Universidad de Alcalá
Escuela Politécnica Superior

Máster Universitario en Analítica de Negocio y Big Data

Trabajo Fin de Máster

Un Enfoque Dinámico de Explicabilidad para los Mapas Cognitivos Difusos

ESCUELA POLITECNICA SUPERIOR

Autor: Diego Javier Benito Gutiérrez

Tutor: Marçal Mora Cantallops

Cotutor: Jose Lisandro Aguilar castro

2025

UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

Máster Universitario en Analítica de Negocio y Big Data

Trabajo Fin de Máster

Un Enfoque Dinámico de Explicabilidad para los Mapas Cognitivos Difusos

Autor: Diego Javier Benito Gutiérrez

Tutor: Marçal Mora Cantallops

Cotutor: Jose Lisandro Aguilar castro

Tribunal:

Presidente: Name of the tribunal president

Vocal 1º: Name of the first vocal

Vocal 2º: Name of the second vocal

Fecha de depósito: 15 de junio de 2025

Agradecimientos

Este trabajo representa una etapa muy importante en mi formación académica y personal, y quiero aprovechar este momento para expresar mi más profundo agradecimiento a todas las personas que han formado parte de este recorrido.

En primer lugar, quiero extender mi gratitud a mi familia, especialmente a mis padres, Gladys y Javier, quienes han sido mi principal apoyo y motivación; sin ellos, no habría sido posible superar esta etapa.

Tampoco puedo dejar de mencionar a mis profesores, quienes, con su dedicación, han contribuido significativamente a mi formación y aprendizaje. Agradecimientos especiales a mi tutor académico, Marçal Mora Cantallops, así como a mi cotutor en IMDEA Networks Institute, José Lisandro Aguilar Castro, y a mi jefe de equipo, Antonio Fernández Anta, por su apoyo, supervisión y confianza, que han sido fundamentales para la realización de este proyecto.

Finalmente, agradezco a mis compañeros del laboratorio por su colaboración, buen ambiente y por compartir conmigo el día a día durante el desarrollo de esta investigación.

Resumen

Este trabajo presenta un enfoque novedoso de explicabilidad dinámica aplicado a los Mapas Cognitivos Difusos (MCD) en tareas de clasificación. El objetivo es analizar las relaciones causales y la evolución temporal de los conceptos a lo largo del proceso de razonamiento. Se propone un método de explicabilidad local que permite evaluar la importancia relativa de las variables a través del tiempo basada en sus relaciones causales, facilitando así una interpretación más precisa y detallada del comportamiento del modelo. El método fue evaluado en cuatro conjuntos de datos: dengue, COVID-19, diabetes y fallos en vehículos submarinos autónomos. Se comparó su rendimiento explicativo con métodos clásicos como SHapley Additive exPlanations (SHAP), Feature Permutatio (FP), y medidas de centralidad basadas en teoría de grafos. Además, se analizo la calidad de las explicaciones generadas por el método propuesto mediante el enfoque ROAR (RemOve And Retrain), y se verifico que cumpliera con las propiedades deseables en los métodos de explicabilidad. Los resultados demuestran que las explicaciones obtenidas son coherentes con la dinámica de los MCD, superando en calidad a las obtenidas con SHAP y FP. Se concluye que la propuesta mejora significativamente la comprensión y la confianza en los MCD en tareas de clasificación, posicionándose como una herramienta valiosa en contextos sensibles donde la explicabilidad es un requisito fundamental.

Palabras clave: Inteligencia Artificial Explicable, Causalidad, Mapas Cognitivos Difusos, Aprendizahe Automático, Clasificación.

Abstract

This work presents a novel approach to dynamic explainability applied to Fuzzy Cognitive Maps (FCM) in classification tasks. The objective is to analyze the causal relationships and the temporal evolution of concepts throughout the reasoning process. A local explainability method is proposed that allows evaluating the relative importance of variables over time based on their causal relationships, thus facilitating a more precise and detailed interpretation of the model's behavior. The method was evaluated on four datasets: dengue, COVID-19, diabetes, and failures in autonomous underwater vehicles. Its explanatory performance was compared with classical methods such as SHapley Additive exPlanations (SHAP), Feature Permutation (FP), and centrality measures based on graph theory. Also, the quality of the explanations generated by the proposed method is analyzed using the ROAR (RemOve And Retrain) approach, and the fulfillment of desirable properties in explainability method is verified. The results demonstrate that the explanations obtained are consistent with the dynamics of FCM, surpassing in quality those obtained with SHAP and FP. It is concluded that the proposal significantly improves the understanding and trust of FCMs in classification tasks, positioning itself as a valuable tool in sensitive contexts where explainability is a fundamental requirement.

