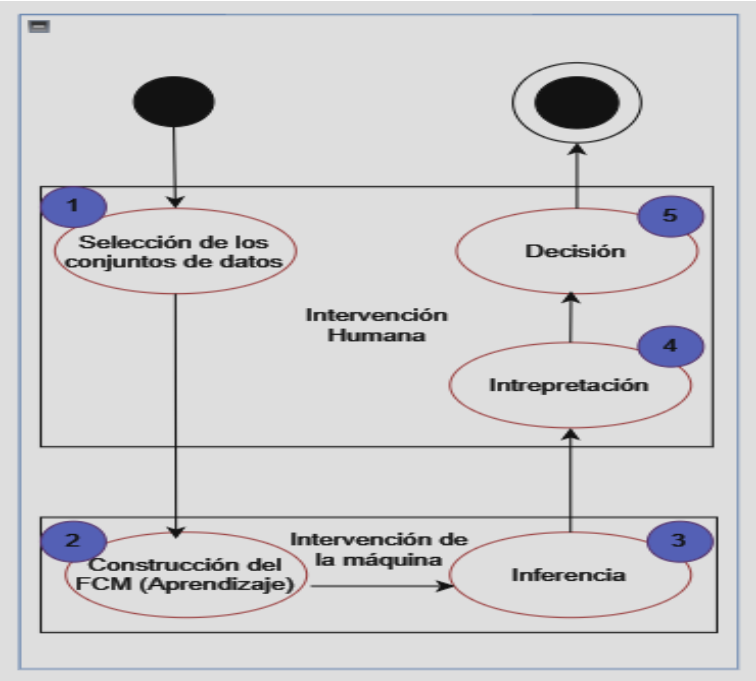

Figura 1.2: Metodología de desarrollo de un MCD

   a) Selección de los conjuntos de datos: Se corresponde con la fase de Entendimiento de los datos de CRISP-DM.

   b) Construcción del FCM: Es la fase de modelado en CRISP-DM.

   c) Inferencia: El proceso de inferencia es explicado en la parte teórica correspondiente a los MCD.

   d) Interpretación: Corresponde con la segunda parte de la fase de Evaluación de CRISP-DM, donde se realiza una análisis de explicabilidad del modelo obtenido.

   e) Decisión: Corresponde a qué decisión realiza el usuario de acuerdo a los resultados ofrecidos por el modelo.

5. Evaluación: Esta fase comprende la evaluación y discusión de los resultados obtenidos con los modelos. Por un lado, se evalúa el rendimiento del modelo empleando métricas clásicas, y por otro lado, se efectúa el análisis de explicabilidad con el enfoque propuesto en este trabajo y otros métodos de explicabilidad usados en la literatura.

6. Implantación: Es la presentación de los resultados obtenidos. Se encuentra vinculado a la redacción de este manuscrito con los resultados del proyecto.

1.4. Estructura

El documento consta de los siguientes capítulos:

1. Introducción: detalla el contexto y los motivos que han llevado al desarrollo del proyecto; se realiza una breve descripción del problema y la solución presentada. Se presentan los objetivos generales y objetivos específicos a conseguir durante el desarrollo del trabajo. Por último, se expone una breve descripción de las metodologías aplicadas, CRISP-DM y la metodología empleada en el diseño de los MCD, y la relación existente entre ambas metodologías.

2. Base teórica: En primer lugar, se expone el estado del arte del proyecto, el cual se divide en dos secciones:

   a) Estudios sobre el desarrollo y creación de MCD basados en conjuntos de datos.

   b) Estudios de enfoques de explicabilidad basados en MCD.

   Por último, se explican los conceptos y técnicas usadas en el desarrollo del proyecto, tales como los MCD y las técnicas de explicabilidad usadas, entre otras.

3. Desarrollo práctico (Experimentación): Aplicación de todos los conceptos y técnicas descritas en la base teórica, para la obtención de los modelos predictivos, y la discusión sobre el rendimiento y explicabilidad de los modelos y técnica empleada.

4. Conclusión y líneas futuras: Se hace un análisis de las ventajas e inconvenientes de las técnicas empleadas en la construcción y explicación de los modelos. Además, se proponen posibles ampliaciones futuras a nivel de las técnicas para la creación de los modelos y para el análisis de explicabilidad, entre otras cosas.

Capítulo 2

Estudio teórico

2.1. Introducción

En este capítulo se presentan los conceptos y técnicas aplicados en las distintas fases de desarrollo del trabajo. Se exponen las distintas técnicas empleadas para el preprocesamiento de los datos, construcción del modelo, y evaluación de rendimiento y explicabilidad del modelo.

2.2. Estado del Arte

En el capítulo 1 se han presentado trabajos que proponen modelos predictivos en los dominios de energía eléctrica y médico empleando los MCD. En su mayoría, estos trabajos se basan en la creación de modelos MCD basados en expertos. En esta sección se analizan estudios que emplean la técnica de los MCD basados en datos. Por otro lado, se realiza un estudio de enfoques de explicabilidad basados en la propia técnica de los MCD u otras técnicas basadas en causalidad.

2.2.1. Mapas Cognitivos Difusos basados en conjuntos de datos

Prácticamente, la totalidad de la literatura en ambos dominios de estudio se centra en la creación de MCD basados en expertos. Solo se han encontrado dentro del dominio médico los trabajos [21] y [22] que crean los MCD a partir de conjuntos de datos, y realizan el aprendizaje por medio del algoritmo de optimización de PSO (que se explica en la sección dedicada a MCD). En la literatura, respecto al dominio energético, no se ha encontrado información.

Es curiosa la existencia de tan pocos trabajos relacionados. De acuerdo a [23] existe un motivo, el problema no reside en los algoritmos de aprendizaje supervisados basados en datos sino en la construcción del propio modelo. Un MCD es un grafo con conceptos y relaciones entre conceptos. Las relaciones establecen el grado de causalidad entre dichos conceptos. Cuando los expertos definen un MCD, establecen los conceptos y las relaciones, y después se aplica un aprendizaje no supervisado para la obtención del MCD. Es aquí donde surge el problema. Por un lado, los MCD definidos por expertos han ofrecido buenos resultados en la literatura, y por otro lado, en términos de rendimiento, los algoritmos de creación de este MCD inicial basados en datos no han demostrado ser tan potentes como los definidos por expertos.

2.2.2. Estudio de enfoques de explicabilidad basados en la técnica de los Mapas Cognitivos Difusos

La naturaleza inherente de los MCD permite crear un ranking de las características más influyentes en el modelo, utilizando el peso de las relaciones causales con respecto a la variable predictiva. Esto posibilita una explicación más clara y comprensible de cómo se toman decisiones dentro del modelo, lo que resulta en una mayor transparencia e interpretabilidad. En la literatura, se ha investigado el uso de MCD como modelos explicativos auxiliares para otros modelos más complejos de caja negra [24],[25].

Además, existen artículos que se centran en un análisis más profundo de la explicabilidad en los MCD, utilizando la sensibilidad local como una herramienta clave. Estos estudios tienen como objetivo comprender en detalle cómo pequeñas variaciones en las variables de entrada pueden afectar significativamente el comportamiento y la estabilidad del modelo. Al emplear el análisis de sensibilidad local, los investigadores obtienen una comprensión completa de cómo responde cada concepto en el MCD a los cambios en las variables de entrada, y cómo estas variaciones se propagan a través de las relaciones causales. Este nivel más profundo de explicabilidad mejora la interpretación y comprensión de los MCD, lo que los convierte en valiosos para diversas aplicaciones, como la toma de decisiones, la modelización de sistemas complejos y el análisis de riesgos.

Algunos estudios exploran la sensibilidad local de conceptos en los MCD. En el artículo [26], el autor demostró que los resultados del MCD dependen particularmente de las relaciones con pesos elevados. Además, se observó que los cambios en los valores de estos pesos pueden tener efectos significativos en los resultados del análisis del MCD. En el artículo [27], los autores miden la robustez de su algoritmo en un MCD utilizando diferentes métodos de análisis de propagación de incertidumbre. En ambos estudios, se evalúa la sensibilidad local de los conceptos como un indicador de la calidad de la robustez, midiendo los cambios en las predicciones al variar los pesos de las relaciones.

2.3. Preparación de los datos

En el ámbito del AA y la ciencia de datos, la obtención de un modelo con un buen rendimiento es un tarea complicada si no se entienden y prepararan adecuadamente los conjuntos de datos. La ingeniería de características es un proceso fundamental para mejorar la calidad de los datos y permitir que los modelos de aprendizaje automático sean más efectivos.

Este es un proceso iterativo que requiere experimentación y pruebas para encontrar la mejor combinación de variables para un problema determinado. El éxito de un modelo de AA depende en gran medida de la calidad de las características utilizadas en el modelo. Las técnicas de preprocesamiento de los datos empleadas en este proyecto se pueden subdividir en los siguientes procesos:

- Análisis exploratorio de los datos: Se usan métodos como la correlación para la visualización de relaciones entre las variables, se exploran los datos en busca de valores faltantes o atípicos, entre otras cosas.
- Transformación de características: Implica modificar las características existentes para mejorar su distribución o expresividad. Algunos ejemplos comunes son la normalización, la estandarización, la discretización, la codificación one-hot y la transformación logarítmica.
- Selección de características: En ocasiones, algunas características pueden no ser relevantes para el modelo. La selección de características implica identificar y mantener aquellas características más informativas para predecir la variable objetivo.

2.3.1. Análisis exploratorio de los datos

2.3.1.1. Correlación de Pearson

La correlación de Pearson es una medida estadística utilizada para evaluar la relación lineal entre dos variables continuas. Una relación lineal es una asociación entre dos variables que puede ser descrita mediante una línea recta en un gráfico. Existen dos tipos de relación lineal:

- Relación lineal creciente: Si el valor de una variable aumenta, el valor de la otra variable también aumenta de forma proporcional; si el valor de una variable disminuye, el valor de la otra variable también disminuye de manera proporcional.
- Relación lineal decreciente: Si el valor de una variable aumenta, el valor de la otra variable disminuye de forma proporcional; si el valor de una variable disminuye, el valor de la otra variable aumenta de manera proporcional.

Ecuación 2.1 muestra la fórmula para el cálculo de la correlación de Pearson, donde S_{XY} es la covarianza de las variables X e Y y S_X e S_Y son las desviaciones típicas de cada una de las variables. El valor del coeficiente de Pearson esta comprendido en el intervalo [-1,1]. Dependiendo del tipo de relación lineal entre las variables, el coeficiente de la correlación de Pearson describe los siguientes escenarios (ver figura 2.1):

- $r > 0$: Indica la existencia de una correlación positiva, si el valor de una de las variables aumenta, el valor de la otra variable también aumenta. Si el valor de una de las variables disminuye, el valor de la otra variable también disminuye. Cuanto más cerca esté el valor del coeficiente a 1, mayor es la correlación existente entre ambas variables. Un coeficiente de correlación de Pearson de 1 indica una correlación positiva perfecta, lo que implica que las dos variables tienen una relación lineal positiva fuerte.
- $r = 0$: Indica que no existe una relación lineal entre las dos variables.
- $r < 0$: Indica la existencia de una correlación negativa, si el valor de una de las variables aumenta, el valor de la otra variable disminuye. Si el valor de una de las variables disminuye, el valor de la otra variable aumenta. cuanto más cerca esté el valor del coeficiente a -1, mayor es la correlación existente entre ambas variables. Un coeficiente de correlación de Pearson de -1 indica una correlación negativa perfecta, lo que implica que las dos variables tienen una relación lineal negativa fuerte,

$$r = \frac{S_{XY}}{S_X S_Y} \tag{2.1}$$

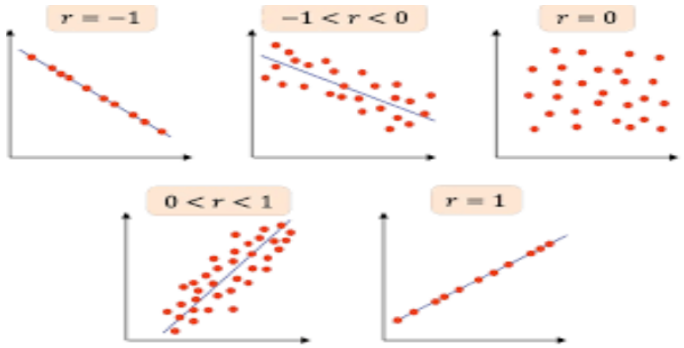

Figura 2.1: Escenarios de Correlación de Pearson

2.3.1.2. Correlación de Spearman

La correlación de Spearman es una medida estadística que evalúa la relación monótona entre dos variables. Una relación monótona se refiere es una asociación entre dos variables en la que una variable tiende a aumentar o disminuir sistemáticamente a medida que la otra variable también lo hace, sin necesariamente seguir una relación lineal, independientemente de si la relación es lineal o no. En una relación monótona, las variables tienden a moverse en la misma dirección relativa, pero no necesariamente a un ritmo constante. A diferencia de la correlación de Pearson, que se basa en las relaciones lineales, la correlación de Spearman se basa en los rangos de los datos.

La ecuación 2.2 muestra la fórmula para el cálculo de la correlación de Spearman, donde d_i es la diferencia de rangos entre dos variables y n es el número de observaciones de la muestra. El valor de la correlación de Spearman esta comprendido en el intervalo [1,-1]. El proceso de cálculo de la correlación de Spearman es el siguiente:

1. Se clasifican los valores de cada variable de manera independiente, asignando un rango a cada valor (el menor valor obtiene el rango más bajo y el mayor valor obtiene el rango más alto). Si hay empates, se asigna el promedio de los rangos correspondientes.
2. Se calcula la diferencia de rangos para cada par de observaciones (X, Y), conocido como d_i.
3. Se aplica la fórmula de la correlación de Spearman.

El valor del coeficiente de Spearman esta comprendido en el intervalo [-1,1], dependiendo del tipo de relación monótona entre las variables, el coeficiente de la correlación de Spearman describe los siguientes escenarios (ver figura 2.2):

- $r_S > 0$: Indica una correlación monótona creciente entre las variables. Si una variable crece, la otra variable nunca decrece. Cuanto más cerca esté el valor del coeficiente a 1, mayor es la correlación existente entre ambas variables. Un coeficiente de correlación de Spearman de 1 significa una perfecta asociación positiva de rangos entre las variables.
- $r_S = 0$: Indica que no existe una correlación monótona entre las variables, lo que no implica que no exista otro tipo de relación entre las variables.
- $r_S < 0$: Indica una relación monótona decreciente entre las variables. Si una variable crece, la otra variable nunca crece. Cuanto más cerca esté el valor del coeficiente a -1, mayor es la correlación existente entre ambas variables. Un coeficiente de correlación de Spearman de -1 significa una perfecta asociación negativa de rangos entre las variables.

$$r_S = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \tag{2.2}$$

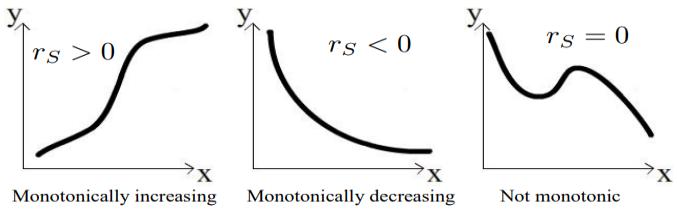

Figura 2.2: Escenarios de Correlación de Spearman

2.3.1.3. V de Cramér

El coeficiente de contingencia de Cramér (V de Cramér), es una medida que cuantifica la relación entre dos variables categóricas dentro de una tabla de contingencia. Es usada con la medida estadística Chi cuadrado cuando es necesario precisar la fuerza de asociación o relación entre variables. El valor del Coeficiente de Cramér está comprendido dentro del intervalo [0,1]. Un valor de V cercano a 0 indica que las variables son independientes o no están asociadas, y un valor cercano a 1 indica una fuerte asociación entre las variables. La ecuación 2.3 muestra la formula para el cálculo del coeficiente de contingencia de Crammer, donde X^2 es el estadístico Chi cuadrado calculado para la tabla de contingencia, n es el tamaño de la muestra, r es el número de filas de la tabla de contingencia y c es el número de columnas de la tabla de contingencia.

$$V = \sqrt{\frac{X^2}{n(min[r,c]-1)}} \tag{2.3}$$

El cálculo del coeficiente de contingencia de Cramér implica crear la tabla de contingencia de **Chi cuadrado**. Chi cuadrado es una medida estadística que evalúa la relación entre dos variables categóricas. Una variable categórica, cualitativa o discreta, puede tomar un valor de un conjunto limitado de valores conocido como categorías o niveles.

Chi cuadrado aplica una prueba de independencia entre dos variables, evaluando si existe una relación significativa entre ellas en función de las distribuciones observadas en una tabla de contingencia. El cálculo de chi cuadrado sigue los siguientes pasos:

1. Creación de la tabla de contingencia: Muestra las frecuencias observadas para las diferentes categorías o niveles de las variables.

2. Definición de hipótesis: Se define la hipótesis nula H_0 y la hipótesis alternativa H_1. La hipótesis nula establece que las variables son independientes, mientras que la hipótesis alternativa sostiene que existe una correlación entre las variables.

3. Cálculo de Chi cuadrado: Se aplica la fórmula de la ecuación 2.4, donde 0_{ij} es la frecuencia observada en la celda i-ésima fila y j-ésima columna de la tabla de contingencia, E_{ij}, es la frecuencia esperada en la misma celda bajo la suposición de independencia entre las variables, y f y c son el número de filas y número de columnas de la tabla.

$$X^2 = \sum_{i=1}^{f}\sum_{j=1}^{c}\frac{(0_{ij}-E_{ij})^2}{E_{ij}} \tag{2.4}$$

4. Cálculo de los grados de libertad: Se aplica la fórmula de la ecuación 2.5, donde p es el número de filas y q el número de columnas de la tabla de contingencia.

$$df = (p-1)(q-1) = 1 \tag{2.5}$$

5. Análisis del valor de Chi cuadrado: Se usan los grados de libertad y valor de Chi cuadrado para encontrar el valor crítico de chi cuadrado en una tabla de distribución chi cuadrado. Si el valor calculado de chi cuadrado es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye concluye que hay una relación entre las variables.

2.3.1.4. Detección de valores atípicos (Sustitución por mediana) y faltantes (Eliminación de los registros)

La presencia de valores atípicos y faltantes en las características de nuestros conjuntos de datos afectan el rendimiento del modelo, por lo que es necesaria su detección y manejo. Los valores atípicos, son observaciones que se desvían significativamente del resto de valores dentro de un conjunto de datos. Son resultado de las siguientes circunstancias:

- Errores de medición.
- Errores cometidos en el muestreo.
- Problemas en la recopilación de los datos.
- Fallos causados por la metodología de la investigación
- Anomalías o eventos inusuales dentro del conjunto de datos que contienen información relevante del comportamiento de los datos.

Existen varias técnicas para la detección valores atípicos:

- Análisis de valores extremos: Se basa en la identificación de observaciones que se encuentran por encima o por debajo de un umbral predefinido. Se usan medidas estadísticas en la definición del umbral como el rango intercuartil (IQR), el z-score, o criterios basados en desviaciones estándar para identificar los valores atípicos.
- Gráficos de dispersión: Pueden revelar relaciones inusuales entre dos variables. Las observaciones que se alejan de la tendencia general pueden ser identificadas como valores atípicos.
- Basados en modelos: Estos técnicas usan modelos de AA en la detección de atípicos. Algunos ejemplos son los métodos de detección de anomalías basados en clustering, árboles de decisión o modelos probabilísticos.

Existen distintos métodos para el tratamiento de valores atípicos:

- No aplicar un tratamiento: Si los valores atípicos contienen información relevante del comportamiento de los datos en el sistema estudiado.
- Eliminación: Se eliminan las observaciones/registros que contienen los valores atípicos, viable si el porcentaje de valores atípicos es pequeño respeto al tamaño total del conjunto de datos.
- Truncamiento: Se establece un límite superior o inferior para los valores extremos.
- Reemplazo: Los valores atípicos son reemplazados por un valor representativo de la característica como la media, mediana o moda.

Los valores faltantes son aquellos que no están presentes en un conjunto de datos. Estos valores pueden ser el resultado de diversos factores:

- Errores en la entrada de los datos.
- Problemas de medición.
- Datos no recopilados.

Existen distintas técnicas para el tratamiento de valores faltantes:

- Eliminación: Se eliminan las observaciones/registros con valores faltantes. Es viable si los datos faltantes son un pequeño porcentaje del conjunto de datos total. Es el usado en este trabajo.
- Reemplazo: los valores faltantes son reemplazados por un valor representativo de la característica como la media, mediana o moda.
- Basados en modelos: Estos técnicas usan modelos de AA para la predicción del valor faltante, algunos ejemplos son la regresión lineal y los árboles de decisión.

Método del Rango intercuartílico: El rango intercuartílico (RIC) es una medida estadística usada en la detección de valores atípicos, se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). La ecuación 2.6 muestra el cálculo del RIC.

$$RIC = Q3 - Q1 \tag{2.6}$$

Utilizando este rango intercuartílico, es posible identificar valores atípicos. Para ello, se pueden establecer los nuevos umbrales superior e inferior que se encuentren a cierta distancia multiplicada por el rango intercuartílico desde el tercer cuartil (Q3) y el primer cuartil (Q1), respectivamente. Estos umbrales proporcionan un marco de referencia para identificar observaciones que se encuentran significativamente por encima del tercer cuartil o por debajo del primer cuartil. Comúnmente se usa para definir la distancia el valor 1.5 pero puede ajustarse según el contexto y la tolerancia a valores atípicos. Las ecuaciones 2.7 y 2.8 muestran el cálculo de ambos umbrales.

$$UmbralInferior = Q1 - 1{,}5RIC \tag{2.7}$$

$$UmbralSuperior = Q3 + 1{,}5RIC \tag{2.8}$$

El rango intercuartil está considerado un estadístico robusto por su baja exposición a valores atípicos, debido a que solo considera las observaciones del primer y tercer cuartil en su cálculo, por lo tanto solo tiene en cuenta los valores más cercanos a la mediana. El método del rango intercuartílico se apoya de manera visual en el uso de diagramas de caja y bigotes.

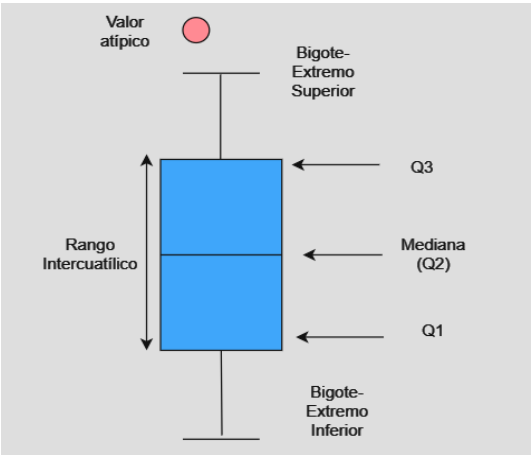

Figura 2.3: Diagrama de Caja y Bigotes

El tratamiento de valores faltantes se basa en la eliminación de los registros que los contienen. Como se verá, en la parte experimental, la cantidad de registros con valores faltantes es mínima en comparación al tamaño total de los conjuntos de datos.

El tratamiento de los valores atípicos se basa en la sustitución por mediana, lo que implica reemplazar estos valores por el valor de la mediana en la variable. Se escogió la mediana en vez de la media ya que es una medida más robusta a los valores atípicos. Por otro lado, la detección de los valores atípicos es realizada empleando el rango intercuartílico.

2.3.2. Transformación de características

2.3.2.1. Discretizacion de Datos (Basada en Rangos)

La discretización de datos es un proceso que transforman valores de una variable continua en valores discretos, los valores continuos se agrupan en rangos o categorías. La discretizacion de variables es útil por distintos aspectos:

- Varios modelos de AA tienen menores tiempo de ejecución al usar valores discretos.
- Reducción de valores atípicos y redundancia: Al agrupar valores continuos en categorías
- Transformación de un problema de predicción en un problema de clasificación.

Los algoritmos de discretización pueden clasificarse en algoritmos de discretización supervisada y discretización no supervisada. En los algoritmos de discretización no supervisados se especifica el número de clases o el número de observaciones en cada clase. En los algoritmos de discretización supervisada, la discretización se ejecuta en función de cálculos basados en la entropía y la pureza. En este trabajo, se usa la discretización para convertir una variable continua dependiente en una variable discreta dependiente, y así convertimos un problema de predicción de valores continuos en un problema de predicción de clases (valores discretos). Debido a la falta de una implementación directa de algoritmos de discretización supervisada, se optó por utilizar algoritmos de discretización no supervisada y emplear un enfoque de prueba y error para determinar el número óptimo de clases en el proceso de discretización. Algunos de los algoritmos de discretización no supervisada más usados en la literatura son:

- Discretización basada en rangos (usado en este trabajo): Divide el rango de valores continuos en k clases del mismo ancho de posibles valores para la variable.
- Discretización basada en cuantiles: Cada clase k contiene la misma cantidad de valores, divididos en función de los cuantiles.
- Discretización basada en densidades: Usan algoritmos de clustering como k-means, DBSCAN para agrupar los datos dentro de cada clase.

2.3.2.2. Normalización de Datos (Min-Max)

La normalización de datos es un proceso utilizado para estandarizar los valores de las variables en un conjunto de datos. Se busca que todas las variables tengan una escala similar de forma que ninguna variable tenga una influencia desproporcionada en el análisis o comportamiento del modelo. Existe una amplia literatura que analiza el impacto de la normalización de los datos en el rendimiento de modelos de clasificación en múltiples conjuntos de datos como [28], [29]. También existe una extensa literatura

que aborda la importancia de la normalización de los datos en modelos de AA específicos como en k-vecinos [30], redes neuronales [31] o máquinas de vectores[32]. En base a esta literatura, se realizó una normalización de los datos aplicando la técnica de normalización Mín-Max. A continuación se presenta la definición matemática de Min-Max y de otras técnicas de normalización muy usadas en la práctica:

- Min-Max: Los valores se escalan al restar por el valor mínimo y dividirlo por la diferencia entre el valor máximo y el valor mínimo, los datos se encuentran en el rango $\in [0, 1]$

$$X_{norm} = \frac{x - min(x)}{max(x) - min(x)} \in [0, 1] \tag{2.9}$$

- Puntuación Z: los valores se escalan para tener una media de 0 y una desviación estándar de 1. Se resta la media de cada valor y se divide por la desviación estándar.

$$X_{norm} = \frac{x - \mu(x)}{\sigma(x)} \in [-\infty, \infty] \tag{2.10}$$

- Decimal: Los valores se escalan al dividirlos por la potencia de 10 del valor máximo absoluto.

$$X_{norm} = \frac{x}{10^d} \in [-1, 1] \tag{2.11}$$

2.3.2.3. Balanceo de Datos (sobremuestreo SMOTE)

El balanceo de datos es un proceso utilizado para ajustar la cantidad de muestras de cada clase. De acuerdo a numerosas investigaciones [33],[34],[35],[36], cuando hay una desproporción de las clases en un conjunto de datos, los modelos de AA tienden a favorecer a la clase mayoritaria. Esto se debe a que los modelos buscan maximizar la precisión global y pueden tener dificultades para aprender patrones en la clase minoritaria. El balanceo de clases permite que el modelo aprenda de manera equitativa en torno a todas las clases. Existen varias técnicas de balanceo de datos comunes, incluyendo:

1. Submuestreo: Reducción del número de observaciones de la clase mayoritaría para equilibrarla respecto al número de observaciones de la clase minoritaria.

2. Sobremuestreo: Aumento del número de observación de la clase minoritaria para equilibrarla respecto al número de observaciones de la clase mayoritaria.

3. Híbrido: Combinan estrategias de submuestreo y sobremuestreo para equilibrar las clases y mejorar el rendimiento del modelo de AA.

SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) es una técnica de sobremuestreo definida en el artículo [37]. A diferencia de otras técnicas de sobremuestro que duplican instancias de la clase minoritaria, SMOTE aumenta el número de observaciones de la clase minoritaria mediante la generación de muestras sintéticas. El proceso de SMOTE es el siguiente:

1. Selección de una instancia de la clase minoritaria.

2. Encontrar los k vecinos más cercanos de la instancia pertenecientes a la misma clase calculando la distancia euclídea (se definen como x_k).

3. Generar una nueva instancia sintética calculando una combinación lineal entre la instancia seleccionada y uno de sus vecinos cercanos.

$$x\prime = x + aleatorio(0,1)|x - x_k| \tag{2.12}$$

4. Repetir los pasos del 1 a 3 hasta generar el número de observaciones necesarias.

2.3.3. Selección de Características (ANOVA)

La selección de características (variables) en el AA es el proceso de elegir un subconjunto relevante de características de entrada para construir un modelo de AA. Los objetivos de la selección de características son:

- Reducir la dimensionalidad del conjunto de datos.
- Eliminar características irrelevantes que no aportan información y actúan como ruido en la construcción del modelo.
- Mejorar el entendimiento de los datos.
- Menor tiempo de ejecución para la construcción del modelo.

A continuación, se hace una presentación de las técnicas más utilizadas en la selección de características en el AA:

- Métodos de Filtrado (ver figura 2.4): Se evalúa la relevancia de cada característica de forma individual, las variables se ordenan de acuerdo con algún índice de relevancia de forma que las variables con los valores de relevancia menores puedan ser eliminadas. El modelo se entrena con el subconjunto de variables seleccionadas (usado en este trabajo).

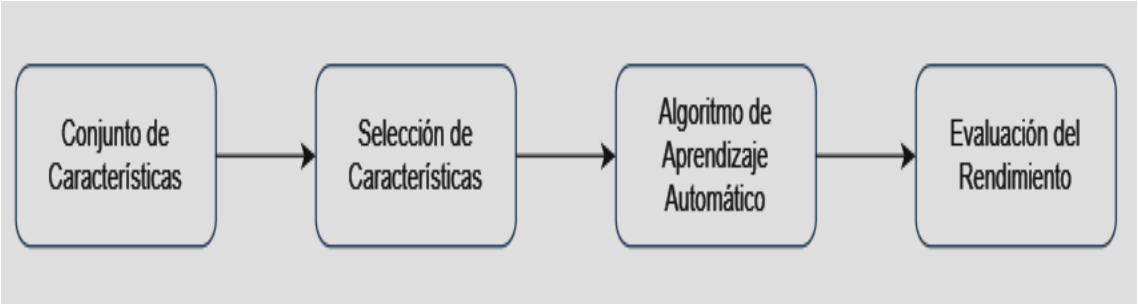

Figura 2.4: Selección de Características basada en Filtros

- Métodos de envoltura (ver figura 2.5) : Usan el modelo de AA para evaluar subconjuntos de variables, seleccionando aquellas que mejor se ajusten al modelo. Implica la construcción y evaluación de manera iterativa de modelos usando diferentes subconjuntos de características.
- Métodos integrados: el algoritmo de selección de características esta integrado dentro del modelo de AA, combina las cualidades de los métodos de filtrado y envoltura.

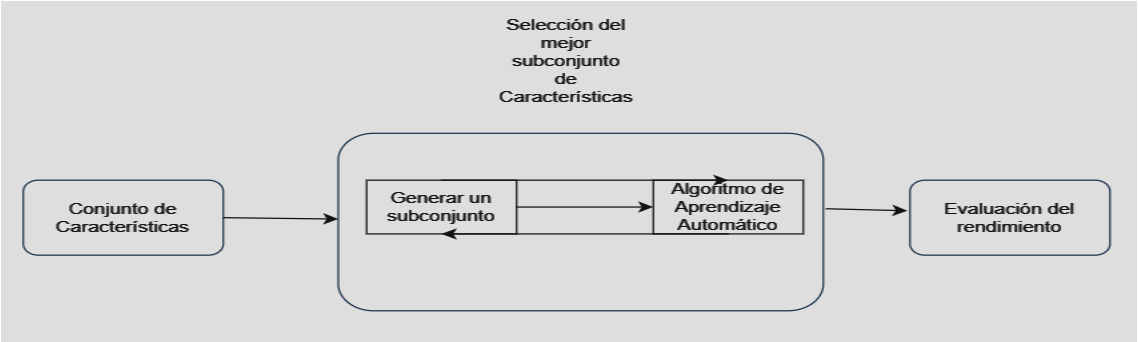

Figura 2.5: Selección de Características basada en Envoltura

2.3.3.1. ANOVA unidireccional modificado

ANOVA es una técnica estadística que compara las medias de tres o más grupos independientes en relación con una variable dependiente continua, Se utiliza típicamente en problemas de selección de características para tareas de clasificación, donde la variable dependiente es categórica y las variables independientes son numéricas. Por lo tanto, las variables numéricas son evaluadas para determinar su importancia en la tarea de clasificación. No evalúa la relación entre las variables dependientes y la variable objetivo, sino que calcula una medida de importancia para cada característica en términos de su capacidad para distinguir entre las clases. El proceso de selección de características usando ANOVA unidireccional modificado es el siguiente:

1. Cálculo de la variabilidad: Por cada una de las variables independientes del conjunto de datos, se calcula la varianza dentro de cada clase (varianza intraclase), que mide cuanto varían los valores de cada variable dentro de cada clase por separado. La ecuación 2.13 muestra el cálculo de la varianza intraclase, donde k es el número de clases de la variable objetivo, n_i es el tamaño de la clase i (número total de instancias en la clase i), x_{ij} es el valor de la instancia j en la clase i y $\bar{x_i}$ es la media de la clase i.

$$MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2}{\sum_{i=1}^{k}(n_i - 1)} \tag{2.13}$$

   Se calcula la varianza total dentro de todas las clases combinadas (varianza total), que es una medida de la variabilidad total de la característica en el conjunto de datos. La ecuación 2.14 muestra el cálculo de la varianza total, donde k es el número de clases de la variable objetivo, n_i es el tamaño de la clase i (número total de instancias en la clase i), x_{ij} es el valor de la instancia j en la clase i, $\bar{x}_{total}$ es la media de todos los valores del conjunto de datos y N es el número total de instancias del conjunto de datos.

$$MSB = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{total})^2}{N - 1} \tag{2.14}$$

2. Estadístico F: Se compara la variabilidad calculando el estadístico F entre la variabilidad entre las clases (varianza total), y la variabilidad dentro de las clases (varianza intraclase). La ecuación 2.15 muestra el cálculo del estadístico F.

$$F = \frac{MSB}{MSW} \tag{2.15}$$

3. Ranking de Características: Ordenación de las características de acuerdo al valor F obtenido para cada una de las variables independientes en el análisis. Cuánto mayor es el valor F mejor posición adquiere la característica dentro del ranking.

4. Selección de características: Selección del subconjunto de características a usar en el entrenamiento del modelo, el subconjunto se encuentra formado por las K primeras características dentro del ranking. Dicho valor K es escogido por el usuario en base al ranking de características creado en el paso anterior.

Por último, destacar que existen otras técnicas como PCA (Análisis de componentes) y MCA (Análisis de correspondencias múltiples) que reducen la dimensionalidad de un conjunto de datos aplicando un proceso matemático, el cuál transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables. El motivo de no explorar este tipo de técnicas es simple, el proyecto se centra en explicar los modelos creados, si encapsulamos las variables/dimensiones en nuevas variables, se pierde toda la capacidad de explicabilidad de los modelos.

2.4. Mapas Cognitivos Difusos

Los MCD fueron introducidos por Kosko en 1986 [38]. Surgen a partir de la lógica difusa definida por Lofti Zadeh en 1965 [39] y los mapas cognitivos creados por Axelrod en 1976 [40]. Axelrod introdujo las mapas cognitivos para representar el conocimiento de los científicos sociales. Luego Kosko amplió su definición al introducir valores de lógica difusa.

Los MCD se emplean para modelar sistemas complejos debido a su facilidad de construcción e interpretación [15]. Un MCD es un grafo dirigido cuyos vértices representan conceptos, mientras que las aristas se utilizan para expresar relaciones causales entre los conceptos, lo que permite realizar un análisis de explicabilidad basado en causalidad en el grafo.

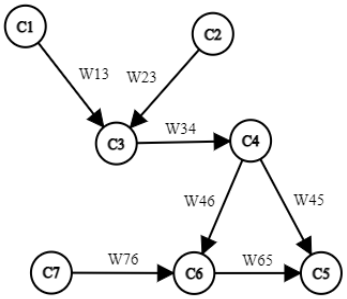

Figura 2.6: Representación gráfica de un Mapa cognitivo difuso simple

La figura 2.6 presenta el grafo de un MCD simple, el cual consta de siete (7) conceptos y siete (7) aristas ponderadas. Cada concepto representa una variable, entidad, evento, condición u otro factor relevante del sistema. Existen relaciones entre conceptos que están representadas por un peso en la arista dirigida, estas relaciones muestran el grado de influencia de un concepto sobre otro [41], básicamente son relaciones causales. Los conceptos y las relaciones causales constituyen los elementos fundamentales de un MCD, lo que le confieren su capacidad autoexplicativa.

Cada concepto se caracteriza por tener asociado un valor a_i. Por lo general, este valor está limitado al intervalo [0,1]. Las relaciones causales entre conceptos permiten grados de causalidad, por lo que los valores de los pesos pueden oscilar en el intervalo [-1,1]. De acuerdo a Aguilar [41], en general, la tarea de crear un MCD y asignar valores a los pesos de las relaciones la realizan expertos en un determinado dominio, pero también podemos usar los datos sin procesar para esta tarea. Existen tres posibles tipos de relación causal entre el concepto c_i y el concepto c_j , según el signo del peso w_{ij}:

- si $w_{ij} < 0$ entonces hay una relación de causalidad negativa entre los conceptos c_i y c_j. La disminución/aumento del valor del concepto c_i produce un aumento/disminución del valor del concepto c_i con intensidad $|w_{ij}|$.
- si $w_{ij} > 0$ entonces hay una relación de causalidad positiva entre los conceptos c_i y c_j. La disminución/aumento del valor del concepto c_i produce una disminución/aumento del valor del concepto c_i con intensidad $|w_{ij}|$.
- Si $w_{ij} = 0$ no hay relación causal entre los conceptos c_i y c_j

El valor del peso w_{ij} indica el grado de influencia entre el concepto c_i y el concepto c_j. Matemáticamente, los MCD se definen por la 4-tupla (C,W,A,F) donde:

- $C = [c_1, ..., c_m]$ es el conjunto de m conceptos que son las variables (nodos del grafo) que componen el sistema. La figura 2.6 muestra un MCD con siete conceptos.
- W es la matriz de adyacencia que representa las relaciones de causalidad (aristas del grafo) entre los conceptos. A continuación, se muestra la matriz de adyacencia del MCD de la figura 2.6.

$$W = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 \end{array} \\ \begin{pmatrix} 0 & 0 & w_{13} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{45} & w_{46} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{65} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{76} & 0 \end{pmatrix} \end{array} \tag{2.16}$$

- $A = (a_i, ..., a_m)$ es el vector de estado que representa el grado de activación de los conceptos, es decir, el valor de cada concepto. En cada instante de tiempo t, cada concepto c_i tiene asignado un grado de activación a_i.
- F o f (·) es la función de umbral (activación) que mantiene los valores de cada concepto dentro de un intervalo. La elección de una función de activación depende del problema a resolver, entre las más usadas en la literatura están:
  - Bivalente
$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{2.17}$$
  - Trivalente
$$f(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} \tag{2.18}$$
  - Sigmoidal
$$f(x) = \frac{1}{1 + e^{-\lambda x}} \tag{2.19}$$
  - Tangente hiperbólica
$$f(x) = tanh(\lambda x) \tag{2.20}$$

La ecuación 2.21 muestra la regla de inferencia de Kosko para el concepto c_j en el momento t+1. Dado un vector inicial A, el proceso de inferencia consiste en aplicar la ecuación 2.21 para calcular el vector de estado A en cada paso t hasta que una condición es alcanzada.

$$A_i^{(t+1)} = f(\sum_{\substack{j=1 \\ i \neq j}}^{M} w_{ji} a_j^t) \tag{2.21}$$

Anteriormente, se ha comentado que la tarea de crear un MCD generalmente la realizan expertos en un dominio determinado, pero también podemos usar un conjunto de datos para entrenar al modelo. En ambos enfoques, el objetivo es extraer la matriz de peso final W. Para los enfoques basados en datos se han creado numerosas técnicas de aprendizaje, que se pueden dividir en tres grupos:

- Basadas en Hebbian: Los algoritmos basados en Hebbian son un aprendizaje no supervisado donde el objetivo es obtener la matriz W basado en el nivel de activación de los conceptos, tal que 2 conceptos que se activan en un momento tienden a fortalecer su relación y en caso contrario a inhibirla.
- Basada en poblaciones: Los algoritmos basados en la población son un aprendizaje supervisado. Consisten en calcular la matriz de peso a partir de datos. El objetivo es minimizar la diferencia entre los resultados esperados y las respuestas del MCD mediante la utilización de algoritmos de optimización para minimizar el error. Para minimizar el error, se modifican los pesos del MCD descritos como individuos de la población.
- Híbridas: Los algoritmos híbridos usan el conocimiento de los expertos para la construcción del MCD y se entrena de manera supervisada usando un conjunto de datos.

En este trabajo se usa la herramienta FCM-Experts, implementada en JAVA y definida por Napoles [42]. Incluye diferentes tipos de aprendizaje supervisado y no supervisado para los enfoques basados en datos. La construcción y entrenamiento de los modelos de clasificación del proyecto está basado en algoritmos de poblaciones, el proceso de optimización de cálculo de matriz W se realiza utilizando Particle Swarm Optimization (PSO) [43], en concreto, se usa la variante del algoritmo Global-best PSO [44].

2.4.1. Algoritmo de Optimización Global-best PSO

Global-best Particle Swarm Optimization [44] es una variante del algoritmo PSO definida en [44] y usada en problemas de optimización. PSO define un enjambre de partículas, cada una de las partículas representa una solución al problema de optimización (en este trabajo, cada partícula representa una matriz W). Cada partícula se define dentro del algoritmo por:

1. Una posición en el espacio de búsqueda X_i.
2. Una velocidad V_i.
3. Pbest: la mejor posición encontrada por la partícula, es decir, la mejor solución encontrada hasta el momento.

PSO también almacena la mejor posición encontrada por cualquier partícula dentro del enjambre (gbest). En el proceso de optimización, las partículas se mueven a través del espacio de búsqueda ajustando su posición (ver ecuación 2.22) y velocidad (ver ecuación 2.23)

$$X_i = X_i + V_i \tag{2.22}$$

$$V_i^{t+1} = wV_i^t + c_1 r_1(Pbest_i - X_i) + c2r2(gbest - X_i) \tag{2.23}$$

En la ecuación 2.23 encontramos varios términos que requieren de una explicación ya que actúan como hiperparámetros dentro del modelo:

- w: Es la inercia, permite que las partículas mantengan la dirección de movimiento de la anterior posición, este movimiento es el que ayuda a encontrar posibles nuevas soluciones.
- c_1 es el coeficiente cognitivo, dirige a la partícula a su mejor posición personal.
- c_2 es el coeficiente social, dirige a las partículas hacia la mejor posición global encontrada en el enjambre.

El algoritmo Global-best PSO continúa iterando hasta que se cumple algún criterio de terminación, como alcanzar un número máximo de iteraciones o lograr una precisión deseada en la solución.

2.5. Evaluación de modelos

El uso de métricas clásicas de evaluación en modelos de aprendizaje automático permiten:

1. Medir el rendimiento del modelo: proporciona una medida de su efectividad en función de criterios y objetivos establecidos.
2. Comparar modelos: ayuda a identificar que cuál de los modelos es el más efectivo en términos de rendimiento. dominio y contexto del problema a resolver.
3. Ajuste de hiper parámetros: Identifica las combinaciones de hiper parámetros que maximizan el rendimiento del modelo.
4. Detectar sobreajuste y subajuste: Identifican problemas de sobreajuste (overfitting) y subajuste (underfitting) en los modelos.
5. Evaluación del impacto de cambios: Miden y evalúan el impacto de cambios dentro de la estructura del modelo.

Las métricas clásicas de evaluación de modelos usadas en el trabajo realizado son:

- Matriz de Confusión
- Precisión
- Sensibilidad
- Puntuación F1
- Curva ROC

Clases	Clase Positiva (Real)	Clase Negativa (Real)
Clase Positiva (Predicha)	TP	FP
Clase Negativa (Predicha)	FN	TN

Tabla 2.1: Matriz de confusión de 2 clases

2.5.1. Matriz de confusión

La matriz de confusión evalúa el rendimiento de un modelo AA en problemas de clasificación. Muestra la relación entre las predicciones de las clases del modelo y las clases reales en forma de una matriz. La matriz de confusión tiene dimensiones $N \times N$, donde N es el número de clases en el problema de clasificación. Cada celda de la matriz representa una de las distintas combinaciones entre predicción de clase y clase real de cada una de las clases. La tabla 2.1 muestra la matriz de confusión de un modelo de clasificación de 2 clases, donde:

- TP (Verdaderos positivos): Número total de instancias de la clase positiva predecidas correctamente.
- FP (Falsos positivos): Número total de instancias en las que el modelo predijo incorrectamente una muestra como positiva cuando realmente era negativa.
- FN (Falsos negativos): Número total de instancias en las que el modelo predijo incorrectamente una muestra como negativa cuando realmente era positiva.
- TN (Verdaderos negativos): Número total de instancias de la clase negativa predecidas correctamente.

Algunos de los modelos construidos realizan tareas de predicción de clases con 3 clases. La tabla 2.2 muestra la matriz de confusión de un modelo de clasificación de 3 clases, donde:

Clases	Clase 1 (Real)	Clase 2 (Real)	Clase 3 (Real)	Falsos positivos
Clase 1 (Predicha)	TP1	E21	E31	FP1=E21+E31
Clase 2 (Predicha)	E12	TP2	E32	FP2=E12+E32
Clase 3 (Predicha)	E13	E23	TP3	FP3=E13+E23
Falsos negativos	FN1=E12+E13	FN2=E21+E23	FN3=E31+E32	

Tabla 2.2: Matriz de confusión de 3 clases

- TP1 (Verdaderos positivos clase 1): Número total de instancias de la clase 1 predichas correctamente.
- TP2 (Verdaderos positivos clase 2): Número total de instancias de la clase 2 predichas correctamente.
- TP3 (Verdaderos positivos clase 3): Número total de instancias de la clase 3 predichas correctamente.
- FN1 (Falsos negativos de la clase 1): Número total de instancias de la clase 1 en las que el modelo predijo incorrectamente como de la clase 2 o clase 3.
- FN2 (Falsos negativos de la clase 2): Número total de instancias de la clase 2 en las que el modelo predijo incorrectamente como de la clase 1 o clase 3.
- FN3 (Falsos negativos de la clase 3): Número total de instancias de la clase 3 en las que el modelo predijo incorrectamente como de la clase 1 o clase 2.
- FP1 (Falsos positivos de la clase 1): Número total de instancias en las que el modelo predijo incorrectamente una muestra como de clase 1 cuando realmente era de clase 2 o clase 3.

- FP2 (Falsos positivos de la clase 2): Número total de instancias en las que el modelo predijo incorrectamente una muestra como de clase 2 cuando realmente era de clase 1 o clase 3.
- FP3 (Falsos positivos de la clase 3): Número total de instancias en las que el modelo predijo incorrectamente una muestra como de clase 3 cuando realmente era de clase 1 o clase 2.

La matriz de confusión permite evaluar el rendimiento del modelo en cada clase y comprender los aciertos y errores de la clasificación. Permite calcular otras métricas de evaluación de modelos como la precisión, la sensibilidad y la puntuación F1.

2.5.2. Precisión

Mide la proporción de predicciones de una clase realizadas correctamente por el modelo sobre el total de predicciones realizadas para una clase en particular. La ecuación 2.24 muestra el cálculo de la precisión de la clase de positivos en un modelo de clasificación de 2 clases.

$$Precision = \frac{TP}{TP + FP} \tag{2.24}$$

La precisión proporciona una medida de qué tan precisas son las predicciones positivas del modelo en comparación con el número total de predicciones positivas, incluyendo tanto los aciertos (verdaderos positivos) como los errores (falsos positivos).

$$Precision = \frac{TP1}{TP1 + FP1} \tag{2.25}$$

La ecuación 2.25 muestra el cálculo de la precisión de la clase 1 en un modelo de clasificación de 3 clases para la clase 1.

2.5.3. Sensibilidad

La sensibilidad, recall o tasa de verdaderos positivos, mide la proporción de instancias positivas que son correctamente identificadas por el modelo, es decir, mide la capacidad del modelo para detectar correctamente las instancias de la clase en relación con el total instancias de la clase predichas correcta o incorrectamente. La ecuación 2.26 muestra el cálculo de la sensibilidad en un modelo de clasificación de 2 clases.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{2.26}$$

La ecuación 2.27 muestra el cálculo de la sensibilidad en un modelo de clasificación de 3 clases para la clase 1.

$$Sensibilidad = \frac{TP1}{TP1 + FN1} \tag{2.27}$$

2.5.4. Puntuación F1

La métrica Puntuación F1 combina la precisión y sensibilidad para proporcionar una medida general del rendimiento del modelo. La ecuación 2.28 muestra la fórmula del cálculo de la puntuación F1.

$$PuntuacionF1 = 2 * \frac{precision * sensibilidad}{precision + sensibilidad} \tag{2.28}$$

El valor de la puntuación F1 varía entre 0 y 1, donde 1 indica un rendimiento óptimo del modelo en términos de precisión y sensibilidad, y 0 indica un rendimiento deficiente. F1 equilibra la importancia de la precisión y sensibilidad del modelo. Es útil en situaciones y problemas donde se requiere tener un equilibrio entre la identificación precisa de los casos positivos y la capacidad de detectar todos los casos positivos.

2.5.5. Curva ROC

La curva ROC es una gráfica que muestra la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) a medida que se varía el umbral de clasificación del modelo.

En la curva ROC, el eje x representa la tasa de falsos positivos y el eje y representa la tasa de verdaderos positivos. Un clasificador perfecto que puede distinguir perfectamente entre las clases tendría una curva ROC que pasaría por el punto (0,1), lo que indicaría una sensibilidad del 100 %. La curva ROC también se utiliza para calcular el Área bajo la curva ROC (AUC-ROC), que es una medida numérica del rendimiento global del modelo. El AUC-ROC varía entre 0 y 1, donde un valor de 1 indica un rendimiento perfecto del modelo y un valor de 0.5 indica un rendimiento similar al azar.

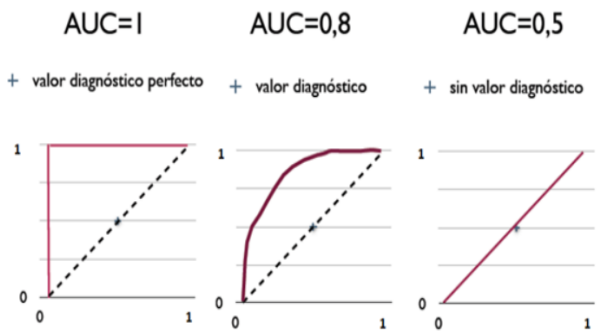

Figura 2.7: Tipos de curva ROC en base a su AUC

2.6. Explicabilidad de modelos

Actualmente, la Inteligencia Ariticial (IA) se usa en una amplia variedad de contextos. Sin embargo, existen áreas críticas en las que las decisiones tomadas y los resultados generados por los modelos tienen un impacto significativo en la vida de las personas. Algunos dominios críticos donde la IA tiene un impacto significativo son:

- Salud: En el campo médico, la IA es usada en el diagnóstico médico, la toma de decisiones clínicas y el desarrollo de tratamientos. Los modelos analizan grandes cantidades de datos clínicos, como imágenes médicas y registros de pacientes. Los médicos necesitan comprender cómo se llega a una determinada decisión por parte del modelo.

- Justicia: En el ámbito de la justicia, los modelos son usados en la evaluación de riesgos, la toma de decisiones legales y la predicción del comportamiento criminal. Es fundamental que los modelos sean explicables y transparentes para garantizar la equidad y la justicia en el sistema legal.

- Finanzas: Los modelos son aplicados en el análisis de riesgos, la evaluación de créditos y la detección de fraudes. Estas operaciones son cruciales para garantizar la eficiencia y la seguridad en el sector

financiero. La explicabilidad de los modelos permite a las instituciones financieras garantizar la equidad y confianza en los resultados generados.