Keywords: ,Explainable Artificial Intelligence, Causality, Fuzzy Cognitive Maps, Machine Learning, Classification.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1 Contexto y Justificación

La creciente adopción de modelos de inteligencia artificial (IA) en entornos críticos como la medicina, la ingeniería de sistemas, la ciberseguridad o la administración pública, entre otros ámbitos, ha generado una atención creciente hacia la necesidad de transparencia, fiabilidad e interpretabilidad en los procesos de decisión automatizados. Si bien los avances recientes en aprendizaje automático han permitido desarrollar modelos de alta capacidad predictiva, estos a menudo presentan un comportamiento opaco que dificulta la comprensión del razonamiento interno que guía sus decisiones. Este fenómeno, ampliamente conocido como el ”problema de la caja negra”, ha impulsado el desarrollo del campo de la Inteligencia Artificial Explicable (IAE), cuya finalidad es dotar a los sistemas inteligentes de mecanismos que permitan entender, auditar y justificar su comportamiento de forma comprensible para humanos, sin renunciar a su potencia técnica.

En los últimos años, la investigación en IAE ha ganado un papel central dentro del desarrollo de sistemas inteligentes, especialmente a raíz del despliegue masivo de modelos altamente complejos como las redes neuronales profundas o las ténicas de ensamblado como *random forests* o *boosting*. A pesar del excelente rendimiento de estas técnicas en tareas de clasificación, predicción y generación, los modelos que generan carecen de interpretabilidad inherente, lo que dificulta su validación, genera desconfianza en usuarios finales, y puede conllevar a riesgos de aceptación en entornos sensibles [1], [2].

Este conflicto entre precisión y transparencia ha motivado la aparición de métodos de explicabilidad, entre las que destacan métodos locales agnósticos como *Local Interpretable Model-agnostic Explanations (LIME)* [3], aproximaciones basadas en teoría de juegos como *SHapley Additive exPlanations (SHAP)* [4], y modelos visuales como *Gradient-weighted Class Activation Mapping (Grad-CAM)* [5]. Sin embargo, gran parte de estas propuestas han sido diseñadas para modelos discriminativos estáticos basado en el comportamiento de los datos, y no consideran el comportamiento iterativo o dinámico de ciertos sistemas de inferencia, limitando así su capacidad explicativa en escenarios más complejos. En este contexto, resulta crucial desarrollar nuevas metodologías de explicabilidad capaces de capturar la evolución temporal y causal de los modelos, particularmente en aquellos que presentan una estructura explícita de razonamiento causal como los Mapas Cognitivos Difusos (MCDs).

Los MCDs se han consolidado como una herramienta de modelado especialmente adecuada para representar sistemas complejos, dinámicos y con incertidumbre inherente. Su estructura basada en grafos dirigidos ponderados permite integrar conocimiento experto e inferencia basada en relaciones causales difusas, lo cual facilita su aplicación en dominios como el diagnóstico médico, la predicción de fenómenos

sociales, el análisis de sistemas industriales, o el diseño de políticas públicas [6]-[8]. A pesar de su reconocida interpretabilidad estructural, el análisis de explicabilidad en MCDs ha estado tradicionalmente limitado a medidas estáticas, como las métricas de centralidad en teoría de grafos, la reducción de la red conceptual, o el estudio de pesos causales sin considerar su evolución temporal [9]-[13]. Estos enfoques, si bien útiles, presentan limitaciones al momento de capturar la dinámica real del modelo durante el proceso de inferencia, donde las influencias causales entre conceptos no son constantes, sino que varían a medida que el sistema se actualiza. Esta laguna metodológica motiva la necesidad de enfoques explicativos que no solo identifiquen qué conceptos son relevantes, sino también cómo y cuándo emergen esas influencias a lo largo del tiempo, en función de la evolución interna del sistema.

El presente trabajo parte de la hipótesis de que, para lograr una comprensión profunda y contextual del comportamiento de los MCDs, es necesario un enfoque de explicabilidad que no solo identifique las relaciones relevantes, sino que también analice cómo evoluciona dinámicamente el comportamiento del modelo a partir de una instancia específica a predecir. En este contexto, se propone un método de explicabilidad dinámico y local para MCDs, diseñado específicamente para analizarlas relaciones causales y la evolución temporal de los conceptos a lo largo del proceso de razonamiento. Este método permite evaluar la importancia relativa de las variables a través del tiempo, facilitando una interpretación más precisa y detallada del comportamiento del modelo.

El objetivo central de esta investigación es diseñar, implementar y validar empíricamente este enfoque, evaluando su capacidad para generar explicaciones consistentes, interpretables y útiles en contextos de clasificación basados en MCDs. Además, se propone comparar el método desarrollado con técnicas de explicabilidad ampliamente utilizadas en la literatura, como SHAP y Feature Permutation (FP), así como evaluar su calidad usando medidas derivadas de la teoría de grafos aplicadas a MCDs.

Con el fin de evaluar la validez y aplicabilidad del método, se han empleado cuatro conjuntos de datos reales: diagnóstico clínico de dengue, diagnóstico de COronaVIrus Disease 2019 (COVID-19), diagnóstico de pacientes con diabetes, y detección de fallos en vehículos submarinos autónomos. Estos casos permiten comprobar la robuztez del enfoque, su capacidad para generar conocimiento útil, y su aplicabilidad en escenarios donde la explicabilidad no es un complemento, sino una necesidad funcional y ética.

Para el desarrollo de este trabajo, se ha seguido una metodología basada en el estándar Cross Industry Standard Process for Data Mining (CRISP-DM), adaptado al contexto de la IAE. Los resultados obtenidos muestran que el método propuesto permite no solo identificar con mayor precisión las variables que más influyen en la salida del modelo para una instancia dada, sino también trazar una trayectoria causal coherente a lo largo de las iteraciones, la cual refleja fielmente el comportamiento dinámico del sistema. Además, las explicaciones generadas han demostrado cumplir con las propiedades esperadas en términos de calidad, robustez y eficiencia en todo método de explicabilidad, lo que respalda su utilidad práctica y su adecuación para tareas de clasificación basadas en MCDs.

Esta tesis, por tanto, propone un nuevo método de explicabilidad, el cual es una contribución completamente original en el campo de la IAE, al plantear un enfoque dinámico y local específicamente diseñado para MCDs, que no tiene precedentes en la literatura actual. A diferencia de los métodos existentes, que se centran en explicaciones estáticas, el enfoque presentado analiza la evolución temporal del modelo, permitiendo identificar no solo qué conceptos influyen en su salida, sino también, cómo varía su influencia a lo largo del proceso iterativo de razonamiento.

Este método ha sido desarrollado íntegramente en el marco de esta investigación, y representa una innovación metodológica sustancial, al integrar propiedades estructurales, dinámicas y causales en la generación de explicaciones. Hasta la fecha, no se ha reportado en la literatura un enfoque con las características definidas en esta investigación aplicado a MCDs, lo que sitúa esta propuesta como un

avance pionero con alto potencial de impacto. Además, este enfoque abre nuevas líneas de investigación en torno a la integración de explicabilidad dinámica en MCDs, con aplicaciones en contextos donde la transparencia, la comprensión y la justificación del modelo son elementos indispensables para su aceptación y uso responsable.

1.2 Objetivos

El presente trabajo tiene como propósito desarrollar un método de explicabilidad dinámico para modelos de clasificación basados en MCDs, centrado en la evolución de la relación causal entre los conceptos a lo largo del proceso de inferencia. Este enfoque propone abordar las limitaciones de los métodos existentes en la literatura, que a menudo ignoran la naturaleza dinámica de los MCD. Los objetivos específicos son:

1. **Analizar las relaciones causales entre las variables:** Desarrollar un sistema que permita observar y estudiar cómo evolucionan las relaciones causales entre conceptos durante el proceso de inferencia ante entradas específicas. Este sistema debe evaluar la importancia relativa de las variables a lo largo del tiempo a partir de alli, proporcionando una comprensión más profunda de la dinámica del modelo y de las interacciones entre conceptos.

2. **Desarrollar un método de explicabilidad local para MCD:** Proponer un método que, a partir de una instancia dada como entrada a un modelo de clasificación basado en un MCD, sea capaz de identificar con precisión las características más relevantes que influyen en su resultado. Este método debe ofrecer explicaciones locales, centradas en instancias particulares, facilitando una interpretación clara de cómo se llega a un especifico resultado.

3. **Observar el comportamiento global del modelo a partir de explicaciones locales:** Aplicar el método de explicabilidad a un conjunto de instancias para obtener una visión general del comportamiento del modelo. Aunque no se trata de una explicación global del modelo, este análisis permite describir cómo ciertas variables afectan de forma recurrente las decisiones, revelando patrones de comportamiento consistentes a lo largo del conjunto de datos.

4. **Comparación del método propuesto con clasicos métodos de explicabilidad:** Evaluar el desempeño del método propuesto comparándolo con otros métodos, particularmente *poshoc* ampliamente utilizados en la literatura para la generación de explicaciones. Esta evaluación permitirá identificar las fortalezas, debilidades y diferencias del enfoque desarrollado respecto a alternativas existentes.

5. **Evaluación de la calidad de las explicaciones generadas:** Verificar la calidad de las explicaciones locales producidas por el método propuesto mediante métricas adecuadas. Se busca determinar su utilidad, precisión e interpretabilidad, así como su impacto en la confianza y comprensión por parte de los usuarios.