- Transporte Autónomo: En el campo del transporte autónomo, la IA desempeña un papel fundamental en el desarrollo de vehículos autónomos. La explicabilidad de los modelos es esencial para comprender cómo toman decisiones críticas en tiempo real, como evitar colisiones o tomar desvíos seguros.

- Energía: En el ámbito de la energía, los modelos son aplicados en la gestión de redes eléctricas y la optimización del consumo energético para maximizar la eficiencia y sostenibilidad en el sector energético.

Se puede apreciar que la confianza en los modelos es un requisito fundamental en todos los ámbitos críticos mencionados anteriormente. De acuerdo a Doshi-Velez [45], las métricas tradicionales que se basan en cuantificar el rendimiento de los modelos no son adecuadas para describir la mayoría de las tareas del mundo real. Es necesario proporcionar una explicación de las decisiones tomadas por un modelo para que los expertos puedan confiar en él. Si un usuario no confía en el modelo, lógicamente no lo utilizará. Numerosos estudios [46][47] han afirmado que los usuarios prefieren un modelo explicativo en el que puedan confiar, incluso si su rendimiento es menor que otras alternativas en las que no se puede confiar. Por otro lado, el Reglamento General de Protección de Datos (GDPR) de la Unión Europea de 2018 establece el derecho a la explicación y a la no discriminación de los usuarios [48].

En los últimos años, ha surgido el campo de la Inteligencia Artifical Explicable (IAE), cuyo objetivo es desarrollar modelos capaces de explicar y justificar sus decisiones y resultados de manera comprensible para los seres humanos, con el fin de proporcionar confianza al usuario en el modelo. La IAE surge como un intento de abordar la falta de explicabilidad de los modelos complejos de ”caja negra” con alto rendimiento, que carecen de transparencia en su funcionamiento interno. La IAE es un campo emergente, es importante destacar que no existe una definición estricta de IAE, y varios documentos han intentado establecer reglas básicas y principios. [49] describió los cuatro principios fundamentales para los sistemas de IA explicables. En primer lugar, los sistemas de IA deben ser capaces de proporcionar explicaciones claras y comprensibles sobre cómo toman sus decisiones. Estas explicaciones deben ser significativas y adaptarse a las necesidades y nivel de experiencia del usuario. Además, es fundamental que las explicaciones sean precisas y reflejen el proceso real de toma de decisiones de la IA. Por último, los sistemas de IA deben ser diseñados para reconocer y comunicar sus propias limitaciones, así como las restricciones inherentes a los datos con los que han sido entrenados. Por otro lado, el trabajo [50] estableció seis propiedades básicas que los modelos de IA deben poseer para ser confiables y transparentes, las cuales son monotocidad, usabilidad, confiabilidad, causalidad, escalabilidad y generalidad. Finalmente, existen diversos enfoques para clasificar los métodos de explicabilidad. Estos enfoques se basan en diferentes criterios y perspectivas. A continuación se exponen algunas clasificaciones comunes de los métodos de explicabilidad. De acuerdo a [51], los métodos de IAE se pueden clasificar atendiendo a los siguientes criterios:

- Aplicación respecto a la construcción del modelo: los métodos de IAE se agrupan en tres categorías principales en función de cuándo se aplican en la construcción del modelo: Pre-Modelo, In-Modelo y Post-Modelo. Los métodos Pre-Modelo se aplican antes de la construcción del modelo y son independientes del modelo en sí. Estas técnicas se basan en realizar un análisis exploratorio de los datos para comprender mejor sus características y patrones. Los métodos In-Modelo hacen referencia a la propia explicabilidad inherente del modelo. Estas técnicas buscan comprender cómo el modelo toma decisiones y explicar sus resultados. Un ejemplo de técnica In-Modelo es MCD, que es una técnica de explicabilidad basada en la causalidad. Por último, los métodos Post-Modelo se

aplican después de construir el modelo. Se centran en analizar y explicar las decisiones tomadas por el modelo en función de las entradas y salidas.

- Explicabilidad en relación con el propio modelo: podemos distinguir entre explicabilidad intrínseca y post hoc. La explicabilidad intrínseca se refiere a la capacidad inherente del modelo para proporcionar explicaciones claras y comprensibles sobre sus decisiones. Esto significa que el modelo en sí mismo tiene la capacidad de explicar cómo llega a ciertas predicciones o resultados. Por otro lado, la explicabilidad post hoc se logra aplicando métodos o técnicas después de la construcción del modelo. Estos métodos son independientes del modelo construido y se utilizan para analizar y comprender cómo el modelo toma decisiones. La distinción entre explicabilidad intrínseca y post hoc radica en si la explicabilidad se logra debido a las características inherentes del modelo o mediante el uso de métodos independientes que analizan el modelo después de su construcción.

- Aplicabilidad a diferentes modelos: los métodos de IAE se pueden agrupar en específicos y agnósticos respecto a su capacidad de aplicación en distintos modelos.

  Los métodos específicos están diseñados con características particulares de ciertos tipos de modelos en mente. Estos métodos suelen aprovechar propiedades específicas de los modelos para proporcionar explicaciones relevantes y precisas. Los métodos específicos están diseñados y limitados para funcionar en técnicas específicas, ya que aprovechan características particulares de la técnica empleada en la construcción del modelo. Estos métodos no son directamente aplicables a otros tipos de técnicas. Por otro lado, los métodos agnósticos al modelo son independientes y aplicables a cualquier modelo, independientemente de su arquitectura o tipo. Estos métodos no asumen conocimiento previo sobre el modelo y lo tratan como una ”caja negra”. Los métodos específicos pueden ofrecer explicaciones más detalladas y específicas para técnicas particulares, aprovechando su estructura y características. Sin embargo, los métodos agnósticos al modelo son más flexibles y pueden proporcionar explicaciones generales para cualquier modelo.

Pre-Modelo	No aplica	No aplica
In-Modelo	Intrínseco	Específico del modelo
Post-Modelo	Post Hoc	Agnóstico al modelo

Tabla 2.3: Relación entre clases de métodos de explicabilidad

De acuerdo con Molnar [52], otro criterio que permite diferenciar los métodos de explicación es el resultado, es decir, el tipo de explicación que produce cada método. Se distinguen entre los siguientes distintos tipos de explicación:

- Basada en características: proporciona un resumen estadístico de cada característica y su influencia en las decisiones tomadas. Se analiza cómo cada variable de entrada afecta las salidas o predicciones del modelo.

- Basada en el comportamiento interno del modelo: la explicación se basa en examinar los componentes y elementos internos de los modelos intrínsecamente interpretables. Estos modelos tienen una estructura y reglas claras que permiten extraer la explicación, los métodos que generan los elementos internos del modelo son específicos del modelo en sí mismo

- Basadas en instancias o ejemplos: proporciona instancias específicas de la clase o ejemplos que ayudan a comprender cómo el modelo toma decisiones. Estos ejemplos se utilizan para ilustrar el razonamiento detrás de las predicciones del modelo.

- Basada en modelos sustitutos: construyen un modelo interpretable a partir de un modelo complejo de "caja negra". La idea es crear un modelo más simple y fácilmente interpretable que capture la esencia del modelo complejo original.

Por último, los métodos de explicabilidad se pueden clasificar de acuerdo al ámbito de aplicación en locales y globales.

1. Explicabilidad local: describen el comportamiento del modelo alrededor de una instancia o ejemplo específico proporcionado explicaciones detalladas y específicas para esa observación en particular. Al examinar el modelo en un contexto local, se puede comprender el razonamiento para esa instancia en particular.
2. Explicabilidad global: describen el comportamiento general del modelo, proporcionan una visión más amplia y generalizada del modelo en su conjunto. Son útiles para comprender las características y patrones generales del modelo.

La tabla 2.4 muestra los métodos Post-Modelo y agnósticos al modelo más empleados en la literatura.

Método de explicación	**Ámbito**	**Resultado**
Gráfico de dependencia parcial	Global	Resumen de características
Expectativa de Condición Individual	Global/Local	Resumen de características
Gráfico de efectos locales acumulados	Glocal	Resumen de características
LIME	Local	Modelo sustituto y resumen de características
Interacción de características	Global	Resumen de características
Importancia de características	Global/Local	Resumen de características
Modelo sustituto local	Local	Modelo sustituto interpretable
Valores de Shapley	Local	Resumen de características
Desglose	Local	Resumen de características
Anchura	Local	Resumen de características
Explicación de contrafactuales	Local	Instancias nuevas
Funciones de influencia	Global/Local	Instancias existentes
Prototipos	Global	Instancias existentes

Tabla 2.4: Comparación de los principales métodos Post-Modelo y agnósticos al modelo

En el presente trabajo, se define un enfoque de explicabilidad propio basada en causalidad usando la técnica de MCD, y se emplean dos métodos post Hoc, un método local basado en el método LIME y otro global basado en la Importancia de características, ambos agnósticos al modelo

2.6.1. LIME (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLICATIONS)

LIME es un método para explicar las predicciones individuales de un modelo mediante la construcción de modelos sustitutos locales basados en cada predicción individual. Los modelos sustitutos se construyen al generar un conjunto de instancias cercanas a la instancia de interés y luego se etiquetan utilizando el modelo de aprendizaje automático original. Estas instancias y etiquetas se utilizan para entrenar el modelo sustituto, que puede ser más interpretable. LIME fue propuesto por Ribeiro [53], y se basa en que un método de explicabilidad debe satisfacer una serie de principios:

- Las explicaciones deben ser fáciles de entender y tener en cuenta las limitaciones del usuario, por lo que deben utilizar artefactos textuales o visuales que faciliten la comprensión de la explicación. En modelos complejos con una gran cantidad de características, estos criterios pueden no ser suficientes para proporcionar una explicación. Un método de explicación debe ser capaz de seleccionar un pequeño conjunto de características que son más importantes para una predicción dada.
- Fidelidad local: capacidad de un método de explicabilidad para proporcionar una explicación del comportamiento de una instancia de manera local, es decir, en su vecindario cercano, Implica comprender cómo se toman las decisiones o se realizan las predicciones para esa instancia en particular, teniendo en cuenta las características y los valores de las instancias cercanas.
- Agnóstico al modelo: un método de explicación debe ser capaz de tratar cualquier modelo como una "caja negra" proporcionando una explicación independiente del modelo.
- Perspectiva global: un modelo de explicabilidad debe ser capaz de proporcionar una perspectiva global sobre el comportamiento del modelo. A partir de un conjunto de explicaciones locales de las predicciones, se puede proporcionar una explicación representativa del modelo.

El objetivo de LIME es explicar la predicción de una instancia calculando la influencia de las características individuales, dada una instancia. LIME descubre qué sucede con la predicción de una instancia cuando ocurren variaciones dentro de las variables de la instancia, estas variaciones se conocen como la vecindad de la instancia. Para cada instancia a explicar, LIME crea un modelo sustituto. La ecuación 2.29 muestra cómo se obtiene la explicación de LIME:

$$\xi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2.29}$$

Donde G es la familia de explicaciones posibles, L es la función de pérdida y mide qué tan cerca está la explicación de g en el modelo sustituto de la predicción del modelo original y $\Omega(g)$ mide la complejidad del modelo sustituto.

2.6.2. Método de Importancia de características basada en permutaciones

La importancia de las características es otro método que permite calcular la importancia relativa de cada característica (variable) en un conjunto de datos, en la construcción de un modelo predictivo. Existen diversos métodos de explicabilidad basados en la importancia de las características. En este trabajo se uso un método basado en la importancia de características usando un criterio de permutación para la sustitución de valores aleatorios para una variable. Este método consiste en modificar aleatoriamente los valores de una característica dentro del rango de valores posibles de la variable y luego observar cómo afecta al rendimiento del modelo. Cuanto mayor sea el impacto en el rendimiento del modelo, se considerará que esa característica es más importante, lo que permite identificar las características más influyentes en las predicciones del modelo. El impacto en el rendimiento es medido usando métricas clásicas, en este trabajo, la métrica usada para medir el impacto en el rendimiento es la precisión (ACC), el algoritmo es mostrado en 2.1.

Data: Conjunto de Datos
Result: Importancia de características
Entrenamiento del modelo;
Test del modelo;
Cálculo de la métrica de rendimiento (precisión, ACC en nuestro caso) ;
for *Variable j en el conjunto de datos* **do**
    Asignar aleatoriamente valores a la variable j dentro del rango;
    Entrenamiento del modelo;
    Test del modelo;
    Cálculo del ACC;
    Cálculo de la diferencia $Df_j = ACC_0 - ACC_j$;
    Hacer el histograma (id de la variable en el eje x y Df_j en el eje y;
    Asignar los valores originales a la variable j

Algorithm 2.1: Macro-algoritmo de importancia de características usando el criterio de permutación

2.6.3. Método basado en causalidad

Un MCD establece las relaciones de influencia entre los conceptos de entrada y los conceptos clase mediante los pesos de las relaciones causales. Estos pesos representan la fuerza o intensidad con la que un concepto influye en otro, y permiten establecer qué características son más influyentes en la toma de decisiones del modelo. Gran parte de los trabajos sobre MCD se enfoca principalmente en la construcción de modelos y en la observación exclusiva de las relaciones causales entre los conceptos de dichos modelos. Existen muy pocos trabajos que realicen un análisis profundo de la explicabilidad en los MCD. Estos trabajos se centran en el uso de la sensibilidad local con la intención de comprender como variaciones en las variables de entradas afectan a la capacidad predictiva del modelo.

En este trabajo, se presenta un método distinto y novedoso. Nuestro método, denominado ”Algoritmo de evaluación de la influencia de características menos Influyentes basada en causalidad”, se fundamenta en el principio de sensibilidad local, que guarda similitudes con enfoques previos, pero con una diferencia importante. En este análisis, se evalúa la sensibilidad de los conceptos de clase eliminando las relaciones causales de menor influencia entre los conceptos de entrada y los conceptos de clase (salida). El objetivo de este método es analizar la influencia causal que ejercen los conceptos de menor influencia en relación con los conceptos de clase (salida) a los que están relacionados. Se busca determinar si esta influencia

tiene un impacto significativo en el comportamiento predictivo del modelo. La eliminación de las relaciones causales de menor influencia se lleva a cabo al establecer umbrales para los pesos de las relaciones causales. Dado un c_i, que es un concepto de entrada, y c_j, que es un concepto de salida del modelo, existe una relación causal con un peso w_{ij}. Si w_{ij} es menor que los umbrales establecidos, la relación causal w_{ij} se elimina del modelo. El análisis de sensibilidad es realizado a través de un conjunto instancias de prueba. Estas instancias son escogidas deliberadamente para abarcar una variedad de escenarios. Si se observan diferencias significativas en las predicciones entre el modelo original y el modelo modificado, se puede concluir que la eliminación de las relaciones causales menos influyentes tiene un impacto en el comportamiento predictivo del modelo. En el algoritmo 2.2 se muestran los pasos del algoritmo del método definido.

Data: Conjunto de datos
Result: Importancia de las características menos influyentes del modelo
Entrenamiento del modelo;
Test del modelo;
Establecer lista de umbrales U;
for *concepto de clase* c_c *en el modelo* **do**
  for *umbral* $u_i \in U$ **do**
    for *relación* W_{ec} *entre el concepto de entrada* c_e *y el concepto de salida* c_c **do**
      if $|w_{ec}| \leq u_i$ **then**
        Eliminación de la relación causal $we3_{ec}$ *;*
    Cálculo de la diferencia en las predicciones entre el modelo original y el modelo modificado por el umbral u_i
  Volver al modelo original

Algorithm 2.2: Algoritmo de evaluación de la influencia de características menos influyentes basada en causalidad

Los umbrales han sido diseñados teniendo en cuenta la distribución de los pesos de las relaciones causales para cada clase específica, lo que los hace específicos y dependientes del modelo analizado. Este análisis permite evaluar tanto la influencia de los conceptos con menor peso como la de aquellos con mayor peso, observando si el comportamiento del modelo cambia o permanece igual.

Capítulo 3

Desarrollo Práctico

3.1. Introducción

En esta sección, se lleva a cabo la experimentación basada en los conceptos teóricos previamente explicados en la sección anterior. En primer lugar, se realiza la selección de los conjuntos de datos, se realiza una explicación de los metadatos de cada uno de los conjuntos (tamaño muestra, número de variables, tipos de variables, origen) y su relevancia en el contexto del proyecto. Posteriormente, procedemos a la preparación de los datos o ingeniería de características. Este paso es fundamental para garantizar que los datos estén limpios, estructurados y listos para ser utilizados en la construcción de los modelos. Realizamos transformaciones y procesamientos que nos permiten extraer la información más relevante y significativa de los datos originales. Luego, se realiza a la construcción de los modelos. Durante esta etapa, se ajustan los hiperparámetros de cada uno de los modelos. Una vez construidos los modelos, se procede a evaluar su rendimiento, esta evaluación permite comprender como de efectivos son los modelos en la tarea que se les asignó. Finalmente, aplicamos los métodos de explicabilidad a los modelos construidos. Estos métodos nos permiten comprender cómo se toman las decisiones dentro de los modelos y qué características son las más influyentes en sus predicciones. A través de la explicabilidad, buscamos obtener una visión más clara y comprensible de la relación entre las variables y cómo afectan a los resultados del modelo.

3.2. Entendimiento de los datos

El propósito de esta fase es lograr una comprensión exhaustiva de los datos con los cuales se va a trabajar. Se recopilan datos provenientes de diversas fuentes con el fin de identificar conjuntos de datos pertinentes, y posteriormente, se realiza una exploración inicial para analizar su contenido.

3.2.1. Conjunto de datos de Energía

El conjunto de datos de electricidad se obtuvo de Kaggle [54], se caracteriza por mostrar la evolución del consumo, generación de energía y precios a lo largo del tiempo, en intervalos de una hora, el conjunto de datos está compuesto de datos procedentes de tres fuentes distintas. Los datos de consumo y generación se recuperaron de ENTSO, un portal público para datos de operadores de servicios de transmisión (TSO). Los datos referentes al precio del consumo eléctrico se han obtenido del TSO español Red Eléctrica España. Finalmente, los datos meteorológicos se adquirieron de un proyecto abierto de Open Weather

API. El conjunto de datos inicial de Energía presenta un total de 43 características y 35145 instancias, recogidas cada hora, desde el 01-1-2015 a las 00:00 hasta el 31-12-2018 a las 23:00. En un primer análisis del conjunto de datos de energía se han comprobado las siguientes observaciones:

- Varias de las características presentan únicamente valores nulos, lo cual sugiere una posible pérdida de información por parte de los organismos responsables de la distribución de los datos. Se ha optado por eliminar estas características ya que no aportan nada en la resolución del problema, en concreto, se han eliminado 15 características en total.

- Se identificaron tres variables relacionadas con la temperatura en el conjunto de datos. Sin embargo, se procedió a eliminar dos de ellas para simplificar y evitar redundancias en la construcción de los modelos.

- Se encontraron cuatro variables de tipo categórico que realizaban una descripción del tiempo en el momento del muestro, La información proporcionada dichas variables era redundantes respecto a las otras variables climáticas del conjunto de datos, se optó por la eliminación de dichas características.

- Varias variables presentan valores nulos y atípicos, se realizar su correspondiente tratamiento en los apartados dedicados a ello.

Al finalizar este análisis, el conjunto de datos se ha visto reducido a un total de 22 características y 35145 instancias, todas las características son de tipo numérico. La tabla 3.1 muestra una descripción de las variables que conforman el conjunto de datos de energía antes de realizar el preprocesamiento.

Nombre de la característica	**Descripción**
Temperatura	Temperatura en grados Kelvin
Presión	Presión en hectopascales
Humedad	Medición de la humedad en porcentaje
Velocidad del viento	Velocidad del aire en m/s
Dirección del viento	Dirección del viento en grados
Generación de biomasa	Generación de biomasa en MW
Generación de carbón marrón/lignito	Generación de carbón marrón/lignito en MW
Generación de gas	Generación de gas en MW
Generación de carbón mineral	Generación de carbón en MW
Generación de petróleo	Generación de petróleo en MW
Generación de hidroelectricidad de bombeo	Generación de energía hidroeléctrica de bombeo en MW
Generación hidroelectricidad de flujo continuo	Generación de energía hidroeléctrica de flujo continuo en MW
Generación hidroelectricidad de embalses	Generación de energía hidroeléctrica en embalses en MW
Generación nuclear	Generación de energía nuclear en MW
Generación de otros tipos	Generación de otros tipos de energía en MW
Generación de otras energías renovables	Generación de otras energías renovables en MW
Generación de energía solar	Generación solar en MW
Generación de residuos	Generación de residuos en MW
Generación eólica terrestre	Generación eólica terrestre en MW
Carga total actual	Demanda eléctrica real
Precio del mercado diario	Precio previsto en el mercado en EUR/MWh
Precio actual	Precio en EUR/MWh

Tabla 3.1: Breve resumen de las características del conjunto de datos de energía en el análisis inicial

Las figuras 3.1, 3.2, 3.3 y 3.4 contienen información detallada sobre los estadísticos fundamentales asociados a cada una de las variables presentes en el conjunto de datos de energía. El número de registros sin valores nulos indica cuántas observaciones en la variable contienen datos completos y no tienen valores faltantes. La media es el promedio aritmético de todos los valores en la variable. La desviación estándar cuantifica cuánto varían los valores con respecto a la media. El valor mínimo es el valor más bajo presente en la variable, mientras que el valor máximo es el valor más alto. Estos valores extremos pueden proporcionar información sobre posibles valores atípicos o excepcionales en los datos. Los cuartiles dividen los datos ordenados en cuatro partes iguales. El primer cuartil (Q1) es el valor que separa el 25 % inferior de los datos, la mediana (Q2) divide los datos en dos partes iguales (50 % cada una), y el tercer cuartil (Q3) separa el 25 % superior de los datos.

	temp	pressure	humidity	wind_speed	wind_deg	generation biomass
count	35145.000000	35145.000000	35145.000000	35145.000000	35145.000000	35126.000000
mean	290.780780	1015.973794	65.145113	2.692815	160.753820	383.494762
std	7.231284	11.927677	19.689276	2.581825	120.436402	85.359940
min	268.830656	969.000000	8.000000	0.000000	0.000000	0.000000
25%	285.150000	1012.000000	51.000000	1.000000	50.000000	333.000000
50%	290.170000	1017.000000	67.000000	2.000000	130.000000	367.000000
75%	296.150000	1021.000000	82.000000	4.000000	280.000000	433.000000
max	311.150000	1087.000000	100.000000	133.000000	360.000000	592.000000

Figura 3.1: Estadísticos conjunto de datos Energía (primer grupo de variables)

	generation fossil brown coal/lignite	generation fossil gas	generation fossil hard coal	generation fossil oil	generation hydro pumped storage consumption
count	35127.000000	35127.000000	35127.000000	35126.000000	35126.000000
mean	448.321491	5623.118456	4257.369488	298.282156	475.168565
std	354.600474	2202.111947	1961.783054	52.525203	791.842394
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	4126.000000	2528.500000	263.000000	0.000000
50%	510.000000	4969.000000	4477.000000	300.000000	68.000000
75%	757.000000	6429.000000	5840.000000	330.000000	615.750000
max	999.000000	20034.000000	8359.000000	449.000000	4523.000000

Figura 3.2: Estadísticos conjunto de datos Energía (segundo grupo de variables)

	generation hydro run-of-river and poundage	generation hydro water reservoir	generation nuclear	generation other	generation other renewable	generation solar
count	35126.000000	35127.000000	35128.000000	35127.000000	35127.000000	35127.000000
mean	972.006719	2604.731289	6263.290765	60.224044	85.637971	1431.093973
std	400.642209	1834.671896	839.879946	20.239146	14.075899	1679.379427
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	637.000000	1077.000000	5757.000000	53.000000	73.000000	70.000000
50%	906.000000	2164.000000	6564.000000	57.000000	88.000000	615.000000
75%	1250.000000	3757.000000	7024.000000	80.000000	97.000000	2573.000000
max	2000.000000	9728.000000	7117.000000	106.000000	119.000000	5792.000000

Figura 3.3: Estadísticos conjunto de datos Energía (tercer grupo de variables)

	generation fossil brown coal/lignite	generation fossil gas	generation fossil hard coal	generation fossil oil	generation hydro pumped storage consumption
count	35127.000000	35127.000000	35127.000000	35126.000000	35126.000000
mean	448.321491	5623.118456	4257.369488	298.282156	475.168565
std	354.600474	2202.111947	1961.783054	52.525203	791.842394
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	4126.000000	2528.500000	263.000000	0.000000
50%	510.000000	4969.000000	4477.000000	300.000000	68.000000
75%	757.000000	6429.000000	5840.000000	330.000000	615.750000
max	999.000000	20034.000000	8359.000000	449.000000	4523.000000

Figura 3.4: Estadísticos conjunto de datos Energía (cuarto grupo de variables)

3.2.2. Conjunto de datos de COVID-19

El dataset de COVID-19 es de [55] Israel y comprende un total de 5,853,480 pruebas realizadas por el gobierno entre marzo y noviembre de 2020. Para agilizar los análisis, se tomó una muestra aleatoria de 100,000 instancias del conjunto completo. Originalmente, el dataset contenía 10 características. Sin embargo, la variable 'fecha de la prueba' se eliminó porque no era relevante para el objetivo del estudio. Las características seleccionadas se centran en síntomas, indicadores de género, edad y razones para realizar la prueba, así como la característica objetivo, que clasifica las pruebas como positivas o negativas. Las características utilizadas son binarias, donde 1 indica presencia y 0 indica ausencia, excepto para la edad donde 0 indica menor que 60 años y 1 lo contrario, y el indicador de la razón de prueba, donde 0 denota otra razón para realizar la prueba, 1 indica contacto con una persona confirmada y 2 indica haber llegado del exterior. Estas variables forman la base para la construcción del modelo de detección y clasificación de COVID-19.