1.3 Contribuciones

A continuación, se presentan las principales contribuciones de este trabajo:

- Se propone un método de explicabilidad dinámica para modelos de clasificación basados en MCD, que permite analizar el comportamiento temporal del sistema durante el proceso de inferencia, y extraer las variables más relevantes asociadas a una predicción específica.

- Se introduce un enfoque de explicabilidad local centrado en instancias particulares, que facilita la interpretación de las decisiones individuales del modelo a partir del método desarrollado.
- Se realiza una evaluación cuantitativa para compararlo con métodos de explicabilidad existentes, evidenciando mejoras en fidelidad, interpretabilidad y utilidad de las explicaciones generadas.
- Se valida el método propuesto en diversos dominios del mundo real, demostrando su aplicabilidad, utilidad y robustez en escenarios prácticos.

1.4 Metodología

Para la realización de este trabajo, se adopta la metodología CRISP-DM [14], la cual estructura el ciclo de vida de un proyecto de análisis de datos. Dado que el enfoque de este estudio se centra en el desarrollo de un método de explicabilidad, se propone una adaptación de la estructura tradicional de CRISP-DM para ajustarse a los objetivos del proyecto. A continuación, se describen brevemente las fases de este proyecto, las cuales serán explicadas con mayor detalle en la sección 3.

1. **Comprensión del negocio:** En esta fase se definen los objetivos del proyecto desde una perspectiva aplicada, identificando los casos de uso pertinentes de los MCD. A su vez, se analiza cómo la interpretabilidad dinámica puede abordar necesidades específicas en sectores como la salud, las finanzas, el ámbito legal, la seguridad y las ciencias sociales.
2. **Desarrollo del método de explicabilidad para MCD:** Se diseña y construye un método de explicabilidad local enfocado en el comportamiento dinámico de los MCD. Este método permitirá identificar variables claves a partir del comportamiento de las relaciones causales a lo largo del tiempo, generando explicaciones comprensibles para instancias individuales.
3. **Entendimiento de los datos:** Esta etapa comprende la recopilación, exploración y análisis de los conjunto de datos a ser usados durante la experimentacion con nuestro método de explicabilidad. Se evalúa la calidad de los datos, entre otras cosas.
4. **Preparación de los datos:** Incluye el procesamiento necesario para limpiar, transformar y seleccionar las variables. También se analizan las relaciones causales entre las variables, que se utilizarán en la construcción de los modelos MCD.
5. **Modelado:** Se desarrollan los modelos de clasificación basados en MCD. Además, se entrenan otros modelos de clasificación usando otras técnicas de aprendizaje automático. En el caso concreto de los MCD, en esta fase se modelan las relaciones entre conceptos y se calibran los pesos difusos para capturar adecuadamente la evolución de las interacciones entre variables.
6. **Evaluación:** Se lleva a cabo la evaluación del método propuesto, comparando su rendimiento explicativo con otros métodos de explicabilidad de tipo *pos- hoc* de la literatura. Asimismo, se analiza la calidad, utilidad e interpretabilidad de las explicaciones generadas, así como su impacto en la confianza de los usuarios en el modelo.
7. **Implantación:** Finalmente, se formalizan los resultados obtenidos mediante la elaboración de informes técnicos y documentos científicos. Esta fase incluye la difusión de los hallazgos y posibles aplicaciones prácticas del método de explicabilidad desarrollado en distintos contextos.

Capítulo 2

Estudio teórico

En este capítulo se presenta un análisis detallado del marco teórico que sustenta el presente trabajo. Se inicia con una revisión del estado del arte, donde se examinan los avances y enfoques más relevantes en los temas relacionados a la tesis. Posteriormente, se profundiza en los MCDs, abordando su fundamento, aplicaciones recientes, y las técnicas y variantes más destacadas. A continuación, se introduce el concepto de IAE, incluyendo sus diversas taxonomías, las propiedades que debe cumplir un método de explicabilidad, junto con una introducción a la explicabilidad causal, su necesidad y los avances logrados en esta área. Luego, se analiza el uso de los MCDs en el contexto de la explicabilidad.

2.1 Estado del Arte

El campo de la IAE ha experimentado un crecimiento significativo en su adopción a lo largo de múltiples sectores [15]. Esta expansión se refleja en la distribución sectorial de sus aplicaciones, presentada en la Figura 2.1, donde se evidencia que el desarrollo y la implementación de técnicas de IAE se han extendido a diversos dominios. Entre ellos, el sector médico destaca por concentrar el 24 % de los casos reportados, constituyendo la mayor proporción de aplicaciones documentadas hasta la fecha [16].

Este crecimiento ha impulsado tanto la creación de frameworks robustos y versátiles, que facilitan la integración de capacidades explicativas en sistemas de IA, como el desarrollo de métodos novedosos orientados a mejorar la transparencia, interpretabilidad y explicabilidad de modelos complejos. Asimismo, la comunidad investigadora ha mostrado un interés creciente en la incorporación de principios causales para fortalecer la fundamentación teórica y práctica de la explicabilidad. Las siguientes subsecciones abordan estos avances recientes, organizados en tres ejes: frameworks, métodos explicativos e integración de causalidad.