El conjunto de datos de COVID-19 está listo y preparado para la construcción de modelos, lo que significa que ya ha pasado por un proceso de preprocesamiento previo. La tabla 3.2 muestra la descripción de las características del dataset de COVID-19

Nombre de la característica	**Descripción**
Tos	Expulsión repentina y aguda de aire de los pulmones actuando como un mecanismo protector para limpiar las vías respiratorias o como síntoma de alteración pulmonar
Fiebre	Aumento de la temperatura corporal
Dificultad al respirar	Sensación incómoda en la que la persona tiene dificultad para respirar normalmente pulmonar
Dolor de garganta	Dolor, picazón o irritación de la garganta que a menudo empeora cuando traga
Dolor de cabeza	Dolor y malestar localizado en cualquier parte de la cabeza
Edad de 60 años y más	Indica si un sujeto tiene más de 60 años de edad
Género	Género del sujeto.
Motivo del test	Razón de realización del test
Resultado del test	Resultado del test (Variable objetivo a predecir)

Tabla 3.2: Breve Descripción de las características del conjunto de datos de COVID-19

3.2.3. Conjunto de datos de Dengue

El conjunto de datos de Dengue corresponde a los datos de pacientes con Dengue recopilados por el Ministerio de Salud de Medellín [56] en 2020. El preprocesamiento del conjunto de datos fue realizado por Hoyos y colaboradores en el artículo [15]. Se seleccionaron 22 variables del conjunto de datos original, consideradas por expertos como las más relevantes en el contexto del Dengue. Las características seleccionadas incluyen síntomas, pruebas de laboratorio y la característica objetivo para clasificar la gravedad del Dengue. Todas las características son binarias, excepto la variable objetivo, que distingue entre tres tipos de Dengue: Dengue sin signos de alarma (10,210 instancias), Dengue con signos de alarma (11,123 instancias) y Dengue grave (11,186 instancias). La tabla 3.3 muestra una descripción de las características finales presentes en el conjunto de datos, recogida en al artículo citado con anterioridad [15].

Nombre de la característica	Descripción
Edad	Tiempo transcurrido en años desde el nacimiento del individuo
Fiebre	Incremento de la temperatura corporal
Cefalea	Dolor y disconformidad localizado en cualquier parte de la cabeza
Dolor DO	Dolor detrás de los ojos
Mialgias	Dolor muscular
Artralgias	Dolor articular
Erupción	Exantema cutáneo
Dolor abdominal	Dolor intenso, ubicado en el epigastrio y/o el hipocondrio derecho
Vómito	Expulsión violenta del contenido estomacal por la boca
Somnolencia	Estado de cansancio y sueño profundo y prolongado
Hipotensión	Presión arterial baja en la pared de la arteria
Hepatomegalia	Condición de tener el hígado agrandado
Hemorragia mucosa	Manifestaciones de sangrado leve a severo en la mucosa nasal, encías, piel, tracto genital femenino, cerebro, pulmones, tracto digestivo y hematuria
Hipotermia	Decremento de la temperatura corporal.
Hematocrito alto	Prueba del aumento indirecto del hematocrito
Plaquetas bajas	Decremento de los niveles de plaquetas en sangre
Edema	Edema causado por el exceso de líquido atrapado en los tejidos del cuerpo
Extravasación	Se caracteriza por derrames serosos a nivel de varias cavidades
Sangrado	La sangre se filtra desde las arterias, venas o capilares por los cuales circula, especialmente cuando se produce en cantidades muy grandes
Shock	Manifestación de gravedad evidenciada por piel fría, pulso débil, taquicardia e hipotensión
Falla orgánica	Afectación de varios órganos debido a la extravasación de líquidos
Severidad	Severidad del Dengue (Variable objetivo a predecir)

Tabla 3.3: Breve Descripción de las características del conjunto de datos de Dengue

3.3. Preparación de los datos

La preparación es un proceso esencial con el objetivo de mejorar la calidad de los datos y la efectividad de los modelos de AA. El resultado final son conjuntos de datos optimizados, los cuales son usados en la construcción de los modelos. En esta sección se aplican todas las técnicas definidas en la correspondiente sección de estudio teórico.

3.3.1. Conjunto de datos de Energía

3.3.1.1. Detección de valores atípicos y faltantes

A continuación, se verifica si alguna de las variables contiene valores faltantes o nulos, la tabla 3.4 muestra el número de valores nulos de cada columna.

Nombre de la característica	**N. valores nulos**
Generación de biomasa	19
Generación de carbón marrón/lignito	18
Generación de gas	18
Generación de carbón mineral	18
Generación de petróleo	19
Generación de hidroelectricidad de bombeo	19
Generación de hidroelectricidad de flujo continuo	19
Generación de hidroelectricidad de embalses	18
Generación de nuclear	17
Generación de otros tipos	18
Generación de otras renovables	18
Generación de solar	18
Generación de residuos	19
Generación de eólica terrestre	18
Carga total actual	36

Tabla 3.4: Número de valores nulos por cada columna del conjunto de datos de energía

Se puede observar que casi todas las variables que presentan valores tiene casi el mismo número de instancias con valores nulos, a excepción de la característica 'carga total actual'. Se procede a comprobar si se tratan de las mismas instancias. En la figura 3.5 se puede visualizar que efectivamente los valores nulos se encuentran dentro de las mismas instancias, se ha tomado la decisión de eliminar esas instancias debido a su escaso número. Además, se ha observado que la variable çarga total actual"también contiene muy pocos valores nulos, por lo que se han eliminado las instancias que los contenían. Como resultado de este proceso, el conjunto de datos de Energía ahora consta de 35099 registros en total.

generation biomass	generation fossil brown coal/lignite	generation fossil gas	generation fossil hard coal	generation fossil oil	generation hydro pumped storage consumption	generation hydro run-of-river and poundage	generation hydro water reservoir	generation nuclear	generation other	generation other renewable	generation solar	generation waste
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figura 3.5: Visualización de valores nulos

La detección de valores atípicos se realiza utilizando el rango intercuartílico, tal como se ha explicado en la sección teórica. La tabla 3.5 muestra el número de valores atípicos en cada una de las características.

Nombre de la característica	N. valores atípicos
Presión	3394
Velocidad del viento	1005
Generación de biomasa	85
Generación gas	2192
Generación de hidroelectricidad de embalses	18
Generación de petróleo	243
Generación de hidroelectricidad de bombeo	3763
Generación de hidroelectricidad de embalses	344
Generación nuclear	69
Generación de otros tipos	1272
Generación de otras renovables	3
Generación de residuos	326
Generación eólica terrestre	380
Precio del mercado diario	842
Precio actual	701

Tabla 3.5: Número de valores atípicos por cada columna del conjunto de datos de energía

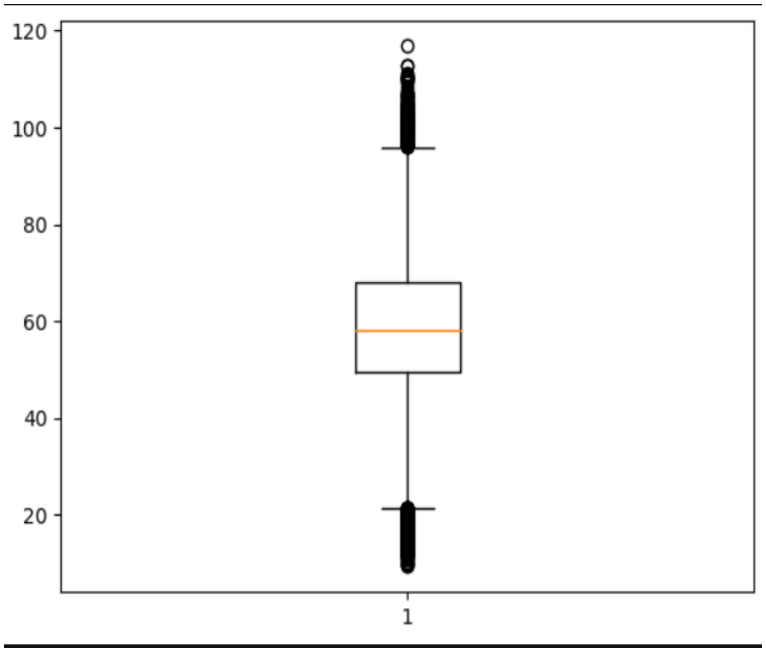

Figura 3.6: Diagrama de caja y bigotes de la variable Precio actual

La figura 3.6 muestra el diagrama de caja y bigotes de la variable 'Precio actual', donde se puede claramente observar cómo algunos valores se encuentran fuera del rango de los 'bigotes'. Cualquier valor que esté más allá de estos límites es considerado un valor atípico.

Se ha procedido a tratar los valores atípicos mediante la sustitución de cada variable que los contenía con su respectiva mediana. Esta elección se basa en el hecho de que la mediana es una medida resistente

y menos influenciada por valores extremos, lo que contribuye a reducir el efecto negativo de los valores atípicos en el análisis o modelado de los datos.

Una vez sustituidos los valores atípicos procedemos a visualizar cambios es los estadísticos básicos de cada variable respecto a los estadísticos mostrados en las figuras 3.1, 3.2, 3.3 y 3.4

	temp	pressure	humidity	wind_speed	wind_deg	generation biomass	generation fossil brown coal/lignite	generation fossil gas
count	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000
mean	290.787414	1017.307160	65.147867	2.423972	160.688624	383.502208	448.378301	5212.805749
std	7.230523	6.173008	19.688745	1.874730	120.444575	84.679051	354.640094	1532.870276
min	268.830656	999.000000	8.000000	0.000000	0.000000	183.000000	0.000000	1518.000000
25%	285.150000	1014.000000	51.000000	1.000000	50.000000	333.000000	0.000000	4126.000000
50%	290.190000	1017.000000	67.000000	2.000000	130.000000	367.000000	510.000000	4968.000000
75%	296.150000	1021.000000	82.000000	4.000000	280.000000	431.000000	757.000000	5923.000000
max	311.150000	1034.000000	100.000000	8.000000	360.000000	583.000000	999.000000	9882.000000

Figura 3.7: Estadísticos conjunto de datos Energía

	generation fossil hard coal	generation fossil oil	generation hydro pumped storage consumption	generation hydro run-of-river and poundage	generation hydro water reservoir	generation nuclear	generation other
count	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000
mean	4256.924442	298.203624	222.317018	971.972392	2544.338614	6269.108607	61.935810
std	1961.603268	51.065366	356.666513	400.616106	1745.906026	830.232554	17.759825
min	0.000000	163.000000	0.000000	0.000000	0.000000	3858.000000	13.000000
25%	2528.000000	263.000000	0.000000	637.000000	1076.500000	5782.500000	54.000000
50%	4475.000000	300.000000	68.000000	905.000000	2163.000000	6562.000000	57.000000
75%	5840.000000	329.000000	289.000000	1250.000000	3677.000000	7024.000000	80.000000
max	8359.000000	430.000000	1540.000000	2000.000000	7767.000000	7117.000000	106.000000

Figura 3.8: Estadísticos conjunto de datos Energía

	generation other renewable	generation solar	generation waste	generation wind onshore	total load actual	price day ahead	price actual
count	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000
mean	85.656685	1431.480099	270.976039	5355.890766	28695.778626	50.650028	57.887698
std	14.042807	1679.582186	48.021950	3057.519925	4573.548850	13.029869	14.195902
min	43.000000	0.000000	135.000000	0.000000	18041.000000	12.940000	9.330000
25%	74.000000	70.000000	243.000000	2935.000000	24808.000000	42.370000	49.355000
50%	88.000000	615.000000	279.000000	4847.000000	28901.000000	50.520000	58.010000
75%	97.000000	2574.000000	310.000000	7283.000000	32186.500000	60.450000	68.010000
max	119.000000	5792.000000	357.000000	14090.000000	41015.000000	89.010000	116.800000

Figura 3.9: Estadísticos conjunto de datos Energía

Se hace evidente que, al sustituir los valores atípicos en variables como 'Presión' y 'Precio actual' utilizando la mediana, los estadísticos básicos de estas características han experimentado cambios significativos, los nuevos valores tienden a estar más cerca del centro de la distribución. Como resultado, los valores mínimos y máximos de estas variables se han ajustado hacia valores más cercanos a la media-

na. Además, los cuartiles, como el primer y tercer cuartil, también se han visto influenciados por esta sustitución de valores atípicos, debido a que se ha modificado la distribución.

3.3.1.2. Discretización de Datos

La variable 'Precio actual' es una variable numérica sobre la que se realizará la predicción. Para abordar el objetivo de pronosticar el precio de la energía, se ha optado por organizar los datos de Precio actual en niveles o categorías. Para lograr esto, fue necesario discretizar la variable y dividirla en diferentes rangos, convirtiéndola así en etiquetas o clases. Esta discretización nos permite trabajar con un enfoque de clasificación, donde cada etiqueta representa un nivel de precio específico. De esta manera, podemos predecir qué nivel de precio se espera para distintas combinaciones de consumo y generación eléctrica.

Durante la construcción de los modelos, se realizaron pruebas con diversos números de niveles de etiquetas para representar los rangos de precios. Después de un análisis y evaluación, se llegó a la conclusión de que utilizar únicamente 3 niveles de precio resultaba ser la opción más adecuada y efectiva. La tabla 3.6 muestra el número de instancias contenidas en cada uno de los niveles de precio creados en el proceso de discretización.

Clases	Número de registros por Clase
Nivel de precio bajo	5775
Nivel de precio medio	28009
Nivel de precio alto	1315

Tabla 3.6: Número de instancias por clase en la discretización

En el proceso de discretización, se eligió implementar el enfoque uniforme, lo que implica la creación de intervalos de valores equitativos para categorizar las instancias en diferentes clases. Se observa en la tabla 3.6 que la clase de 'nivel de precio medio' agrupa un número de instancias mayor que las otras dos clases, por lo que es necesario realizar un proceso de balanceo que se comentara después.

3.3.1.3. SMOTE

En la Tabla 3.6, se muestran el número de instancias en cada nivel de precio obtenido durante el proceso de discretización. Se observa que varias instancias pertenecen a la clase "precio medio", mientras que las clases "precio altoƶ "precio bajo"tienen menos instancias en comparación. Esta situación puede ser problemática al realizar el entrenamiento de un modelo, ya que el modelo tiende a ajustarse a la clase dominante, como se mencionó en el apartado teórico.

En esta situación, tenemos dos opciones para abordar el desbalanceo de clases. Una opción es seleccionar un subconjunto de solo 1000 o 1500 instancias del conjunto de datos para el entrenamiento. Esto puede ayudar a equilibrar las clases, pero también puede llevar a la pérdida de información al descartar un gran número de instancias. La otra opción es utilizar una técnica de balanceo de clases, como SMOTE, la cuál, genera instancias sintéticas para las clases minoritarias, lo que aumenta la cantidad de datos en esas clases y equilibra el conjunto de datos. La siguiente tabla muestra el número de instancias por clase después del balanceo con SMOTE.

Clases	Número de registros por Clase
Nivel de precio bajo	28009
Nivel de precio medio	28009
Nivel de precio alto	28009

Tabla 3.7: Número de instancias por clase en el balanceo

3.3.1.4. Correlación de Pearson

Se obtienen todos los coeficientes de correlación de Pearson entre todas las combinaciones de variables para analizar las relaciones lineales existentes. Estos coeficientes se representan en una matriz, tal como se ilustra en la Figura 3.10. En la tabla 3.8, se presentan aquellas correlaciones que superan el umbral establecido de 0.2 con respecto a la variable objetivo.

Variable	**Nivel de coste de la energía**
Generación de carbón marrón/lignito	0.30
Generación de gas	0.32
Generación de carbón mineral	0.37
Generación de petróleo	0.23
Generación de otras renovables	0.23
Carga total actual	0.31
Precio del mercado diario	0.50
Precio actual	0.78

Tabla 3.8: Correlaciones de Pearson por encima del umbral de 0.2 respecto a la variable objetivo

Las correlaciones restantes con respecto a la variable objetivo son bajas. Sin embargo, no serán descartadas definitivamente hasta el estudio Spearman y ANOVA. Algunas de las variables podrían estar relacionadas de manera no lineal o tener efectos significativos en combinación con otras variables. Por ello, el uso de correlaciones de Spearman, que no se limitan a relaciones lineales, y el análisis de varianza nos permitirán permiten tener una visión más completa del conjunto de datos.

Finalmente en la matriz de correlación existen algunas combinaciones de variables con correlaciones altas entre sí, estas correlaciones no son lo suficientemente fuertes como para descartar alguna de las variables del conjunto de datos. Esto indica que las variables están interrelacionadas en cierta medida, pero ninguna de ellas tiene una influencia tan dominante sobre las demás como para justificar su eliminación del análisis.

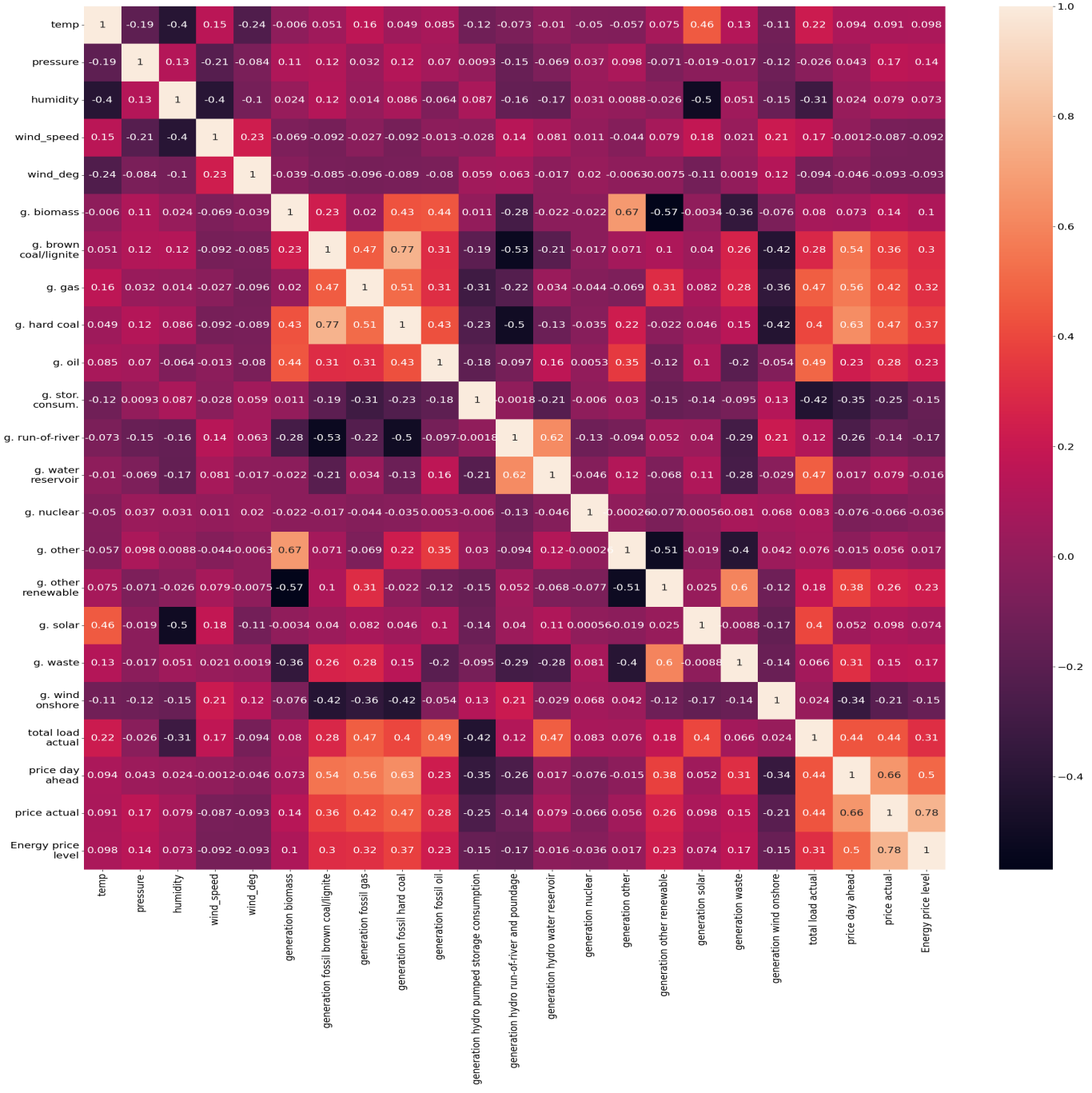

Figura 3.10: Correlación de Pearson conjunto de datos Energía

3.3.1.5. Correlación de Spearman

En el análisis de las relaciones monotónicas entre variables, se calculan los coeficientes de correlación de Spearman para todas las combinaciones entre las variables. Estos coeficientes se organizan en la matriz mostrada en la figura 3.11.

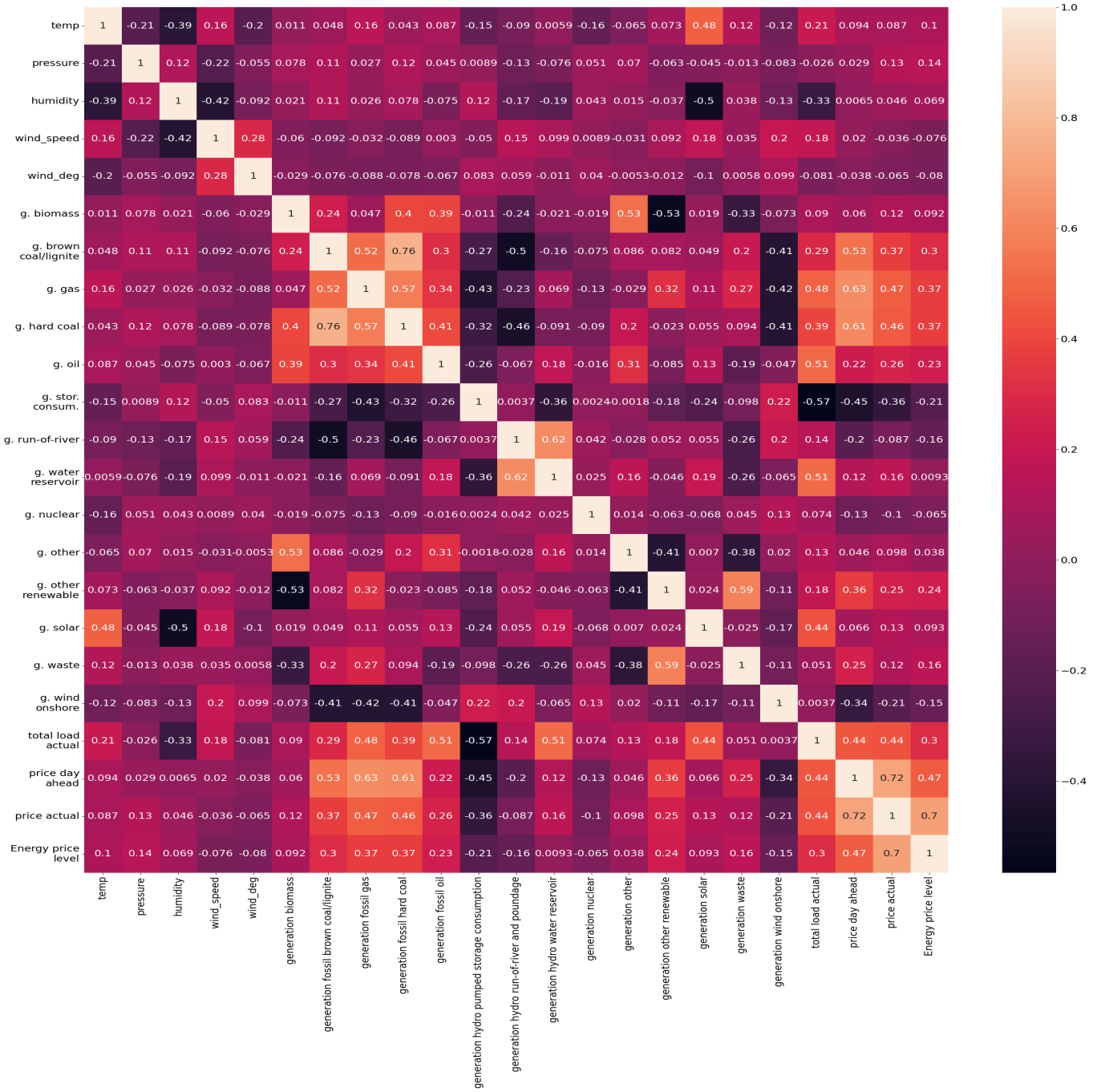

Figura 3.11: Correlación de Spearman conjunto de datos Energía

La tabla 3.9, contiene aquellas correlaciones que superan el umbral establecido de 0.2 con respecto a la variable objetivo.

Variable	Nivel de coste de la energía
Generación de carbón marrón/lignito	0.30
Generación de gas	0.37
Generación de carbón mineral	0.37
Generación de petróleo	0.23
Generación de hidroelectricidad de bombeo	-0.21
Generación de otras renovables	0.24
Carga total actual	0.30
Precio del mercado diario	0.47
Precio actual	0.70

Tabla 3.9: Correlaciones de Spearman por encima del umbral de 0.2 respecto a la variable objetivo

A pesar de que el resto de las correlaciones de Spearman indican relaciones monotónicas más débiles entre las variables en aún pueden ser relevantes en el contexto del análisis estadístico. El estudio de ANOVA decidirá que variables serán usadas en la construcción del modelo.

3.3.1.6. ANOVA unidireccional modificado

La tabla 3.10 presenta el ranking de las 12 mejores variables obtenidas en el análisis de ANOVA unidireccional modificado. Estas 12 variables han sido seleccionadas en el proceso de selección de características debido a que poseen las puntuaciones F más altas, mientras que las demás variables han sido descartadas.

Variable	Puntuación F
Precio actual	26487.59
Precio del mercado diario	5988.70
Generación de carbón mineral	2796.51
Generación de gas	2045.50
Carga total actual	1999.52
Generación de carbón marrón/lignito	1832.21
Generación de otras energías renovables	985.80
Generación de petróleo	966.12
Generación de hidroelectricidad en embalses	550.77
Generación de solar	537.10
Generación de eólica terrestre	455.98
Generación de hidroelectricidad de bombeo	424.91

Tabla 3.10: Ranking de características ANOVA

3.3.1.7. Normalización Min-Max

Se normalización los datos mediante el método Min-Max, como se mencionó en el apartado teórico. La normalización Min-Max ajusta los valores de las variables al rango entre 0 y 1 preservando la relación relativa entre los valores originales. Las figuras 3.12 y 3.13 muestran el cambio en los estadístico básicos de cada variable al aplicar la normalización.

	generation fossil brown coal/lignite	generation fossil gas	generation fossil hard coal	generation fossil oil	generation hydro pumped storage consumption	generation hydro water reservoir	generation other renewable
count	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000
mean	0.448827	0.441751	0.509262	0.506381	0.144362	0.327583	0.561272
std	0.354995	0.183270	0.234670	0.191256	0.231602	0.224785	0.184774
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.311813	0.302429	0.374532	0.000000	0.138599	0.407895
50%	0.510511	0.412482	0.535351	0.513109	0.044156	0.278486	0.592105
75%	0.757758	0.526662	0.698648	0.621723	0.187662	0.473413	0.710526
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figura 3.12: Estadísticos básicos después de la normalización

	generation solar	generation wind onshore	total load actual	price day ahead	price actual
count	35099.000000	35099.000000	35099.000000	35099.000000	35099.000000
mean	0.247148	0.380120	0.463776	0.495728	0.451826
std	0.289983	0.216999	0.199075	0.171288	0.132092
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.012086	0.208304	0.294550	0.386881	0.372430
50%	0.106181	0.344003	0.472708	0.494019	0.452964
75%	0.444406	0.516891	0.615718	0.624556	0.546013
max	1.000000	1.000000	1.000000	1.000000	1.000000

Figura 3.13: Estadísticos básicos después de la normalización

3.3.2. Conjunto de datos de COVID-19

El conjunto de datos de COVID ya está preprocesado y listo para la construcción de modelos.