2.1.1 Frameworks y Herramientas Recientes

La industria y la academia han promovido el desarrollo de herramientas orientadas a la IAE, incluyendo frameworks capaces de integrar capacidades explicativas en sistemas complejos. Por ejemplo, Wang y otros [17] propusieron un framework modular basado en microservicios y APIs abiertas que permite generar explicaciones configurables y reproducibles a lo largo del ciclo de vida de un modelo de Aprendizaje Automático (AA). Otro enfoque híbrido propuesto en [18] combina imágenes médicas con datos tabulares para detectar cáncer de mama, incorporando mecanismos interpretables que identifican variables clínicas clave.

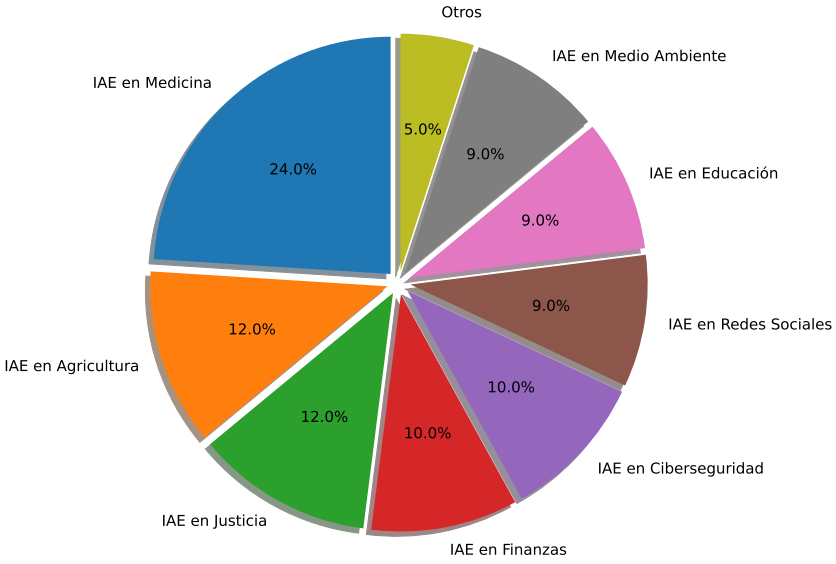

Figura 2.1: Distribución porcentual de las aplicaciones de IAE según el dominio.

En el ámbito de los sistemas IoT, Gummadi y otros [19] propusieron una solución orientada a la detección de anomalías que integra técnicas de AA con siete métodos explicativos para evaluar la relevancia de distintas características, resultando especialmente útil en la monitorización de sensores industriales y la identificación de ataques tipo botnet. Finalmente, la literatura reciente reporta diversos frameworks adicionales [20]-[24], lo que evidencia un interés creciente por desarrollar soluciones explicativas adaptadas a distintas aplicaciones y dominios.

2.1.2 Métodos Recientes en Explicabilidad

Paralelamente al desarrollo de herramientas, se observa una intensa actividad de investigación orientada a la creación de nuevos métodos, o a la mejora de los métodos existentes de explicabilidad.

Entre los avances recientes en métodos de explicabilidad se incluyen mejoras en técnicas de atribución, refinamientos de algoritmos conceptuales, eliminación de artefactos en explicaciones sintéticas, y el desarrollo de técnicas que garantizan explicaciones más robustas y fiables frente a variaciones en datos o modelos [25]. Destacan, por ejemplo, propuestas como un método basado en SHAP aplicado a análisis de grafos que integra correlaciones mediante grafos no dirigidos, superando en precisión y eficiencia a enfoques tradicionales [26]. De forma similar, se ha desarrollado una estrategia iterativa para la selección explicable de características, especialmente útil en conjuntos de datos pequeños y de alta dimensionalidad [27]. Asimismo, los métodos contrafactuales han ampliado su campo de aplicación a nuevos dominios como grafos [28] e imágenes [29], extendiendo considerablemente su utilidad.

Por otra parte, se han propuesto métodos recientes como el Análisis de Casos Cercanos (ACC) [30], que examina etiquetas con probabilidades similares en clasificación de imágenes, generando grafos y agrupaciones jerárquicas para construir conceptos interpretables y explicaciones verbales. De manera complementaria, las Explicaciones Calibradas (EC) [31] incorporan un método basado en Venn-Abers que calibra las salidas del modelo, asigna pesos confiables a las características, y cuantifica con precisión la incertidumbre.

Finalmente, el interés por adaptar los métodos explicativos a modelos emergentes, tales como los generativos o aquellos basados en aprendizaje distribuido y colaborativo, ha ido en aumento, dada la complejidad particular que presentan en términos de interpretabilidad [32]-[35].

2.1.3 Explicabilidad Causal

En los últimos años, la incorporación explícita de la causalidad en los métodos de explicabilidad ha cobrado gran relevancia, superando las limitaciones de los enfoques correlacionales tradicionales (véase 2.3.4). El objetivo es generar explicaciones que no solo describan *cómo* se llegó a una decisión, sino también *por qué*, considerando las relaciones causa-efecto subyacentes. Un elemento clave en esta integración son los Modelos Causales Estructurales (SCMs), que formalizan el conocimiento causal de forma matemática, permitiendo no solo explicar decisiones basadas en relaciones causa-efecto, sino también simular intervenciones hipotéticas y prever cómo cambios en ciertas variables impactan los resultados del modelo.

Desde un punto de vista conceptual y teórico, algunos autores han desarrollado marcos que conectan la contrafactualidad de la inferencia causal con la explicabilidad en inteligencia artificial, promoviendo una convergencia entre ambas disciplinas [36]. Asimismo, se ha propuesto la redefinición de la causalidad y contrafactualidad actual como *explicaciones accionables*, proporcionando una base filosófico-metodológica robusta para el desarrollo de sistemas explicativos causales [37]. En ese contexto, también se han planteado enfoques como el de Explicabilidad Emergente (EE) [38], que integra cadenas causales directamente en el flujo de inferencia de redes neuronales, para facilitar explicaciones más estructuradas y reveladoras de las relaciones internas entre variables.

A su vez, la causalidad se ha incorporado en el diseño de modelos intrínsecamente interpretables, como árboles de decisión causales o redes neuronales con estructuras causales predefinidas, mejorando la transparencia y evitando divisiones o correlaciones espurias [39], [40].

En el plano metodológico, se han desarrollado técnicas que integran conocimiento causal en etapas concretas del proceso explicativo. Por ejemplo, se ha adaptado el cálculo de valores de Shapley para preservar dependencias causales entre características, logrando explicaciones más fieles [41], y se han extendido métodos locales como LIME con muestreos guiados por relaciones causales para mejorar la coherencia y estabilidad de las explicaciones [42]. Estas ideas se están implementando en áreas críticas, como modelos predictivos para cuidados intensivos que emplean descubrimiento causal para mejorar la interpretabilidad y la generalización [43], así como en el análisis tridimensional de imágenes médicas, donde el razonamiento contrafactual proporciona explicaciones más precisas que métodos tradicionales [44]

2.2 Introducción a los Mapas Cognitivos Difusos

Los MCDs son una técnica de inteligencia artificial utilizada para representar y analizar conocimiento en dominios caracterizados por la incertidumbre, la complejidad y la ambigüedad [6], [45]. Los MCDs permiten modelar sistemas complejos mediante una representación gráfica que combina conceptos y relaciones causales, incorporando la lógica difusa para manejar la incertidumbre e imprecisión inherente a estos sistemas [46]-[48]. Los MCDs fueron introducidos por Kosko en 1986 [6], [48], basándose en la lógica difusa definida por Lofti Zadeh en 1965 [46] y en los Mapas Cognitivos (MC) desarrollados por Axelrod en 1976 [47], [48]. Axelrod propuso los MC como una herramienta para representar el conocimiento en ciencias sociales. Kosko amplió esta formulación al permitir valores difusos tanto en los conceptos como

en las relaciones causales entre ellos. Esta ampliación otorgó a los MCDs una mayor expresividad para capturar la incertidumbre y la ambigüedad propias de muchos sistemas reales.

Los MCDs se emplean para modelar sistemas complejos debido a su facilidad de construcción e interpretación, especialmente en dominios como sistemas sociales, ecológicos o económicos, donde las relaciones causales suelen ser inciertas y difíciles de cuantificar [49]. Un MCD es un grafo dirigido donde cada vértice representa un concepto relevante del sistema (una variable, entidad, evento o condición), y cada arista dirigida indica una relación causal entre conceptos, con un peso que expresa el grado e intensidad de esa influencia [8]. La Figura 2.2 ilustra un ejemplo de un MCD simple compuesto por siete (7) conceptos y siete (7) aristas ponderadas. Estas aristas reflejan cómo un concepto influye sobre otro, constituyendo las relaciones causales. Cada concepto c_i tiene asociado un valor de activación a_i, que suele estar acotado en el intervalo $[0, 1]$. Este valor indica su nivel de importancia o estado en un instante dado. Las conexiones causales están definidas por pesos w_{ij}, los cuales pueden tomar valores en el intervalo $[-1, 1]$. Estos valores permiten modelar distintos tipos de relaciones:

- Si $w_{ij} > 0$: existe una relación de causalidad positiva; un aumento en c_i provoca un aumento en c_j, con intensidad proporcional a $|w_{ij}|$.
- Si $w_{ij} < 0$: hay una relación de causalidad negativa; un aumento en c_i produce una disminución en c_j, también proporcional a $|w_{ij}|$.
- Si $w_{ij} = 0$: no hay relación causal entre los conceptos c_i y c_j.