3.3.2.1. V de Cramér

Las variables del conjunto de datos de COVID-19 son de tipo categórico, no se pueden utilizar las correlaciones de Spearman y Pearson. En su lugar, podemos visualizar las relaciones entre estas variables mediante el coeficiente de Cramer (V de Cramér). El coeficiente de Cramer es una medida de asociación adecuada para evaluar la relación entre dos variables categóricas y se basa en la estadística del chi-cuadrado. Estos coeficientes se representan en una matriz, tal como se ilustra en la Figura 3.14

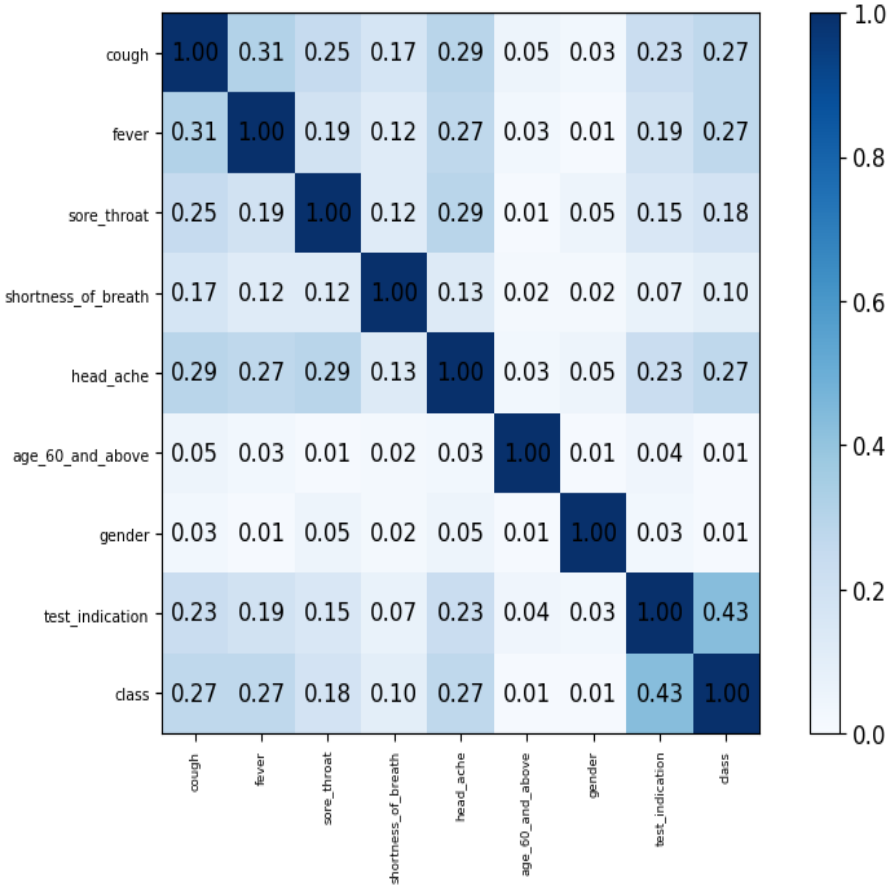

Figura 3.14: Correlación de Cramér para el conjunto de datos COVID-19

La tabla 3.11 muestra aquellas correlaciones de Cramér con umbral mayor de 0.2 sobre la variable objetivo clase.

Variable	Clase COVID
Tos	0.27
Fiebre	0.27
Dolor de cabeza	0.27
Motivo del test	0.43

Tabla 3.11: Mayores correlaciones de Cramér respecto a la clase objetivo COVID-19

En la tabla de correlación de Cramer, se observa que de las 8 variables categóricas, 4 de ellas tienen coeficientes superiores a 0.20, indicando que existe una asociación relevante respecto a la variable objetivo. También se observa como la variable Dolor de garganta con un coeficiente de 0.18 se encuentra cerca del umbral. Por último, las variables Edad de 60 años y más y Género tiene una correlación despreciable respeto a la variable objetivo.

3.3.3. Conjunto de datos de Dengue

En el caso del dataset de Dengue, al igual que con el dataset de Israel, ya ha sido preprocesado y está listo para la construcción de modelos. Las variables en este dataset son de tipo categórico, no es apropiado utilizar las correlaciones de Pearson y Spearman, ya que estas métricas están diseñadas para variables numéricas. Se recurre al uso del coeficiente de Cramér adecuada para medir la asociación entre variables del dataset de Dengue.

3.3.3.1. V de Cramér

La figura 3.15 muestra la matriz de correlaciones de Cramér del conjunto de datos de Dengue

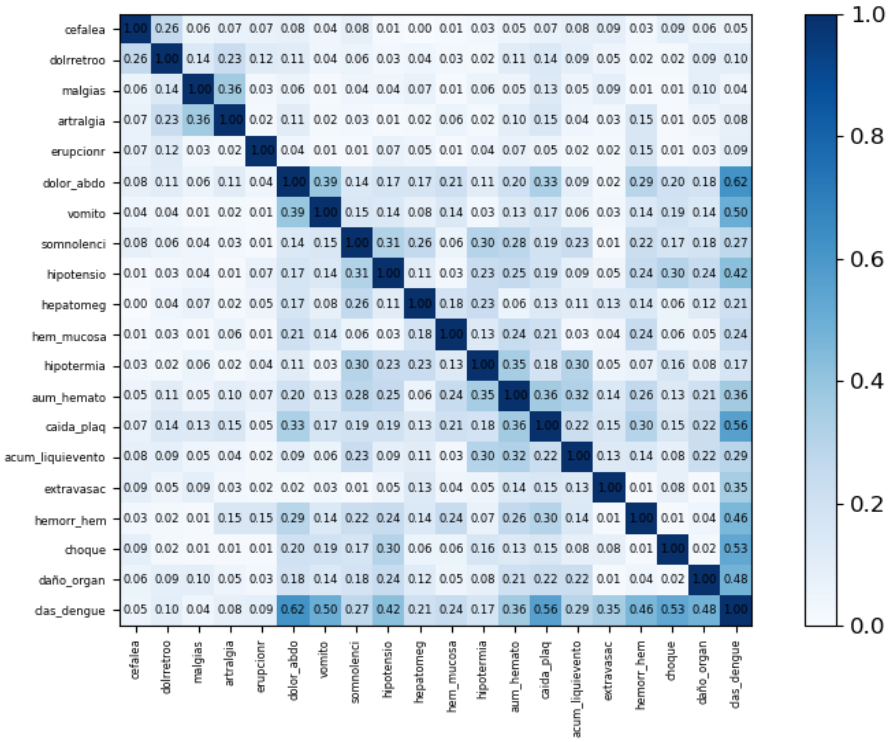

Figura 3.15: Correlación de Cramér conjunto de datos Dengue

La tabla 3.12 muestra las correlaciones que superan el umbral de 0.2 respecto a la variable objetivo (clase Dengue)

Variable	Clase Dengue
Dolor abdominal	0.62
Vómito	0.50
Somnolencia	0.27
Hipotensión	0.42
Hepatomegalia	0.21
Hemorragia mucosa	0.24
Hematocrito alto	0.36
Plaquetas bajas	0.56
Edema	0.29
Extravasación	0.35
Sangrado	0.46
Shock	0.53
Falla orgánica	0.48

Tabla 3.12: Mayores correlaciones de Cramér respecto a la clase objetivo Dengue

Se observa que las variables 'Dolor abdominal', 'Vómito', 'Plaquetas bajas', y 'Shock' muestran una alta correlación, superando un valor de 0.50. Asimismo, las variables 'Hipotensión', 'Sangrado', y 'Falla orgánica' también tienen una correlación cercana a 0.50 respecto a la variable objetivo. Por último, 'Cefalea', 'Dolor Detrás de los ojos', 'Mialgias', 'Artralgia' y 'Erupción' tienen una correlación prácticamente despreciable.

3.4. Mapas Cognitivos Difusos

La tabla 3.13 proporciona una breve descripción de los metadatos de cada conjunto de datos finales, después de ser sometidos a un proceso de preprocesamiento, los cuales se utilizaron en la construcción de los modelos. Se presentan los detalles esenciales de cada conjunto de datos, que incluyen aspectos como el número de variables, la cantidad de instancias, la naturaleza de las variables (numéricas, categóricas) y la cantidad de instancias por cada una de las clases de los conjuntos de datos.

Conjunto de Datos	Registros	No. de variables	No. de variables categóricas	No. de variables numéricas	No. de clases	No. de registros por clase
Dengue [15]	32559	22	22	0	3	10210 11123 11186
Energía [54]	30000	13	1	12	3	10000
COVID-19 [55]	100000	9	9	0	2	50000

Tabla 3.13: Breve descripción de los conjuntos de datos

Como mencionamos previamente, en la etapa de modelado se ha empleado la técnica de los MCD. El cálculo de la matriz de pesos final W, se ha realizado usando un algoritmo basado en poblaciones conocido PSO, el cual ha demostrado ser eficaz en esta tarea. Antes de ejecutar el algoritmo de aprendizaje, es crucial seleccionar ciertos hiperparámetros que afectarán el rendimiento y comportamiento del

algoritmo. Estos hiperparámetros son valores que se eligen de antemano y afectan cómo el PSO busca el óptimo en el espacio de soluciones. La elección de estos hiperparámetros varia según el problema específico y el conjunto de datos utilizado. Fue necesario realizar y experimentar con diferentes valores para encontrar la configuración óptima que produzca el mejor rendimiento del modelo. También fue necesaria la realización de validaciones cruzadas para evaluar su rendimiento de manera más confiable y robusta. Los hiperparámetros ajustados en cada uno de los modelos construidos son el tamaño de la población, el número máximo de iteraciones, el coeficiente cognitivo y el coeficiente social.

En la literatura relacionada con PSO, existe una amplia variedad de artículos que abordan la cuestión de cómo elegir los valores óptimos de los hiperparámetros para lograr una convergencia efectiva de los modelos. En el artículo [57], se llevó a cabo un estudio para investigar cómo los valores de los coeficientes cognitivo y social afectan la convergencia de los modelos en PSO. Tras analizar exhaustivamente diferentes combinaciones de estos hiperparámetros, se llegó a la conclusión de que un valor aproximado de 2 para ambos coeficientes logra los mejores resultados, el uso del mismo valor en ambos coeficientes indica que se da un equilibrio entre la influencia de la información local y global en el movimiento de las partículas durante la búsqueda del óptimo. Esto significa que cada partícula considera tanto su mejor posición local como la mejor posición global del enjambre para guiar su movimiento hacia soluciones potencialmente mejores.

Por otro lado, el artículo [58] realiza un estudio sobre el número óptimo de partículas en distintos conjuntos de datos con diferentes números de variables, con el objetivo de determinar la cantidad adecuada de partículas para lograr la convergencia del modelo. Los resultados de la investigación no son directamente aplicables a los conjuntos de datos de este proyecto, debido a que los resultados obtenidos ofrecían pésimos rendimientos, ha sido necesario un ajuste de dicho hiperparámetro. El proceso llevado a cabo para la construcción de cada uno de los modelos ha sido el siguiente:

1. Se crearon cinco validaciones cruzadas, en las cuales los conjuntos de datos se dividieron en dos partes. El 70 % de los datos se asignó al conjunto de entrenamiento, utilizado para entrenar el modelo, mientras que el 30 % restante se destinó al conjunto de prueba, utilizado para evaluar el rendimiento del modelo. Este enfoque de validación cruzada permite una evaluación más robusta del modelo al repetir el proceso de entrenamiento y evaluación en diferentes particiones de los datos, lo que ayuda a mitigar variaciones en los resultados.

2. Selección de los hiperparámetros del modelo, los cuales comprenden el tamaño de la población, el número máximo de iteraciones, y los coeficientes social y cognitivo.

3. Entrenamiento de los modelos.

4. Evaluación del rendimiento del modelo por medio del uso de las métricas clásicas explicada, si los resultados obtenidos no son satisfactorios, se debe regresar al paso de ajuste de hiperparámetros (paso 2) y realizar nuevos experimentos para encontrar una configuración que mejore el rendimiento del modelo.

3.4.1. Conjunto de datos de Energía

El ajuste de los hiperparámetros del modelo predictivo de Energía con el mejor rendimiento son mostrados en la tabla 3.16.

Hiperparámetro	Valor Asignado
Número máximo de iteraciones	100
Número de partículas de la población	50
Coeficiente Cognitivo	2.01
Coeficiente Social	2.01

Tabla 3.14: Hiperparámetros del conjunto de datos de Energía

El MCD obtenido se muestra en la figura 3.16

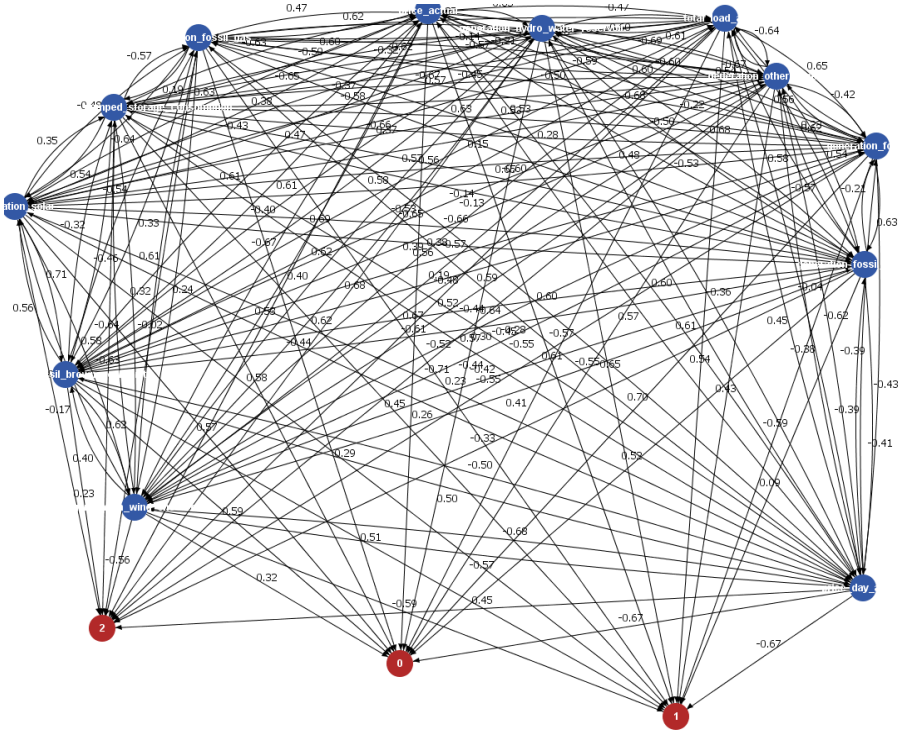

Figura 3.16: Mapa Cognitivo Difuso para predecir costo Energía

Los conceptos representados en color azul corresponden a las variables de entrada, es decir, las características utilizadas para realizar las predicciones.Por otro lado, los conceptos de color rojo representan las clases a predecir, donde el concepto 0 representa el nivel de precio bajo, el concepto 1 representa el nivel de precio medio y el concepto 2 representa el nivel de precio alto.

En la figura 3.16 no se logra apreciar el peso de las relaciones causales, la matriz de pesos de la figura 3.17 proporciona información sobre las relaciones causales entre los diferentes conceptos. Estos pesos indican la importancia o la influencia que cada variable de entrada tiene en la predicción de las clases. Valores más altos de peso sugieren una relación más fuerte entre la variable de entrada y la clase, mientras que valores cercanos a cero indican una relación más débil o insignificante.

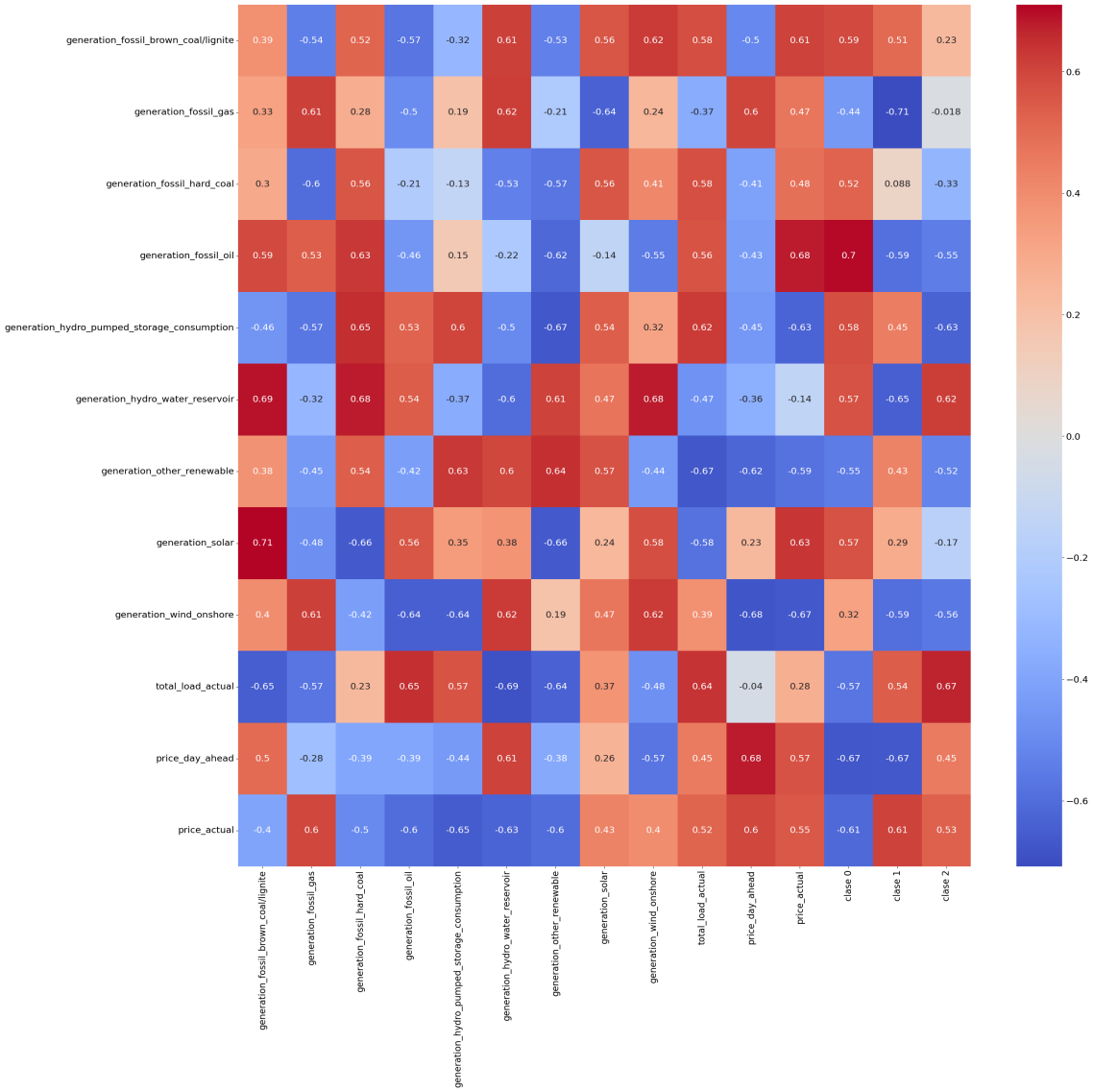

Figura 3.17: Matriz de Pesos para Energía

Al analizar la matriz, se puede observa que va son más relevantes para la predicción y cómo afectan a cada clase en dentro del modelo. La tabla 3.15 muestra los conceptos de entrada con mayor influencia en cada uno de los conceptos de clase.

Orden de Importancia	Clase 0	Clase 1	Clase 2
1	Petróleo	Gas	Carga total actual
2	Precio del mercado diario	Precio del mercado diario	Hidroelectricidad de bombeo
3	Precio actual	Hidroelectricidad de embalses	Hidroelectricidad de embalses
4	Carbón marrón/lignito	Precio actual	Eólica terrestre
5	Hidroelectricidad de bombeo	Petróleo	Petróleo

Tabla 3.15: Variables más importantes basadas en el peso de las relaciones causales para el modelo de Energía

3.4.2. Conjunto de datos de COVID-19

El ajuste de los hiperparámetros del modelo predictivo de COVID-19 con el mejor rendimiento son mostrados en la tabla 3.16.

Hiperparámetro	Valor Asignado
Número máximo de iteraciones	100
Número de partículas de la población	30
Coeficiente Cognitivo	2.01
Coeficiente Social	2.01

Tabla 3.16: Hiperparámetros del conjunto de datos de COVID-19

El MCD obtenido respecto al conjunto de datos de COVID-19 es mostrado en la figura 3.18, donde el concepto 0 indica negativo en el test y el concepto 1 positivo en el test.

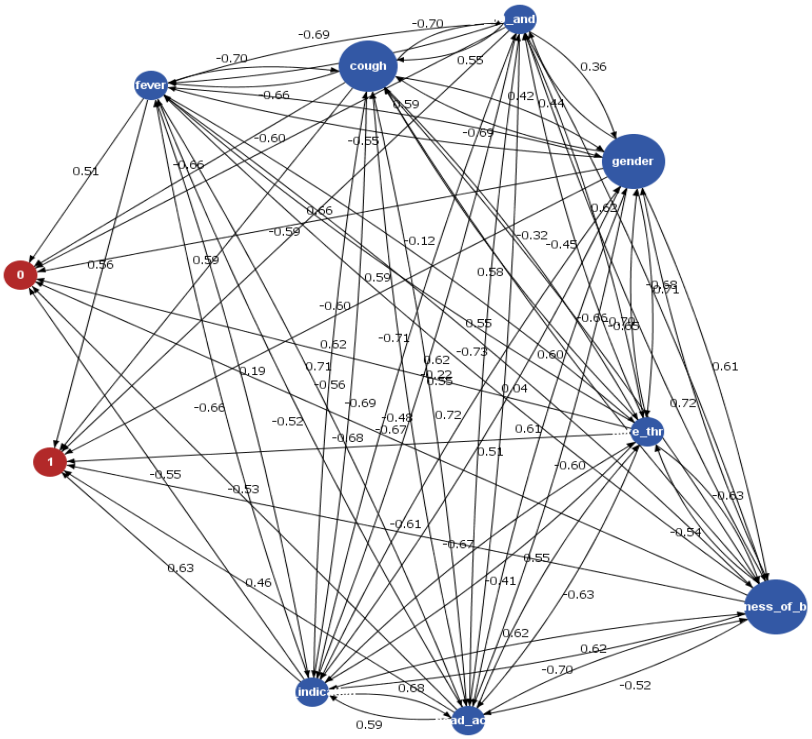

Figura 3.18: Mapa Cognitivo Difuso para predecir el COVID-19

Al igual que en el conjunto de datos de Energía, se presenta la matriz de pesos del conjunto de datos de COVID-19 en la figura 3.19, con el fin de analizar que conceptos de entrada son más influyentes en los conceptos de clase.

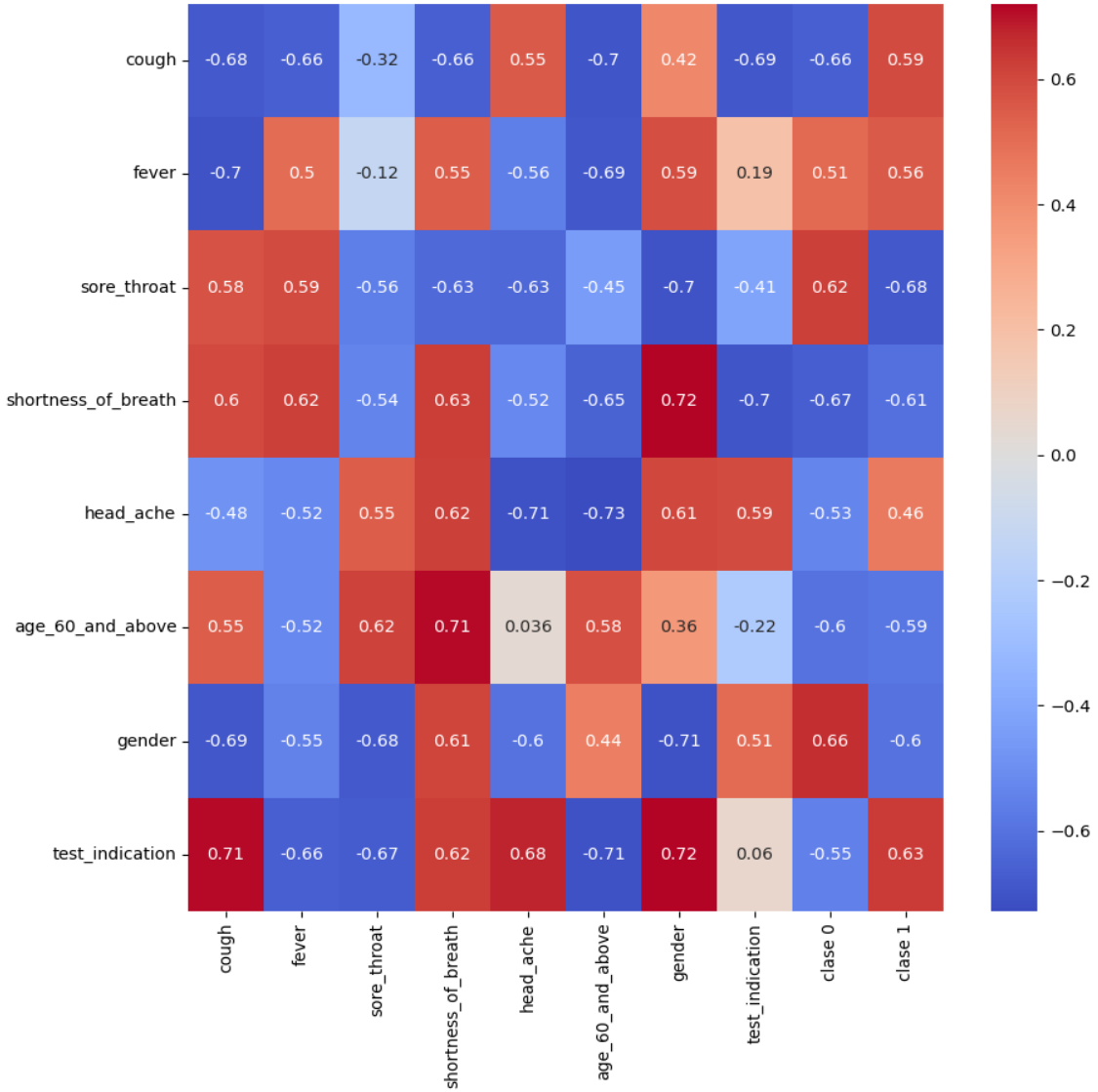

Figura 3.19: Matriz de Pesos para COVID-19

La tabla 3.17 resume cuáles son los conceptos de entrada con mayor influencia respecto a los conceptos de clase.