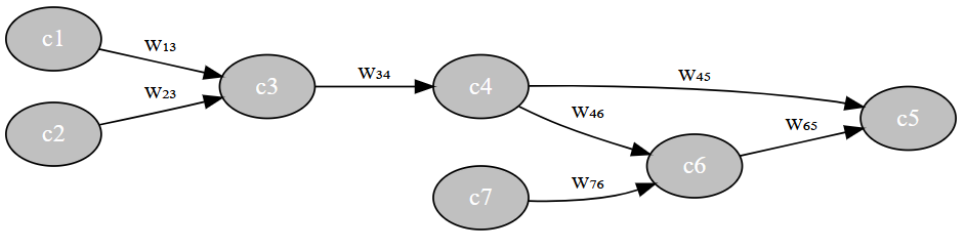

Figura 2.2: Representación gráfica de un MCD simple.

Como se ha mencionado, el valor del peso w_{ij} indica el grado de influencia entre el concepto c_i y el concepto c_j. Formalmente, un MCD se representa mediante una cuádrupla (C, W, A, f), donde:

- $C = [c_1, \ldots, c_m]$ es el conjunto de m conceptos que representan las variables o nodos del grafo que conforman el sistema. La figura 2.2 muestra un MCD con siete conceptos.
- W es la matriz de adyacencia que indica las relaciones de causalidad entre los conceptos, es decir, las aristas del grafo. A continuación se presenta la matriz de adyacencia correspondiente al MCD

mostrado en la figura 2.2:

$$W = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 & c_7 \end{array} \\ \begin{pmatrix} 0 & 0 & w_{13} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{45} & w_{46} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{65} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{76} & 0 \end{pmatrix} \end{array} \tag{2.1}$$

- $A = (a_1, \ldots, a_m)$ es el vector de activación que indica el nivel de activación o estado de cada concepto. En un instante de tiempo t, el valor a_i representa el grado de activación del concepto c_i.
- $f(\cdot)$ es la función de umbral o activación, que se utiliza para mantener los valores de activación dentro de un rango definido. La selección de esta función depende del problema específico a resolver. Las funciones más comunes en la literatura son:
  - **Bivalente**:

$$f(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{2.2}$$

  - **Sigmoidal**:

$$f(x) = \frac{1}{1 + e^{-\lambda x}} \tag{2.3}$$

  - **Trivalente**:

$$f(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \tag{2.4}$$

  - **Tangente hiperbólica**:

$$f(x) = \tanh(\lambda x) \tag{2.5}$$

La elección de la función de activación depende del tipo de análisis deseado: la función sigmoidal es útil en tareas donde se requiere suavizar la salida, mientras que la bivalente o trivalente es preferible cuando se necesita una interpretación lógica categórica.

La actualización del valor de activación del concepto c_i en el instante $t + 1$ se realiza aplicando la función de activación $f(\cdot)$ sobre la suma ponderada de las influencias recibidas desde todos los conceptos c_j que tienen una relación causal dirigida hacia c_i (es decir, desde los nodos c_j con aristas que apuntan a c_i):

$$a_i(t+1) = f\left(\sum_{j=1}^{M} w_{ji} \cdot a_j(t)\right) \tag{2.6}$$

En la ecuación 2.6, w_{ji} representa el peso de la arista que conecta el concepto c_j con c_i, indicando la intensidad y tipo de influencia que c_j ejerce sobre c_i. Este proceso de actualización se repite iterativamente hasta que el sistema alcanza una condición de estabilidad o convergencia.

La construcción de un MCD y la asignación de los pesos a las relaciones pueden realizarse con apoyo de expertos del dominio o mediante métodos de AA. Existen tres enfoques principales para el aprendizaje de la matriz de pesos W:

- **Métodos basados en Hebbian:** Aprendizaje no supervisado que ajusta pesos según si los dos conceptos se activan simultaneamente (refuerza pesos) o no (debilita pesos) [8], [50]-[52].