Orden de Importancia	Clase 0	Clase 1
1	Dificultad al respirar	Dolor de garganta
2	Tos	Motivo del test
3	Género	Dificultad al respirar
4	Dolor de garganta	Género
5	Edad de 60 años y más	Tos

Tabla 3.17: Variables más importantes basadas en el peso de las relaciones causales para el modelo de COVID-19

3.4.3. Conjunto de datos de Dengue

El ajuste de los hiperparámetros del modelo predictivo de Dengue con el mejor rendimiento son mostrados en la tabla 3.18.

Hiperparámetro	Valor Asignado
Número máximo de iteraciones	200
Número de partículas de la población	65
Coeficiente Cognitivo	2.01
Coeficiente Social	2.01

Tabla 3.18: Hiperparámetros del conjunto de datos de Dengue

El modelo del MCD obtenido al entrenar el conjunto de datos de Dengue es mostrado en la figura 3.20. Cada uno de los conceptos de clase representa las distintas clases dentro del conjunto de datos, el concepto 0 indica el nivel de severidad del Dengue 0, el concepto 1 el nivel de severidad de Dengue 1 y el nivel de 2 indica el nivel de severidad de Dengue 2.

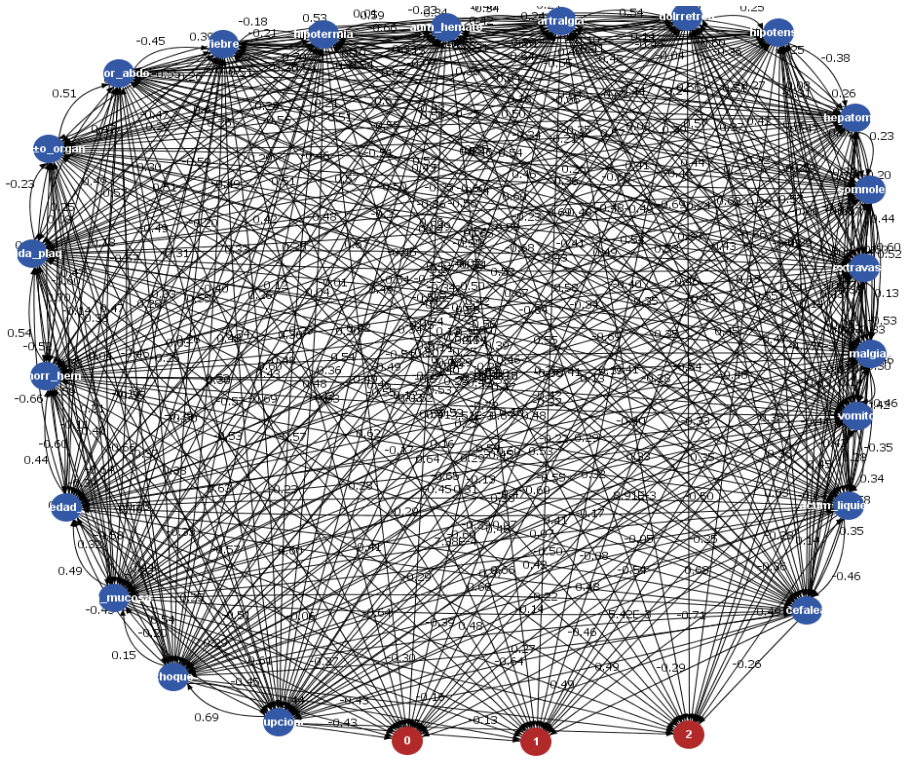

Figura 3.20: Mapa Cognitivo Difuso Dengue

La figura 3.20 presenta dificultades para visualizar las relaciones causales. Por lo tanto, al igual que en los anteriores modelos, se proporciona la matriz de pesos en la figura 3.21 con el fin de analizar las relaciones más influyentes con respecto a los conceptos de clase.

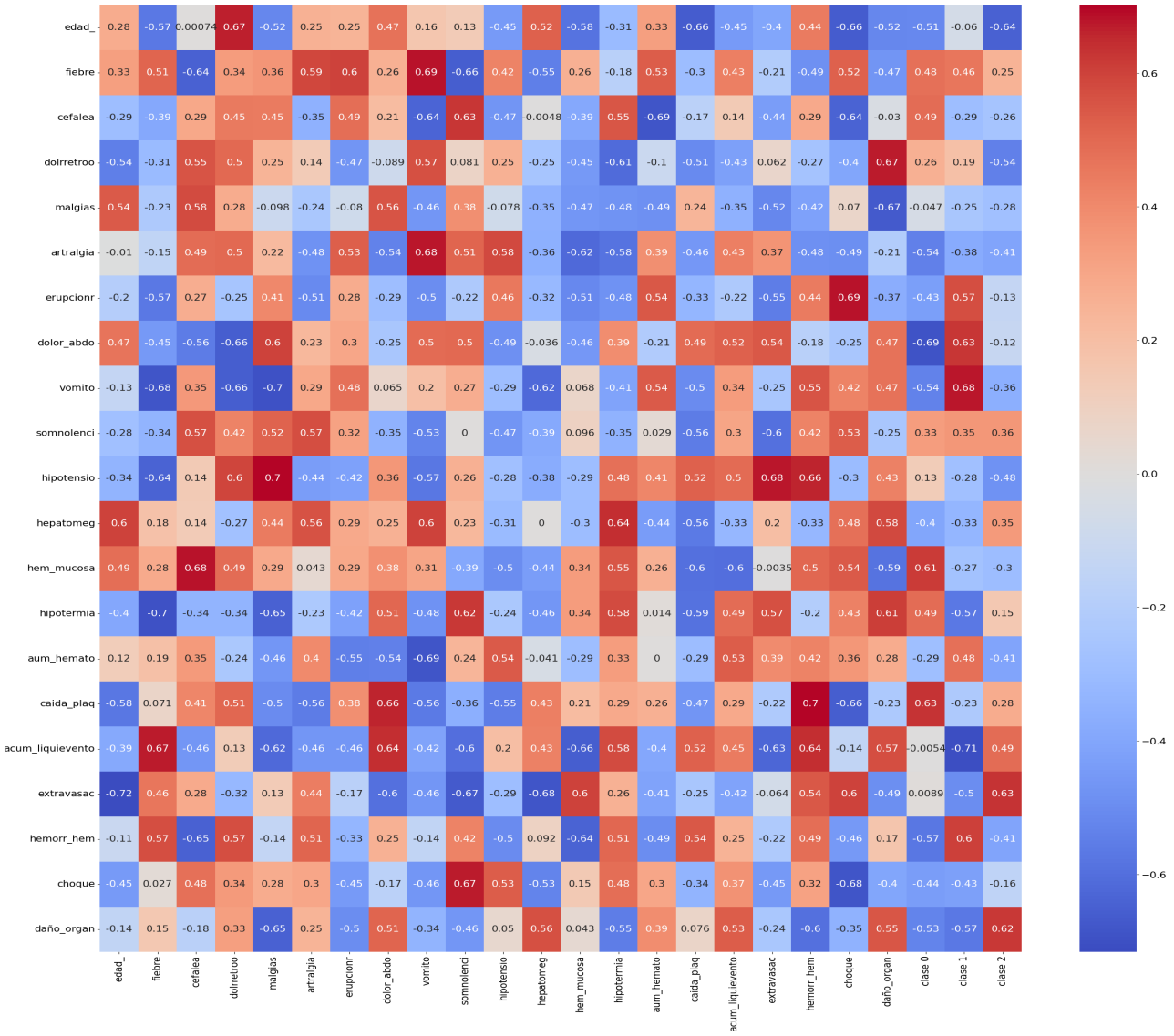

Figura 3.21: Matriz de Pesos Dengue

La tabla 3.19 resume cuáles son los conceptos de entrada con mayor influencia respecto a los conceptos de clase del modelo de Dengue.

Orden de Importancia	Clase 0	Clase 1	Clase 2
1	Dolor abdominal	Edema	Edad
2	Plaquetas bajas	Vómito	Extravasación
3	Hemorragia mucosa	Vómito	Falla orgánica
4	Sangrado	Sangrado	Dolor DO
5	Artralgia	Erupción	Edema

Tabla 3.19: Variables más importantes basadas en el peso de las relaciones causales para el modelo de Dengue

3.5. Evaluación de modelos

En esta sección, se presentan los resultados de rendimiento obtenidos de los modelos finales evaluados utilizando las métricas clásicas previamente definidas en la sección teórica. Estas métricas nos permiten medir la calidad y eficacia de los modelos en función de su capacidad para realizar predicciones precisas y coherentes.

3.5.1. Conjunto de datos de Energía

La matriz de confusión del modelo de Energía, mostrada en la tabla 3.20, revela que la tasa de verdaderos positivos de las clases 0 y 2 es significativamente mayor que la de la clase 1. Por otro lado, la tasa de falsos positivos en todas las clases se encuentra bastante igualada, estos resultados se pueden observar en la tabla 3.21

Matriz Confusión	Clase 0 (Predicha)	Clase 1 (Predicha)	Clase 2 (Predicha)
Clase 0 (Real)	2280	653	51
Clase 1 (Real)	624	1435	985
Clase 2 (Real)	58	366	2548

Tabla 3.20: Matriz de confusión para el modelo de Energía

	Verdaderos Positivos	Falsos positivos
Clase 0	0.76	0.11
Clase 1	0.47	0.17
Clase 2	0.86	0.17

Tabla 3.21: Ratios de verdaderos positivos y falsos positivos para el modelo de Energía

Los resultados de la evaluación de las métricas clásicas del modelo de Energía se presentan en la tabla 3.22. Se observa que las clases 0 y 2, que representan los niveles de precio de energía alto y bajo, respectivamente, muestran valores cercanos de calidad de predicción según las métricas clásicas analizadas. Sin embargo, la clase 1, que corresponde al nivel medio del precio de energía, muestra unos valores considerablemente inferiores en todas las métricas evaluadas. Estos resultados sugieren que el modelo puede tener dificultades para predecir con precisión la clase intermedia del precio de energía. Por otro lado, debido a la baja sensibilidad de la clase 1 en el modelo de Energía, el área bajo la curva ROC (AUC) es menor en comparación con las otras clases. La figura 3.22 respalda la conclusión previamente mencionada. En ella, se puede observar visualmente cómo la curva ROC correspondiente a la clase 1 abarca menos área bajo la curva en comparación con las curvas de las otras clases.

Energía	**Puntuación F1**	**Precisión**	**Sensibilidad**	**AUC**
Clase 0	0.77	0.77	0.76	0.81
Clase 1	0.52	0.58	0.47	0.65
Clase 2	0.78	0.71	0.86	0.82

Tabla 3.22: Resultados de la evaluación para el modelo de Energía

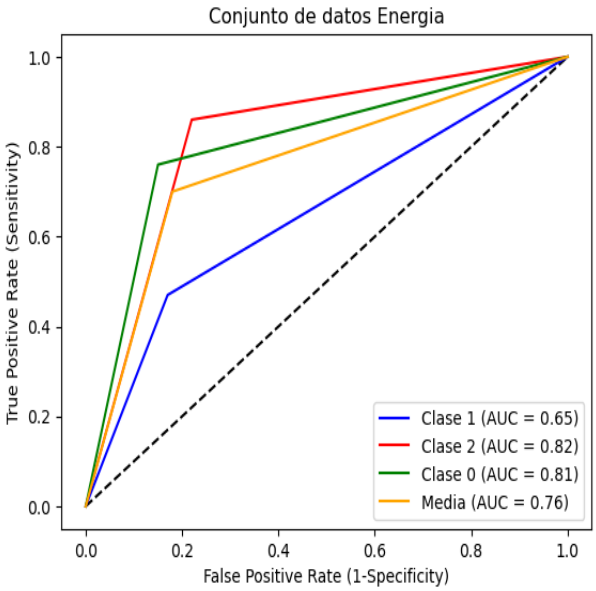

Figura 3.22: Curva ROC del modelo de Energía

3.5.2. Conjunto de datos de COVID-19

La tabla 3.23 presenta la matriz de confusión del modelo de COVID-19, y a partir de sus datos, podemos deducir que la tasa de verdaderos positivos de la clase 1 es menor que la de la clase 1, debido a que en comparación con la clase 0, se clasifican como falsos negativos una gran cantidad de instancias. Como resultado de esto, la tasa de falsos positivos de la clase 0 se ve afectada y se incrementa en tamaño. La tabla 3.24 evidencia los resultados descritos.

Matriz Confusión	Clase 0 (Predicha)	Clase 1 (Predicha)
Clase 0 (Real)	12739	2218
Clase 1 (Real)	5732	9311

Tabla 3.23: Matriz de confusión para el modelo de COVID-19

	Verdaderos Positivos	Falsos positivos
Clase 0	0.85	0.38
Clase 1	0.62	0.15

Tabla 3.24: Ratios de verdaderos positivos y falsos positivos para el modelo de COVID-19

La tabla 3.25 muestra los resultados de evaluación del modelo COVID-19. En cuanto al rendimiento en términos de Puntuación F1 ambas clase están igualdas. La precisión de la clase 0, que se corresponde con los negativos, es bastante menor que la precisión de la clase 1, que se corresponde con los positivos. Sin embargo, la sensibilidad de la clase 0 es mayor que la sensibilidad de la clase 1. En términos de la curva ROC de la figura 3.23 se traduce en que la clase 0 ocupa menor área que la clase 1.

COVID-19	Puntuación F1	Precisión	Sensibilidad	AUC
Clase 0	0.76	0.69	0.85	0.81
Clase 1	0.70	0.81	0.62	0.68

Tabla 3.25: Resultados de la evaluación para el modelo de COVID-19

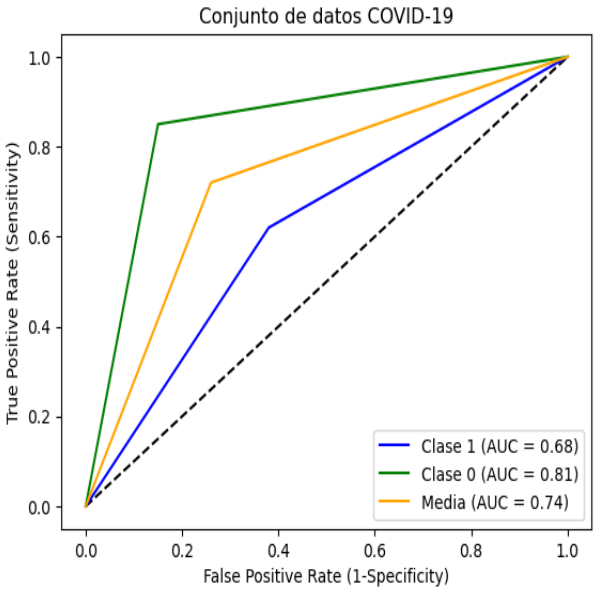

Figura 3.23: Curva ROC para el modelo de COVID-19

3.5.3. Conjunto de datos de Dengue

Con respecto a las matrices de confusión presentadas en las tablas 3.26 y 3.27, es evidente que las tasas de verdaderos positivos de las clases 0 y 2 son más altas en comparación con la tasa de verdaderos positivos de la clase 1. Esto indica que el modelo tiene una mejor capacidad para clasificar correctamente las instancias de las clases 0 y 2. Por otro lado, debido a que la tasa de verdaderos positivos en la clase 1 es menor que en las demás clases, esto tiene un impacto en las tasas de falsos positivos en el resto de las clases.

Matriz Confusión	Clase 0 (Predicha)	Clase 1 (Predicha)	Clase 2 (Predicha)
Clase 0 (Real)	2855	49	186
Clase 1 (Real)	235	2029	1098
Clase 2 (Real)	28	636	2652

Tabla 3.26: Matriz de confusión para el modelo de Dengue

	Verdaderos Positivos	Falsos positivos
Clase 0	0.92	0.04
Clase 1	0.6	0.11
Clase 2	0.8	0.2

Tabla 3.27: Ratios de verdaderos positivos y falsos positivos para el modelo de Dengue

Dengue	Puntuación F1	Precisión	Sensibilidad	AUC
Clase 0	0.92	0.92	0.92	0.93
Clase 1	0.67	0.75	0.6	0.75
Clase 2	0.73	0.67	0.8	0.80

Tabla 3.28: Resultados de la evaluación para el modelo de Dengue

Con lo mencionado anteriormente, podemos notar en la tabla de evaluación de rendimiento en la 3.28 que la sensibilidad de la clase 1 es inferior a la de las clases 0 y 2. Esto implica que el modelo tiene dificultades para identificar correctamente los casos positivos de la clase 1, lo que afecta directamente la curva ROC de esta clase, haciendo que tenga un área menor que el resto de las clases, como se confirma en la figura 3.24. Por otro lado, es notable que la curva ROC de la clase 2 muestra un rendimiento destacado, lo cual significa que el modelo es capaz de distinguir muy bien los casos positivos de esta clase.

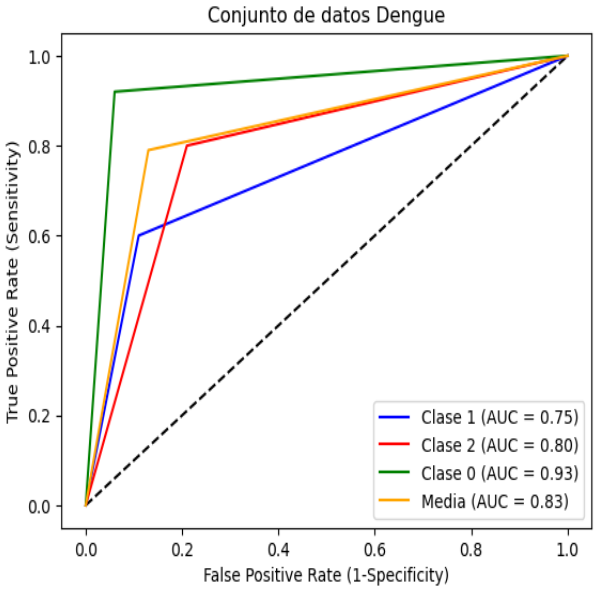

Figura 3.24: Curva ROC para el modelo de Dengue

3.6. Explicabilidad de modelos

En esta sección, se lleva a cabo el análisis basado en explicabilidad de los modelos utilizando los tres enfoques descritos en la sección teórica. Posteriormente, se discuten los resultados obtenidos en cada uno de los métodos aplicados.

3.6.1. LIME

La figura 3.25 ilustra el cálculo de los valores LIME para una predicción en el modelo MCD de Energía para la clase 1, como ejemplo de funcionamiento.

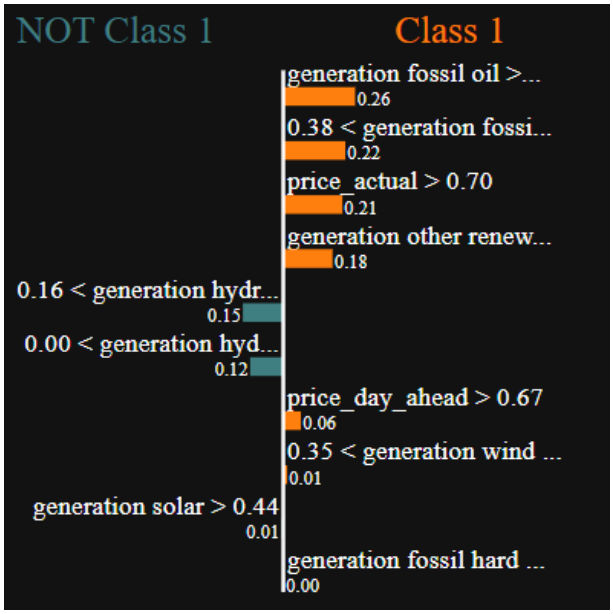

Figura 3.25: Valores LIME para el modelo de Energía de Mapas Cognitivos Difusos

Para analizar los resultados presentados, se toma en consideración lo siguiente:

- El orden de las características refleja su peso en la predicción del modelo, lo que significa que muestra un ranking de cuáles características son más importantes para obtener esa predicción. En este caso, la variable 'Generación de petróleo' tiene una importancia del 26 %, 'Generación de gas' con un 22 % de influencia en la predicción, 'Precio actual' con un peso de un 21 %, y 'Generación de otras renovables' con un 18 %. Estas son las características más influyentes.
- El color de cada característica y el eje x indican cómo influyen positiva o negativamente en la predicción. Las variables 'Generación de petróleo', 'Generación de gas' y 'Precio actual' tienen una influencia positiva en el valor de predicción obtenida, mientras que las variables 'Generación hidroelectricidad de embalses' y 'Generación de hidroelectricidad de bombeo' tienen una influencia negativa en el valor de predicción obtenida.
- El valor al lado de cada característica indica el umbral a partir del cual contribuye positiva o negativamente' a la predicción de la clase 1. Por ejemplo, un valor de 'Generación petróleo' >0.62 y valores de la variable 'Generación gas' comprendidos en el intervalo (0.38,0,58] tienen un impacto positivo en la predicción de clase 1. También, los valores dentro de la característica 'Generación hidroelectricidad de embalses' dentro del intervalo (0.16,0.31] tienen una contribución negativa en la predicción de la clase 1.

Basándonos en la información explicada anteriormente y en la Figura 3.25, podemos extraer las siguientes conclusiones en términos de explicabilidad a partir de la predicción:

- 'Generación de petróleo' es la variable con el peso positivo más alto en la predicción, los valores mayores al umbral de 0.70 contribuyen a la predicción de la clase 1 del precio de consumo de energía.
- 'Generación gas' es la segunda característica con mayor peso positivo en la predicción. Los valores en el intervalo (0.38, 0.58] contribuyen a la predicción de la clase 1.
- 'Precio actual' y 'Generación de otras renovables' son la tercera y cuarta características con mayor peso positivo en la predicción de la clase 1. Valores mayores a 0.70 y 0.71, respectivamente, contribuyen a la predicción.
- 'Generación hidroelectricidad de embalses' y 'hidroelectricidad de bombeo' son las variables con el impacto negativo más significativo en la predicción; sin embargo, su influencia y peso son menores que el resto de las variables ya discutidas. Los valores en los intervalos (0.16,0.31] y (0,0.04] contribuyen negativamente en la predicción realizada.
- El resto de las variables tienen un impacto poco significativo en la predicción realizada.

En resumen, el análisis de explicabilidad realizado a partir de la figura 3.25 nos permite identificar las características más importantes para el modelo, su influencia positiva o negativa, y los umbrales a partir de los cuales comienzan a impactar en la predicción.

Basándonos en las explicaciones de las predicciones individuales, podemos extraer las características más importantes de cada uno de los modelos creados. Al analizar cómo cada modelo realiza predicciones para instancias específicas, podemos identificar las características que tienen una influencia más significativa en el resultado de la predicción. Estas características clave son esenciales para comprender el proceso de toma de decisiones de los modelos y pueden proporcionar información valiosa para refinar y interpretar los modelos de manera más efectiva. La tabla 3.29 muestra las características con mayor influencia en el resultado de las predicciones de cada uno de los modelo construidos.

Orden de importancia	**COVID-19**	**Energía**	**Dengue**
1	Motivo del test	Generación de carbón mineral	Erupción
2	Tos	Carga total actual	Vómito
3	Género	Generación de gas	Shock
4	Fiebre	Precio actual	Hipotensión
5	Dificultad al respirar	Generación de otras renovables	Somnolencia

Tabla 3.29: Características más importantes extraídas a partir de LIME

Basado en el análisis LIME realizado en el modelo de COVID-19, se observa que las características 'Motivo del test', 'Tos' y 'Género' son las más importantes en las predicciones del modelo, con un 33 %, 23 % y 22 % de influencia, respectivamente. Por otro lado, las características 'Fiebre' y 'Dificultad al respirar' tienen una influencia más baja del 7 %.

En el modelo de Energía, la influencia entre las características respecto a las predicciones está más distribuida de manera equitativa. 'Generación de carbón' tiene una influencia del 15 %, 'Carga total actual' tiene una influencia del 13 %, mientras que las variables 'Generación de gas', 'Precio actual' y 'Generación de otras renovables' tienen cada una una influencia del 12 %. Las demás características tienen una influencia cercana al 8 %.

Finalmente, en el conjunto de datos de Dengue, el síntoma que ejerce la mayor influencia en las predicciones es 'Erupción' con un 23 %, seguido de cerca por 'Vómito' con un 19 %. A partir de este punto, los síntomas tienen una influencia más baja, con 'Shock' teniendo un 9 %, 'Somnolencia' y 'Plaquetas Bajas' con un 8 % de influencia en las predicciones.

3.6.2. Importancia de características basada en permutación

La figura 3.26 muestra la importancia de cada característica en el modelo de Energía. Podemos observar que la variable 'Precio actual' es la más sensible a la permutación y, por lo tanto, la más influyente en las predicciones del modelo. Le siguen en importancia las variables 'Generación hidroelectricidad de bombeo' y 'Generación gas'. Por otro lado, las variables 'Generación de carbón mineral' y 'Carga total actual' son las menos influyentes en las predicciones del modelo.

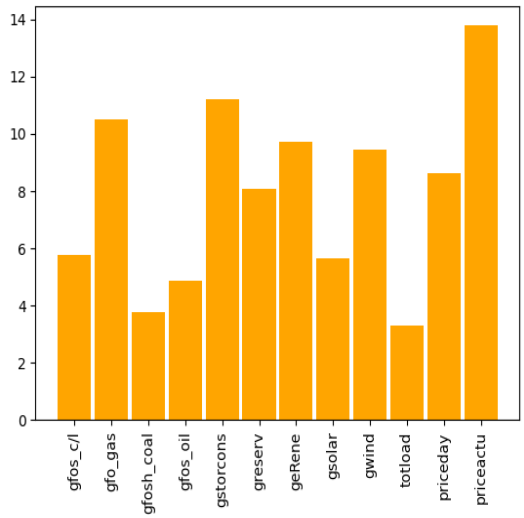

Figura 3.26: Importancia de características permutación Energía

En el modelo de COVID-19 3.26, se destaca claramente que la característica más influyente es 'Motivo del test', seguida de los síntomas 'Dolor de cabeza' y 'Tos'. En contraste, las variables 'Dolor de garganta' y 'Edad' son las menos influyentes en las predicciones del modelo, por lo que tienen un impacto menor dentro del modelo.

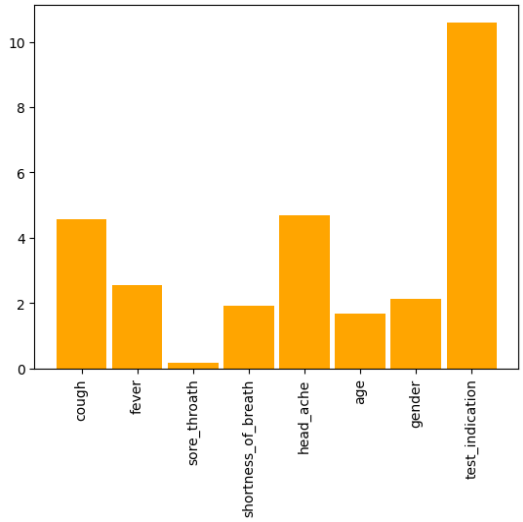

Figura 3.27: Importancia de características permutación COVID-19

En el modelo de Dengue, según se muestra en la Figura 3.28, se puede apreciar que los síntomas 'Hipotermia' y 'Extravasación' tienen la mayor influencia, siendo los más influyentes en el modelo. Por otro lado, los síntomas 'cefalea' y 'Somnolencia' son los menos influyentes y tienen un impacto menor en las predicciones del modelo.

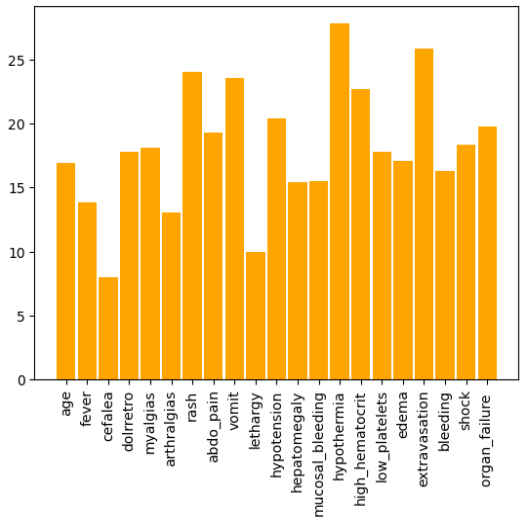

Figura 3.28: Importancia de características permutación Dengue

La siguiente tabla 3.30 resume las características más importantes de cada uno de los modelos.

Orden de Importancia	COVID-19	Energía	Dengue
1	Motivo del test	Precio actual	Hipotermia
2	Dolor de cabeza	Hidroelectricidad de bombeo	Extravasación
3	Tos	Generación de gas	Erupción
4	Fiebre	Generación de otras renovables	Vómito
5	Género	Generación eólica terrestre	Hematocrito alto

Tabla 3.30: Características más importantes extraídas en Importancia de características basada en permutación

3.6.3. Basado en la sensibilidad de las relaciones causales

Es posible obtener una clasificación que indique qué conceptos son más relevantes para cada una de las clases, basándose en el grado de influencia de los conceptos de entrada, es decir, la causalidad. En la sección de modelado se presentan las tablas 3.15, 3.17, 3.19 en las que se muestran las variables más importantes respecto a los conceptos clase en cada uno de los modelos construidos.

El análisis de sensibilidad tiene como objetivo explorar cómo la eliminación de relaciones con menor influencia hacia los conceptos de clase afecta el comportamiento del modelo. Se han definido filtros basados en el peso de las relaciones causales para eliminar conexiones entre los conceptos de entrada y los conceptos de clase. Si el peso de una relación es menor que el valor del filtro, la relación causal entre el concepto y la clase correspondiente se elimina. Estos filtros se aplican a diferentes clases del modelo en varios escenarios. Por ejemplo, en el escenario uno, aplicamos los filtros a la clase 1, mientras que en el escenario dos, los aplicamos a la clase 2.

Los filtros han sido seleccionados teniendo en cuenta la distribución de los pesos de las relaciones causales para cada clase específica, lo que los hace específicos y dependientes del modelo analizado. Al aplicar estos filtros, podemos identificar y eliminar las conexiones menos relevantes entre los conceptos de entrada y las clases, lo que nos permite observar cómo cambia el comportamiento del modelo después de realizar estas modificaciones. La siguiente tabla 3.31 resume los filtros escogidos para cada uno de los modelos. Cada uno de los filtros es más restrictivo que el filtro anterior, en el contexto del modelo de COVID-19, el filtro 1 se encarga de la eliminación de todas las relaciones causales asociadas con conceptos clase cuyo peso sea igual o inferior a 0,55. El filtro 2 suprime las relaciones causales con pesos menores o iguales a 0,62, mientras que el filtro 3 elimina todas las relaciones causales que presenten un peso menor o igual a 0,65

	COVID-19	**Dengue**	**Energía**
Filtro 1	<=0.55	<=0.15	<=0.35
Filtro 2	<=0.62	<=0.30	<=0.50
Filtro 3	<=0.65	<=0.50	<=0.55

Tabla 3.31: Filtros basados en los pesos de las relaciones causales respecto a los conceptos clase

Existe un gran número de relaciones causales entre los conceptos de entrada y los conceptos de clase con pesos más bajos que los filtros definidos. A modo de ejemplo, la figura 3.29 ilustra el porcentaje de relaciones causales eliminadas al aplicar cada uno de los filtros al modelo de Dengue. Se puede observar que al aplicar el filtro de mayor peso, la clase 3 pierde aproximadamente el 80 % de sus relaciones causales, mientras que el resto de las clases pierden en torno a un 70 % de sus relaciones.

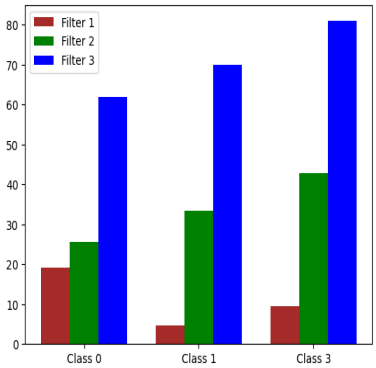

Figura 3.29: Porcentaje de relaciones causales eliminadas respecto a los conceptos clase al aplicar los filtros en Dengue

La tabla 3.32 presenta los porcentajes de relaciones eliminadas al aplicar los filtros para cada clase en los modelos creados.

	COVID-19		**Dengue**			**Energía**		
	Clase 0	Clase 1	clase 0	clase 1	clase 2	clase 0	clase 1	clase 2
Filtro 1	37.5	12.5	19.05	4.76	9.52	8.34	16.7	33.33
Filtro 2	62.5	62.5	25.57	33.33	42.86	16.7	33.33	42.86
Filtro 3	75	75	62	70	80.95	25	50	66.7

Tabla 3.32: Porcentaje de relaciones entre características con respecto a las clases al aplicar filtros a los conjuntos de datos

En el modelo de COVID-19, se observa que al aplicar el primer filtro en la clase 1, se eliminan casi un 40 % de las relaciones causales, mientras que en la clase 0 solo se eliminan alrededor de un 13 %. Por otro lado, la aplicación de los filtros 2 y 3 en ambas clases resulta en la eliminación del mismo número de relaciones causales.

En el modelo de Dengue, el filtro 1 es más restrictivo en la clase 0, seguido de la clase 2 y la clase 1. Por otro lado, el filtro 2 es menos restrictivo en la clase 0, pero es más restrictivo en el resto de las clases. Finalmente, el filtro 3 es el más restrictivo en la clase 2, eliminando casi un 80 %, mientras que en las clases 0 y 1 elimina alrededor del 62 % y 70 % de las relaciones causales, respectivamente.

Por último, en el modelo de Energía, se observa que la aplicación de los filtros es menos restrictiva en la clase 0 en comparación con las clases 1 y 2. Por ejemplo, el filtro de nivel 3 solo elimina un 25 % de las relaciones causales de la clase 0, mientras que en las clases 1 y 2 elimina un 55 % y 67 % de las relaciones, respectivamente. Estos resultados indican que los filtros tienen un impacto más significativo en las clases 1 y 2 y que las variables en la clase 0 tienen una mayor influencia en las predicciones de la clase.

Una vez que se aplicaron los filtros a cada clase en cada modelo, se realizó el análisis de sensibilidad a través de un conjunto instancias de prueba, se observa si existen diferencias significativas en las predicciones entre el modelo original y el modelo modificado. La tabla 3.33 muestra el porcentaje de predicciones que han cambiado en el modelo modificado respecto al modelo original después de eliminar los pesos menos significativos.

	COVID-19		Dengue			Energía		
	clase 0	clase 1	clase 0	clase 1	clase 2	clase 0	clase 1	clase 2
Filtro 1	26.70 %	26.70 %	20 %	20 %	20 %	20 %	33.40 %	26.70 %
Filtro 2	26.70 %	20 %	13.40 %	13.40 %	13.40 %	20 %	26.70 %	26.70 %
Filtro 3	26.70 %	20 %	26.70 %	26.70 %	26.70 %	33.40 %	46.70 %	66.70 %

Tabla 3.33: Porcentaje de instancias cuya predicción cambió debido a los filtros

En el análisis de sensibilidad de los modelos MCD, se puede observar que la eliminación de relaciones causales con menor influencia tiene un impacto significativo. En el modelo de COVID-19, la eliminación de relaciones causales de menor peso en la clase 0 afecta a la predicción en un 26.70 % de los casos, mientras que en la clase 1 afecta al 20 %. En el modelo de Dengue, la eliminación de relaciones causales de menor peso afecta al comportamiento del modelo, cambiando el 26.70 % de las predicciones en todas las clases. El impacto del análisis de sensibilidad es mucho más evidente en el modelo de Energía, donde la eliminación de las relaciones con menor peso tiene un impacto del 33.40 % en las predicciones al eliminarlas de la clase 0. Al eliminarlas de la clase 1, el impacto es del 46.70 %, y al eliminarlas de la clase 2, del 66.70 %.

3.6.4. Análisis de Explicabilidad

La tabla 3.34 proporciona un resumen de las características más influyentes extraídas del análisis de explicabilidad realizado en los modelos MCD, donde el porcentaje en el análisis de LIME indica la influencia de la característica en las predicciones, en Importancia de características basada en permutaciones indica la disminución en la precisión respecto al modelo original y, en el análisis basado en la influencia de las relaciones causales muestra los conceptos con mayor peso (influencia) respecto a los conceptos de clase.

Modelo	LIME	Importancia de características basada en permutación	Basado en la influencia de las relaciones causales
Energía	Generación de carbón mineral 15 % Carga total actual 13 % Generación de gas 12 % Precio actual 12 % Otras renovables 12 %	Precio actual 14 % Hidroelectricidad de bombeo 11 % Generación de gas 11 % Otras renovables 10 % Generación eólica terrestre 9 %	Precio del mercado diario 0.45 Generación de petróleo 0.43 Hidroelectricidad en embalses 0.41 Precio actual 0.41 Generación de gas 0.38
COVID-19	Motivo del test 33 % Tos 23 % Género 22 % Fiebre 7 % Dificultad al respirar 7 %	Motivo del test 11 % Dolor de cabeza 5 % Tos 5 % Fiebre 3 % Género 2 %	Dolor de garganta 0. 66 Dificultad al respirar 0.64 Género 0.63 Tos 0.63 Motivo del test 0.60
Dengue	Erupción 23 % Vómito 19 % Shock 9 % Somnolencia 8 % Plaquetas Bajas 8 %	Hipotermia 28 % Extravasación 26 % Erupción 24 % Hematocrito alto 23 % Vomito 23 %	Dolor abdominal 0.44 Vómito 0.41 Hemorragia mucosa 0.39 Falla orgánica 0.37 Hipotermia 0.35

Tabla 3.34: Resumen de características más importantes en al análisis de explicabilidad realizado

En el modelo de Energía, la característica 'Precio actual' ha sido identificada como importante en todos los análisis realizados. La segunda característica más importante es 'Generación de gas', aunque su posición en los diversos rankings de características de los métodos utilizados no destaca. La variable

'Generación de otras renovables' es considerada importante tanto en LIME como en la importancia de características, pero no destaca en su influencia en los MCD.

.En el modelo de COVID-19, las características con mayor impacto en la predicción del modelo son 'Motivo del test', el síntoma 'Tos' y la variable 'Género'. Las tres variables son reconocidas como importantes en todos los métodos. Por otro lado, la característica 'fiebre' también es considerada importante según los análisis de LIME e importancia de características, pero su influencia dentro del MCD es mínima. Por su parte, la característica 'Dificultad al respirar' es considerada importante según LIME y MCD, aunque no destaca en términos de importancia de características basada en permutación.

Finalmente, en el modelo de Dengue, la característica más influyente encontrada es 'Vómito', la cual es reconocida como importante en todos los métodos. La siguiente característica más importante es 'Erupción', aunque su importancia dentro del modelo MCD es mínima. El síntoma 'Hipotermia' es considerada importante en el análisis de importancia de características basado en permutación y tiene influencia en los pesos del modelo MCD.

El análisis de sensibilidad desarrollado en los MCD nos permite observar que las relaciones causales con menor peso entre los conceptos de entrada y los conceptos de clase tienen un impacto significativo en las predicciones del modelo, y esta dependencia es específica del modelo. Estos hallazgos difieren de los resultados obtenidos en el trabajo [26].

Capítulo 4

Conclusiones y Trabajos Futuros

4.1. Conclusiones

Los modelos MCD construidos pueden no tener un rendimiento de calidad en métricas clásicas en comparación con otros modelos de caja negra. Sin embargo, su ventaja es que son autoexplicativos, lo que nos permite obtener las relaciones causales entre conceptos y ver cómo afectan a las predicciones del modelo. Esta capacidad de explicabilidad es valiosa para comprender el funcionamiento interno del modelo y entender qué factores o características influyen más en las predicciones. Aunque puedan no tener un rendimiento tan alto como otros modelos más complejos, su transparencia y facilidad de interpretación pueden ser beneficiosas en ciertos contextos y escenarios.

El empleo de múltiples métodos de explicabilidad en un mismo modelo permite obtener una comprensión más completa y detallada de cómo se toman decisiones dentro del modelo y qué características tienen un mayor impacto en sus predicciones. A partir del análisis de explicabilidad empleando los distintos métodos se ha obtenido que en el modelo de Energía las variables más importantes son 'Precio actual' y 'Generación de gas'. En el modelo de COVID-19, las variables más importantes son 'Motivo del test', el síntoma 'Tos' y la variable 'Género'. Por ultimo, en el modelo de Dengue el síntoma 'Vómito' es la variable más importante.

El análisis de sensibilidad de la influencia de características menos influyentes basada en causalidad permite identificar y evaluar la importancia relativa de las características menos influyentes en un MCD. En lugar de centrarse exclusivamente en las características más influyentes, este algoritmo se enfoca en determinar cómo las características de menor peso afectan las predicciones y los resultados del modelo. La base fundamental de este enfoque es comprender cómo las relaciones causales entre los conceptos de entrada y los conceptos de clase impactan en la toma de decisiones. Como se ha comprobado, las características menos influyentes aún pueden tener un impacto significativo en las predicciones. En el modelo de Energía, se revela cómo la eliminación de las relaciones causales de menor peso tiene un impacto de considerable magnitud en las predicciones generadas por dicho modelo. Estos hallazgos tienen importantes implicaciones para el diseño, desarrollo y ajuste de modelos ya que subrayan la necesidad de considerar no solo las relaciones causales con más influencia dentro del modelo, sino también aquellas con menor influencia, en la toma de decisiones relacionadas con el modelo. Este análisis, que se basa en las características menos influyentes tiene un potencial bastante prometedor, ya que puede ofrecer una perspectiva valiosa para comprender cómo incluso las características que inicialmente se consideran de menor importancia pueden desempeñar un papel crucial en los resultados de un modelo.

En última instancia, el uso modelos autoexplicativos como los MCD en contraposición a los modelos de caja negra, junto con la implementación de métodos de explicabilidad, brinda a los usuarios la posibilidad de obtener una mayor confianza en los modelos y posibilita su incorporación en dominios críticos. La capacidad de comprender y validar el funcionamiento interno de los modelos construidos en el trabajo permite su aplicación como sistemas de apoyo en la toma de decisiones en los dominios energético y médico, lo que permite a los usuarios tomar decisiones más informadas y acertadas.

4.2. Trabajos Futuros

En futuras líneas de investigación, se podría profundizar en los siguientes aspectos relacionados con el uso de modelos autoexplicativos como los MCD y métodos de explicabilidad:

- Estudios comparativos entre distintas técnicas de construcción de modelos en términos de rendimiento y explicabilidad: la comparación de los métodos de interpretabilidad en diferentes técnicas de IA explicables proporciona una mejor comprensión de los conjuntos de datos con los que estamos trabajando.
- Mejora del rendimiento: Explorar distintas estrategias para mejorar el rendimiento de los MCD. Se puede investigar cómo optimizar la construcción de relaciones causales entre conceptos.
- Integración de técnicas: Investigar cómo combinar modelos autoexplicativos con modelos de caja negra para aprovechar las ventajas de ambos enfoques como en los estudios del estado de arte del proyecto.
- Evaluación de explicaciones: Desarrollar métricas y criterios para evaluar la calidad de las explicaciones proporcionadas por los métodos de explicabilidad.

En resumen, en futuras investigaciones se debería seguir explorando uso de modelos autoexplicativos como los MCD y métodos de explicabilidad para fomentar una mayor confianza y comprensión en la toma de decisiones basadas en modelos de aprendizaje automático.

Bibliografía

[1] N. Panwar, S. Kaushik y S. Kothari, «Role of renewable energy sources in environmental protection: A review,» *Renewable and Sustainable Energy Reviews*, vol. 15, n.º 3, págs. 1513-1524, 2011.

[2] Danish, M. A. Baloch, N. Mahmood y J. W. Zhang, «Effect of natural resources, renewable energy and economic development on CO2 emissions in BRICS countries», *Science of The Total Environment*, vol. 678, págs. 632-638, 2019.

[3] D. Ormandy y V. Ezratty, «Thermal discomfort and health: protecting the susceptible from excess cold and excess heat in housing», *Advances in Building Energy Research*, vol. 10, págs. 84-98, 2016.

[4] K. K. Jaiswal, C. R. Chowdhury, D. Yadav et al., «Renewable and sustainable clean energy development and impact on social, economic, and environmental health», *Energy Nexus*, vol. 7, págs. 100-118, 2022.

[5] M. Shahbaz, M. Zakaria, S. J. H. Shahzad y M. K. Mahalik, «The energy consumption and economic growth nexus in top ten energy-consuming countries: Fresh evidence from using the quantile-on-quantile approach», *Energy Economics*, vol. 71, págs. 282-301, 2018.

[6] F. Lui, Y. Khan y T. Hassan, «Does excessive energy utilization and expansion of urbanization increase carbon dioxide emission in Belt and Road economies?», *Environ Sci. Pollut Res*, vol. 30, págs. 60 080-6010, 2023.

[7] M. Akcin, A. Kaygusuz, A. Karabiber, S. Alagoz, B. B. Alagoz y C. Keles, «Opportunities for energy efficiency in smart cities», *2016 4th International Istanbul Smart Grid Congress and Fair (ICSG), IEEE*, págs. 1-5, 2016.

[8] A. González-Vidal, A. P. Ramallo-González, F. Terroso-Sáenz y A. Skarmeta, «Data driven modeling for energy consumption prediction in smart buildings», en *IEEE International Conference on Big Data (Big Data)*, 2017, págs. 4562-4569.

[9] D. Mariano-Hernández, L. Hernández-Callejo, A. Zorita-Lamadrid, O. Duque-Pérez y F. Santos García, «A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect diagnosis», *Journal of Building Engineering*, vol. 33, pág. 101 692, 2021.

[10] A. González-Vidal, F. Jiménez y A. F. Gómez-Skarmeta, «A methodology for energy multivariate time series forecasting in smart buildings based on feature selection», *Energy and Buildings*, vol. 196, págs. 71-82, 2019.

[11] A. G. Billé, A. Gianfreda, F. Del Grosso y F. Ravazzolo, «Forecasting electricity prices with expert, linear, and nonlinear models», *International Journal of Forecasting*, vol. 39, n.º 2, págs. 570-586, 2023.

[12] R. Syah, M. Rezaei, M. Elveny et al., «Day-ahead electricity price forecasting using WPT, VMI, LSSVM-based self adaptive fuzzy kernel and modified HBMO algorithm», vol. 11, págs. 2045-2322, 2021.

[13] Z. Meng, H. Sun y X. Wang1, «Forecasting Energy Consumption Based on SVR and Markov Model: A Case Study of China», *Frontiers in Environmental Science*, vol. 10, 2022.

[14] D. Ioannis, D. J. Apostolopoulos y P. P. Groumpos, «Advanced fuzzy cognitive maps: state-space and rule-based methodology for coronary artery disease detection», *Biomedical Physics Engineering Express*, vol. 7, n.º 4, pág. 045 007, 2021.

[15] W. Hoyos, J. Aguilar y M. Toro, «A clinical decision-support system for dengue based on fuzzy cognitive maps», *Health Care Manag Sci*, vol. 25, n.º 4, págs. 666-681, 2022.

[16] E. Puerto, J. Aguilar, C. López y D. Chávez, «Using Multilayer Fuzzy Cognitive Maps to diagnose Autism Spectrum Disorder», *Applied Soft Computing*, vol. 75, págs. 58-71, 2019.

[17] V. Mpelogianni, P. Marnetta y P. P. Groumpos, «Fuzzy Cognitive Maps in the Service of Energy Efficiency», *IFAC-PapersOnLine*, vol. 48, n.º 24, págs. 1-6, 2015.

[18] K. Poczeta, Papageorgiou y E. I. Papageorgiou, «Energy Use Forecasting with the Use of a Nested Structure Based on Fuzzy Cognitive Maps and Artificial Neural Networks», *Energies*, vol. 15, n.º 20, pág. 7542, 2022.

[19] M. Alipour, R. Hafezi, E. Papageorgiou, M. Hafezi y M. Alipour,

[20] C. Shearer, «The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing», *Data Warehousing*, vol. 5, págs. 13-22, 2000.

[21] J. L. Salmeron, S. A. Rahimi, A. M. Navali y A. Sadeghpour, «Medical diagnosis of Rheumatoid Arthritis using data driven PSO–FCM with scarce datasets», *Neurocomputing*, vol. 232, págs. 104-112, 2017.

[22] J. B. Raja y S. C. Pandian, «PSO-FCM based data mining model to predict diabetic disease», *Computer Methods and Programs in Biomedicine*, vol. 196, pág. 105 659, 2020.

[23] G. Nápoles, M. Leon Espinosa, I. Grau, K. Vanhoof y R. Bello, «Fuzzy Cognitive Maps Based Models for Pattern Classification: Advances and Challenges», en 2018, págs. 83-98.

[24] A. D. Ioannis y P. P. Groumpos, «Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence», *Applied Sciences*, vol. 13, n.º 6, 2023.

[25] M. Taha y V. Sunil, «Explainable fault prediction using learning fuzzy cognitive maps», *Expert Systems*, e13316,

[26] C. Murungweni, M. Van Wijk, J. Andersson, E. Smaling y K. Giller, «Application of Fuzzy Cognitive Mapping in Livelihood Vulnerability Analysis», *Ecology and Society - ECOL SOC*, vol. 16, dic. de 2011.

[27] A. Al Farsi, D. Petrovic y F. Doctor, «A Non-Iterative Reasoning Algorithm for Fuzzy Cognitive Maps based on Type 2 Fuzzy Sets», *Information Sciences*, vol. 622, págs. 319-336, 2023.

[28] S. García y F. Herrera, «An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons», *Journal of Machine Learning Research*, vol. 9, págs. 2677-2694, 2008.

[29] D. Singh y B. Singh, «Investigating the impact of data normalization on classification performance», *Applied Soft Computing*, vol. 97, pág. 105 524, 2020.

[30] S. Aksoy y R. M. Haralick, «Feature normalization and likelihood-based similarity measures for image retrieval», *Pattern Recognition Letters*, vol. 22, n.º 5, págs. 563-582, 2001.

[31] J. de Carlos Sola y J. Sevilla, «Importance of input data normalization for the application of neural networks to complex industrial problems», *IEEE Transactions on Nuclear Science*, vol. 44, págs. 1464-1468, 1997.

[32] C.-W. Hsu, C.-C. Chang y C.-J. Lin, «A Practical Guide to Support Vector Classication», 2008.

[33] N. Japkowicz y S. Stephen, «The class imbalance problem: A systematic study», *Intell. Data Anal.*, vol. 6, págs. 429-449, 2002.

[34] A. Orriols-Puig y E. Bernadó-Mansilla, «The class imbalance problem in learning classifier systems: a preliminary study», en *Annual Conference on Genetic and Evolutionary Computation*, 2005.

[35] A. Ali, S. M. H. Shamsuddin y A. L. Ralescu, «Classification with class imbalance problem: A review», en *Soft Computing Models in Industrial and Environmental Applications*, 2015.

[36] R. Potolea y C. Lemnaru, «A Comprehensive Study of the Effect of Class Imbalance on the Performance of Classifiers», en *International Conference on Enterprise Information Systems*, 2016.

[37] N. Chawla, K. Bowyer, L. Hall y W. Kegelmeyer, «SMOTE: Synthetic Minority Over-sampling Technique», *J. Artif. Intell. Res. (JAIR)*, vol. 16, págs. 321-357, 2002.

[38] B. Kosko, «Fuzzy cognitive maps», *International Journal of Man-Machine Studies*, vol. 24, n.º 1, págs. 65-75, 1986.

[39] L. Zadeh, «Fuzzy sets», *Information and Control*, vol. 8, n.º 3, págs. 338-353, 1965.

[40] R. Axelrod, *Structure of decision: The cognitive maps of political elites*. 2015, págs. 1-405.

[41] A. Jose, «A survey about Fuzzy Cognitive maps papers», *International Journal of Computational Cognition*, vol. 3, ene. de 2005.

[42] N. Gonzalo, M. Leon Espinos, G. Isel, V. Koen y B. Rafael, «Fuzzy Cognitive Maps Based Models for Pattern Classification: Advances and Challenges», en 2018, págs. 83-98.

[43] P. Riccardo, K. James y B. Tim, «Particle Swarm Optimization: An Overview», *Swarm Intelligence*, vol. 1, oct. de 2007.

[44] Y. Shi y R. Eberhart, «A modified particle swarm optimizer», en *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, 1998, págs. 69-73.

[45] F. Doshi-Velez y B. Kim, «Towards A Rigorous Science of Interpretable Machine Learning», 2017.

[46] N. M. Jhon Lee, «Trust, control strategies and allocation of function in human-machine systems», págs. 1243-1270, 1992.

[47] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce y H. P. Beck, «The role of trust in automation reliance», *International Journal of Human-Computer Studies*, vol. 58, n.º 6, págs. 697-718, 2003.

[48] B. Goodman y S. Flaxman, «European Union Regulations on Algorithmic Decision-Making and a Right to Explanation,» *AI Magazine*, vol. 38, n.º 3, págs. 50-57, 2017.

[49] P. J. Phillips, C. Hahn, P. Fontana, D. Broniatowski y M. Przybocki, «Four Principles of Explainable Artificial Intelligence», 2020.

[50] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti y D. Pedreschi, «A Survey of Methods for Explaining Black Box Models», *ACM Comput. Surv.*, vol. 51, n.º 5, 2018.

[51] D. V. Carvalho, E. M. Pereira y J. S. Cardoso, «Machine Learning Interpretability: A Survey on Methods and Metrics», *Electronics*, vol. 8, n.º 8, 2019.

[52] C. Molnar, *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, 2.ª ed. 2022.

[53] R. M. Tulio, S. Sameer y G. Carlos, «"Why Should I Trust You?": Explaining the Predictions of Any Classifier», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, págs. 1135-1144.

[54] N. JHANA, «Hourly energy demand generation and weather», 2019.

[55] E. Mathieu, H. Ritchie, L. Rodés-Guirao et al., «Coronavirus Pandemic (COVID-19)», *Our World in Data*, 2020.

[56] «Dengue and Dengue Grave», *Secretaría de Salud de Medellín*, 2020.

[57] Y. V. R. Naga Pawan y K. B. Prakash, «Impact of Inertia Weight and Cognitive and Social Constants in Obtaining Best Mean Fitness Value for PSO», en *Soft Computing for Problem Solving*, Springer Singapore, 2020, págs. 197-206.

[58] A. P. Piotrowski, J. J. Napiorkowski y A. E. Piotrowska, «Population size in Particle Swarm Optimization», *Swarm and Evolutionary Computation*, vol. 58, pág. 100 718, 2020.

Apéndice A

Tecnologías utilizadas

A lo largo del desarrollo de este proyecto, se han empleado diversas tecnologías para lograr sus objetivos. A continuación, se presenta una breve descripción de dichas tecnologías:

A.1. Recursos Hardware

Un ordenador con sus respectivos periféricos

A.2. Recursos Software

- Sistema Operativo Windows 10.
- Fuentes de datos abiertas como Kaggle.
- Lenguaje de programación Python 3.
- Jupyter Notebook.
- FCM Expert
- Distribución de Python y Jupyter Notebook de Anaconda.
- Sistema de Composición de textos Latex.

Apéndice B

Código Fuente

A continuación, se incluye el código fuente que ha sido utilizado en el desarrollo de este proyecto. Este código se presenta en diversos segmentos, ordenados de acuerdo a su uso a lo largo del proyecto.

Listado B.1: Carga de paquetes

```
#Carga de las librerias
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import pandas as pd
import random
import seaborn as sns
import association_metrics as am
import lime
import lime.lime_tabular
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import f_classif
from collections import Counter as Counter
from scipy import stats
from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import KBinsDiscretizer
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE, ADASYN


#Opciones de visiualización de los dataframes
pd.options.mode.chained_assignment = None
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows',None)
```

Listado B.2: Carga Conjunto de datos de Energía

```
#Comentario
datasetTiempo=pd.read_csv('C:/Users/diego/Desktop/MapasCognitvosDifusos/datasets/Enteros/tiempo.csv')
datasetEnergia=pd.read_csv('C:/Users/diego/Desktop/MapasCognitvosDifusos/datasets/Enteros/energia.csv')
datasetTiempo.rename(columns = {'dt_iso':'time'}, inplace = True)
dataset=pd.merge(datasetTiempo, datasetEnergia, on="time")
datasetValencia=dataset[dataset['city_name'] == "Valencia"]
datasetValencia=datasetValencia.drop(["city_name"], axis=1)
datasetValencia=datasetValencia.drop(["time"], axis=1)
filasDatasetValencia=datasetValencia.shape[0]
columnasDatasetValencia=datasetValencia.shape[1]
print("Número de registros dentro de la muestra: "+str(filasDatasetValencia)+" muestras.")
print("Número de características dentro de la muestra: "+str(columnasDatasetValencia)+" características.")
#Visualización de los estadísticos principales del conjunto de datos
datasetValencia.describe()
```

Listado B.3: Eliminación de variables que contienenen unicamente valores nulos o con información redundante

```
datasetValencia=datasetValencia.drop(["generation fossil coal-derived gas"], axis=1)
datasetValencia=datasetValencia.drop(["generation fossil oil shale"], axis=1)
datasetValencia=datasetValencia.drop(["generation fossil peat"], axis=1)
datasetValencia=datasetValencia.drop(["generation geothermal"], axis=1)
datasetValencia=datasetValencia.drop(["generation marine"], axis=1)
datasetValencia=datasetValencia.drop(["generation wind offshore"], axis=1)
datasetValencia=datasetValencia.drop(["forecast wind offshore eday ahead"], axis=1)
datasetValencia=datasetValencia.drop(["generation hydro pumped storage aggregated"], axis=1)
datasetValencia=datasetValencia.drop(["rain_1h"], axis=1)
datasetValencia=datasetValencia.drop(["rain_3h"], axis=1)
datasetValencia=datasetValencia.drop(["snow_3h"], axis=1)
datasetValencia=datasetValencia.drop(["total load forecast"], axis=1)
datasetValencia=datasetValencia.drop(["clouds_all"], axis=1)
datasetValencia=datasetValencia.drop(["weather_id"], axis=1)
datasetValencia=datasetValencia.drop(["forecast solar day ahead"], axis=1)
datasetValencia=datasetValencia.drop(["forecast wind onshore day ahead"], axis=1)
datasetValencia=datasetValencia.drop(["weather_main"], axis=1)
datasetValencia=datasetValencia.drop(["weather_description"], axis=1)
datasetValencia=datasetValencia.drop(["weather_icon"], axis=1)
datasetValencia=datasetValencia.drop(["temp_min"], axis=1)
datasetValencia=datasetValencia.drop(["temp_max"], axis=1)
```

Listado B.4: Visualizacion del número valores nulos de cada una de las variables

```
datasetValencia.isnull().sum()
```

Listado B.5: Visualizacion de la fila donde se encuentra cada uno de los valores nulos

```
datasetValencia[datasetValencia.isnull().any(axis=1)]
```

Listado B.6: Eliminación de valores nulos

```
datasetValencia=datasetValencia.dropna(axis=0)
```

Listado B.7: Función de visualizacion de Outliers de cada una de las variables por el método del rango intercuartílico

```
def visualizacionOutliersDimensiones(dataframe,indiceColumnaInicio,indiceColumnaFinal):
    for indiceColumna in range(indiceColumnaInicio,indiceColumnaFinal):
        outliers=0
```

```
55          percentil25=dataframe.iloc[:,indiceColumna].quantile(0.25)
56          percentil75=dataframe.iloc[:,indiceColumna].quantile(0.75)
57          iqr=percentil75-percentil25
58          LimiteSuperior=percentil75+1.5*iqr
59          LimiteInferior=percentil25-1.5*iqr
60
61          outliers+=dataframe[dataframe.iloc[:,indiceColumna]>LimiteSuperior].shape[0]
62          outliers+=dataframe[dataframe.iloc[:,indiceColumna]<LimiteInferior].shape[0]
63
64
65          print("Dimension actual: "+str(dataframe.columns[indiceColumna]))
66          print("Numero de Outliers en la dimension: "+str(outliers))
67          print("")
```

Listado B.8: LLamada a la función de visualización de valores atípicos

```
68 visualizacionOutliersDimensiones(datasetValencia,0,22)
```

Listado B.9: Diagrama de Caja y Bigotes de una de las variables con valores atípicos

```
69 plt.boxplot(datasetEnergia["price actual"])
70 fig = plt.figure(figsize =(10, 7))
71 plt.show()
```

Listado B.10: Función de dustitución de valores atípicos por la mediana

```
72 def sustitucionOutliersDimensionesMediana(dataframe):
73     numeroDimensiones=dataframe.shape[1]
74     dimensionInicial=0
75     for indice in range (dimensionInicial,numeroDimensiones-1):
76         percentil25=dataframe.iloc[:,indice].quantile(0.25)
77         percentil75=dataframe.iloc[:,indice].quantile(0.75)
78         iqr=percentil75-percentil25
79         LimiteSuperior=percentil75+1.5*iqr
80         LimiteInferior=percentil25-1.5*iqr
81         outliersLimiteSuperior=dataframe[dataframe.iloc[:,indice]>LimiteSuperior]
82         outliersLimiteInferior=dataframe[dataframe.iloc[:,indice]<LimiteInferior]
83         dataframeOutliers=pd.concat([outliersLimiteSuperior,outliersLimiteInferior])
84         dataframeOutliers.reset_index(drop=True, inplace=True)
85
86         if(dataframeOutliers.shape[0]>0):
87             mediana=dataframe.iloc[:,indice].median()
88             dataframeFiltrado=dataframe[(dataframe.iloc[:,indice]<=LimiteSuperior)
89             & (dataframe.iloc[:,indice]>=LimiteInferior)]
90             dataframeOutliers.iloc[:,indice]=mediana
91
92             dataframe=pd.concat([dataframeFiltrado,dataframeOutliers])
93             dataframe.reset_index(drop=True,inplace=True)
94
95     return dataframe
```

Listado B.11: Llamada a la función de sustitución de valores Atípicos por la mediana

```
96 datasetEnergia=sustitucionOutliersDimensionesMediana(datasetValencia)
```

Listado B.12: Discretización de la variable price actual

```
97 clases=KBinsDiscretizer(n_bins=3,encode='ordinal',strategy="uniform").
98 fit_transform(datasetEnergia[['price actual']])
99 clases=pd.DataFrame(clases)
```

```
datasetEnergia = pd.concat([datasetEnergia, clases], axis=1)
datasetEnergia.rename(columns = {0:'Energy price level'}, inplace = True)
Counter(datasetEnergia['Energy price level'])
```

Listado B.13: Correlación de Pearson

```
cadena=['temp', 'pressure', 'humidity', 'wind_speed', 'wind_deg',
       'g. biomass', 'g. brown\ncoal/lignite',
       'g. gas', 'g. hard coal',
       'g. oil', 'g. stor.\nconsum.',
       'g. run-of-river',
       'g. water\nreservoir', 'g. nuclear',
       'g. other', 'g. other\nrenewable', 'g. solar',
       'g. waste', 'g. wind\nonshore', 'total load\nactual',
       'price day\nahead', 'price actual', 'Energy price\nlevel']
correlacionPearson=datasetEnergia.corr()
plt.figure(figsize=(30,30))
plt.rc("font",size=17)
sns.heatmap(correlacionPearson, annot=True)
plt.yticks(np.arange(0.5,len(cadena)+0.5,1),cadena,)
```

Listado B.14: Correlación de Spearman

```
cadena=['temp', 'pressure', 'humidity', 'wind_speed', 'wind_deg',
       'g. biomass', 'g. brown\ncoal/lignite',
       'g. gas', 'g. hard coal',
       'g. oil', 'g. stor.\nconsum.',
       'g. run-of-river',
       'g. water\nreservoir', 'g. nuclear',
       'g. other', 'g. other\nrenewable', 'g. solar',
       'g. waste', 'g. wind\nonshore', 'total load\nactual',
       'price day\nahead', 'price actual', 'Energy price\nlevel']
correlacionSpearman=datasetEnergia.corr("spearman")
plt.figure(figsize=(30,30))
plt.rc("font",size=17)
sns.heatmap(correlacionSpearman, annot=True)
plt.yticks(np.arange(0.5,len(cadena)+0.5,1),cadena,)
```

Listado B.15: Función que crea el ranking de características de ANOVA

```
def seleccionDimensionesAnova(dataframe,numeroCaracteristicas):
    numeroColumnas=dataframe.shape[1]
    X=dataframe.iloc[:,0:numeroColumnas-1]
    Y=dataframe.iloc[:,numeroColumnas-1:numeroColumnas]
    bestfeatures=SelectKBest(score_func=f_classif,k=numeroCaracteristicas)
    fit=bestfeatures.fit(X,Y)
    dfscores=pd.DataFrame(fit.scores_)
    dfcolumns=pd.DataFrame(X.columns)
    featureScores=pd.concat([dfcolumns,dfscores],axis=1)
    featureScores.columns = ['Nombre de la dimension','Puntuacion']
    print(featureScores.nlargest(numeroCaracteristicas,'Puntuacion'))
```

Listado B.16: Función que crea el ranking de características de ANOVA

```
seleccionDimensionesAnova(datasetEnergia,12)
```

Listado B.17: Selecciónd de las características del conjunto de datos basado en el ranking de ANOVA

```
datasetEnergia=datasetEnergia[['generation fossil brown coal/lignite',
'generation fossil gas','generation fossil hard coal',
```

```
'generation_fossil_oil','generation_hydro_pumped_storage_consumption',
'generation_hydro_water_reservoir','generation_other_renewable',
'generation_solar','generation_wind_onshore','total_load_actual',
'price_day_ahead','price_actual','Energy_price_level']].copy()
```

Listado B.18: Normalización del conjunto de datos MIN-MAX

```
datos=datasetEnergia
datasetEnergia=(datasetEnergia-datasetEnergia.min())/(datasetEnergia.max()-datasetEnergia.min())
datasetEnergia["Energy_price_level"]=datos["Energy_price_level"]
```

Listado B.19: Descripción de estadísticos del conjunto de datos después de la normalización

```
datasetEnergia.describe()
```

Listado B.20: Balanceo de datos con SMOTE

```
sm = SMOTE(random_state=42)
datasetEnergia, y_resampled = SMOTE().fit_resample(datasetEnergia,datasetEnergia['Energy_price_level'] )
print(sorted(Counter(y_resampled).items()))
```

Listado B.21: Carga y visualización de V de Crámer en el conjunto de datos de COVID-19

```
datasetIsrael=pd.read_csv('C:/Users/diego/Desktop/
MapasCognitvosDifusos/datasets/ValidacionCruzadaIsrael/Israel.csv')
df = datasetIsrael.apply(lambda x: x.astype("category") if x.dtype == "int64" else x)
cramers_v = am.CramersV(df)
cfit = cramers_v.fit().round(2)
# Instantiating a figure and axes object
fig, ax = plt.subplots(figsize = (10, 6))
cax = ax.imshow(cfit.values, interpolation='nearest', cmap='Blues', vmin=0, vmax=1)
ax.set_xticks(ticks = range(len(cfit.columns)),labels = cfit.columns)
ax.set_yticks(ticks = range(len(cfit.columns)),labels = cfit.columns)
ax.tick_params(axis = "x", labelsize =6, labelrotation = 90)
ax.tick_params(axis = "y", labelsize =8, labelrotation = 0)
fig.colorbar(cax).ax.tick_params(labelsize = 12)
for (x, y), t in np.ndenumerate(cfit):
    ax.annotate("{:.2f}".format(t),
                xy = (x, y),
                va = "center",
                 ha = "center").set(color = "black", size = 6)
```

Listado B.22: Carga y visualización de V de Crámer en el conjunto de datos de Dengue

```
datasetDengue=pd.read_csv('C:/Users/diego/Desktop/
MapasCognitvosDifusos/datasets/ValidacionCruzadaDengue/Dengue.csv')
df = datasetDengue.apply(lambda x: x.astype("category") if x.dtype == "int64" else x)pandas.DataFrame (df)
cramers_v = am.CramersV(df)
df = datasetDengue.apply(lambda x: x.astype("category") if x.dtype == "int64" else x)pandas.DataFrame (df)
cramers_v = am.CramersV(df)of the passed
cfit = cramers_v.fit().round(2)
cfit
fig, ax = plt.subplots(figsize = (10, 6))
cax = ax.imshow(cfit.values, interpolation='nearest', cmap='Blues', vmin=0, vmax=1)
ax.set_xticks(ticks = range(len(cfit.columns)),labels = cfit.columns)
ax.set_yticks(ticks = range(len(cfit.columns)),labels = cfit.columns)
ax.tick_params(axis = "x", labelsize =6, labelrotation = 90)
ax.tick_params(axis = "y", labelsize =6, labelrotation = 0)
fig.colorbar(cax).ax.tick_params(labelsize = 12)
for (x, y), t in np.ndenumerate(cfit):
```

```
194     ax.annotate("{:.2f}".format(t),
195                 xy = (x, y),
196                 va = "center",
197                 ha = "center").set(color = "black", size = 6)
```

Listado B.23: Curva ROC Modelo Energía

```
198 sensibilidadClase1=[0,0.47,1]
199 OpuestoEspecificidadClase1=[0,0.17,1]
200
201 sensibilidadClase2=[0,0.86,1]
202 OpuestoEspecificidadClase2=[0,0.22,1]
203
204 sensibilidadClase0=[0,0.76,1]
205 OpuestoEspecificidadClase0=[0,0.15,1]
206
207 sensibilidadMedia=[0,0.70,1]
208 OpuestoEspecificidadMedia=[0,0.18,1]
209
210 diagonalReferenciaX=[0,1]
211 diagonalReferenciaY=[0,1]
212
213 plt.plot(diagonalReferenciaX,diagonalReferenciaY,"r--",color="black")
214 plt.plot(OpuestoEspecificidadClase1,sensibilidadClase1,color="BLUE",label = "Clase␣1␣(AUC␣=␣0.65)")
215 plt.plot(OpuestoEspecificidadClase2,sensibilidadClase2,color="RED",label = "Clase␣2␣(AUC␣=␣0.82)")
216 plt.plot(OpuestoEspecificidadClase0,sensibilidadClase0,color="GREEN",label = "Clase␣0␣(AUC␣=␣0.81)")
217 plt.plot(OpuestoEspecificidadMedia,sensibilidadMedia,color="ORANGE",label = "Media␣(AUC␣=␣0.76)")
218 plt.ylabel("True␣Positive␣Rate␣(Sensitivity)")
219 plt.xlabel("False␣Positive␣Rate␣(1-Specificity)")
220 plt.title("Conjunto␣de␣datos␣Energia")
221 plt.legend()
222 plt.show()
```

Listado B.24: Curva ROC Modelo COVID-19

```
223 sensibilidadClase0=[0,0.85,1]
224 OpuestoEspecificidadClase1=[0,0.38,1]
225
226 sensibilidadClase1=[0,0.62,1]
227 OpuestoEspecificidadClase2=[0,0.15,1]
228
229
230 sensibilidadMedia=[0,0.72,1]
231 OpuestoEspecificidadMedia=[0,0.26,1]
232
233 diagonalReferenciaX=[0,1]
234 diagonalReferenciaY=[0,1]
235
236 plt.plot(diagonalReferenciaX,diagonalReferenciaY,"r--",color="black")
237
238 plt.plot(OpuestoEspecificidadClase1,sensibilidadClase1,color="BLUE",label = "Clase␣1␣(AUC␣=␣0.68)")
239 plt.plot(OpuestoEspecificidadClase0,sensibilidadClase0,color="GREEN",label = "Clase␣0␣(AUC␣=␣0.81)")
240 plt.plot(OpuestoEspecificidadMedia,sensibilidadMedia,color="ORANGE",label = "Media␣(AUC␣=␣0.74)")
241 plt.ylabel("True␣Positive␣Rate␣(Sensitivity)")
242 plt.xlabel("False␣Positive␣Rate␣(1-Specificity)")
243 plt.title("Conjunto␣de␣datos␣COVID-19")
244 plt.legend()
245 plt.show()
```

Listado B.25: Curva ROC Modelo Dengue

```
sensibilidadClase1=[0,0.60,1]
OpuestoEspecificidadClase1=[0,0.11,1]

sensibilidadClase2=[0,0.80,1]
OpuestoEspecificidadClase2=[0,0.21,1]

sensibilidadClase0=[0,0.92,1]
OpuestoEspecificidadClase0=[0,0.06,1]

sensibilidadMedia=[0,0.79,1]
OpuestoEspecificidadMedia=[0,0.13,1]

diagonalReferenciaX=[0,1]
diagonalReferenciaY=[0,1]

plt.plot(diagonalReferenciaX,diagonalReferenciaY,"r--",color="black")

plt.plot(OpuestoEspecificidadClase1,sensibilidadClase1,color="BLUE",label = "Clase 1 (AUC = 0.75)")
plt.plot(OpuestoEspecificidadClase2,sensibilidadClase2,color="RED",label = "Clase 2 (AUC = 0.80)")
plt.plot(OpuestoEspecificidadClase0,sensibilidadClase0,color="GREEN",label = "Clase 0 (AUC = 0.93)")
plt.plot(OpuestoEspecificidadMedia,sensibilidadMedia,color="ORANGE",label = "Media (AUC = 0.83)")
plt.ylabel("True Positive Rate (Sensitivity)")
plt.xlabel("False Positive Rate (1-Specificity)")
plt.title("Conjunto de datos Dengue")
plt.legend()
plt.show()
```

Listado B.26: Ejecución método basado en LIME

```
#Creacion del objeto explicador LIME
#En class names especificamos la lista de los distintos target del dominio en el que nos encontremos
#Se debe tener en cuenta el orden con el que espeficiamos los etiquetas del
#target al creado el explicador cuando pasemos
#la matriz de probabilidades de la
explainer=lime.lime_tabular.LimeTabularExplainer(matrizDatos,feature_names=columnasDatos,
class_names=['Class 0','Class 1','Class 2'],verbose=True, mode='classification')
instancias=muestrasEnergia.iloc[:, :-1]
#Vecinos es lo que vamos a pasar nuestros modelos para obtener la etiqueta de target
#El resto de valores devueltos son usados por otros metodos internos de la clase
#Recibe como parametros la instancia a explicar y el numero de vecinos que deseamos
#crear respecto a esa instancia
data,scaled_data,vecinos,distances=explainer.crearVecinos(instancia,10)
vecinos=pd.DataFrame(vecinos)
vecinos.columns=columnasDatos
display(vecinos)
#Creacion de la explicacion de la instancia, recibe como parametros:
#La instancia sobre la que vamos a realizar la explicacion
#Numero de caracteristicas que queremos que se tengan en cuente a la hora de realizar la explicacion
#Matriz con las probabilidades creada anteriormnete
#Variables con datos auxiliares que son usadas por el explicador
#Matriz de definicion de probabilidades de pertenencia a la clase de cada uno de los vecinos creados.
matriz=[[1,0,0],[0,0,1],[1,0,0],[0,1,0],[1,0,0],[1,0,0],[0,1,0],[0,1,0],[1,0,0],[0,1,0]]
exp=explainer.explain_instance(instancia, num_features=10,matriz=matriz,distances=distances,
data=data,scaled_data=scaled_data)
exp.show_in_notebook(show_table=True)
```

Listado B.27: Código método de permutación basada en rangos

```
298 #Debido a que cada los modelos se encuentran en java, es necesario realizar la permutaciíon
299 #de cada una de las variables del conjunto de datos de forma manual,
300 #motivo por el que se usa la variable indice
301
302 indice=1
303 df1=datasetEnergia
304
305
306 columna=df1.iloc[:,indice]
307 maximoV=columna.max()
308 minimoV=columna.min()
309
310 for i in range(0, len(columna)):
311     columna.loc[i]=random.uniform(minimoV,maximoV)
312
313 columna=pd.DataFrame(columna)#Convierto la columna en Dataframe
314 df1=df1.drop(df1.columns[indice], axis=1)#Elimino la columna barajeada
315 columna=columna.reset_index(drop=True)
316 df1=pd.concat([df1,columna], axis=1)#Agregamos la columna barajeada
317 df1=df1[['generation␣fossil␣brown␣coal/lignite',
318 'generation␣fossil␣gas','generation␣fossil␣hard␣coal',
319 'generation␣fossil␣oil','generation␣hydro␣pumped␣storage␣consumption',
320 'generation␣hydro␣water␣reservoir','generation␣other␣renewable',
321 'generation␣solar','generation␣wind␣onshore',
322 'total␣load␣actual','price_day_ahead',
323 'price_actual','class']] #Reodenamos el dataframe
324 entrenamiento,validacion=train_test_split(df1,test_size = 0.3)
325 entrenamiento.to_csv(('C:/Users/diego/Desktop/
326 MapasCognitvosDifusos/Explicabilidad/metodoCarlos/entrenamiento.csv'),index=False)
327 validacion.to_csv(('C:/Users/diego/Desktop/
328 MapasCognitvosDifusos/Explicabilidad/metodoCarlos/validacion.csv'),index=False)
```

Universidad de Alcalá
Escuela Politécnica Superior

ESCUELA POLITECNICA
SUPERIOR

Universidad
de Alcalá