- **Métodos basados en Expertos:** Aprendizaje que ajusta pesos según el conocimiento experto sin necesidad de datos. Son útiles para pequeños ajustes que mantienen el significado de las relaciones causales, pero su baja flexibilidad y dependencia del experto limitan su desempeño en problemas de clasificación complejos [50], [51].
- **Métodos basados en poblaciones:** En este caso, el aprendizaje es supervisado y se emplean algoritmos de optimización para ajustar la matriz de pesos. Estos algoritmos buscan reducir la discrepancia entre las salidas esperadas y las predicciones generadas por el modelo conceptual difuso, optimizando los pesos para mejorar el rendimiento del sistema [53].
- **Métodos híbridos:** Esta estrategia combina el conocimiento experto para la inicialización del modelo conceptual difuso con un proceso de aprendizaje supervisado/ no supervisado basado en datos históricos. El objetivo es ajustar las matrices de pesos en dos etapas, partiendo de la experiencia previa y refinando el modelo con datos reales. Aunque esta aproximación resulta prometedora, la literatura sobre métodos híbridos en MCD es limitada y su aplicación práctica en problemas reales aún no está ampliamente difundida ni aceptada [54]-[56].

2.2.1 Técnicas Recientes y Variantes en Mapas Cognitivos Difusos

La investigación en MCD ha avanzado significativamente mediante mejoras algorítmicas y el desarrollo de nuevas variantes que amplían su eficacia y ámbito de aplicación. Entre estas, destacan enfoques basados en Aprendizaje Federado (AF) para entrenar modelos colaborativos sin necesidad de compartir datos sensibles, lo cual garantiza la privacidad y seguridad, especialmente en contextos médicos. Por ejemplo, se aplicaron tres esquemas de aprendizaje federado con MCD para predecir la mortalidad y prescribir tratamientos en casos de dengue severo, logrando mejoras respecto a modelos centralizados [57].

En [58], se presentó Prescriptive Fuzzy Cognitive Maps (PRV-FCM), una técnica que combina MCD con algoritmos metaheurísticos, como los genéticos, para generar modelos prescriptivos capaces de describir, y predecir el comportamiento del sistema y recomendar acciones. Esta técnica fue validada en diversos escenarios, mostrando resultados cercanos a los valores deseados para variables clave, y una alta eficacia en la toma de decisiones automatizada. Otra propuesta relevante es el modelo Fuzzy General Grey Cognitive Map (FGGCM), que incorpora la incertidumbre de datos intervalares múltiples o números difusos dentro del marco de los MCD, mejorando así el manejo de la imprecisión inherente a muchos sistemas reales.

Para entornos distribuidos, se diseñó el algoritmo Federated Fuzzy Cognitive Maps (F-FCM) [59], orientado al aprendizaje no supervisado. Este preserva la privacidad de los datos, optimiza globalmente los prototipos mediante gradientes federados y demuestra eficiencia en la construcción de estructuras globales. Finalmente, se han desarrollado variantes híbridas que integran MCD con redes neuronales profundas para modelar relaciones complejas [60], así como extensiones basadas en lógica difusa para capturar mayores niveles de incertidumbre en las relaciones causales [61]. También se han propuesto integraciones con computación cuántica, que buscan aprovechar el paralelismo cuántico para modelar sistemas caracterizados por incertidumbres [62].

2.2.2 Aplicaciones Recientes de Mapas Cognitivos Difusos

En los últimos años, los MCD se han consolidado como herramientas efectivas para analizar y resolver problemas en múltiples áreas. En Ecuador, por ejemplo, se utilizaron para identificar factores clave en el desarrollo municipal. Mediante el uso de algoritmos genéticos, se diseñaron estrategias que destacan el

papel del liderazgo, las transferencias gubernamentales y el aprovechamiento de recursos naturales [63]. En la industria del gas, un MCD construido con la participación de expertos identificó la protección catódica como el factor principal para mitigar la corrosión en ductos. Los resultados se validaron utilizando teoría de Z-números, apoyando así la gestión de riesgos [64].

La combinación de MCD y Modelado Basado en Agentes (MBA) permitió simular el impacto del comportamiento individual en la propagación del COVID-19 en Bengaluru, India, subrayando la importancia de considerar factores conductuales en las políticas sanitarias [65]. De forma similar, se desarrolló un sistema de apoyo para el diagnóstico del dengue, el cual clasifica su severidad con una precisión del 89.4 %, facilitando la evaluación de variables clínicas [49].

En el ámbito de la sostenibilidad, los MCD se emplearon para evaluar la influencia de la economía circular en las cadenas de suministro, superando la subjetividad y apoyando la toma de decisiones estratégicas [66]. En Turquía, mediante minería de texto y mapeo cognitivo difuso, se priorizaron acciones para mejorar la gestión de residuos farmacéuticos, destacando la necesidad de sistemas confiables y conciencia social [67]. Finalmente, también en Turquía, se aplicaron MCD para analizar cómo las actitudes agrícolas afectan la inflación alimentaria. A partir de escenarios construidos con entrevistas y revisión bibliográfica, se formularon recomendaciones para la formulación de políticas públicas [68].

2.3 Explicabilidad

Esta sección introduce la explicabilidad, abordando su importancia en la IA. Se presentan las diferentes taxonomías empleadas en la literatura para clasificar los métodos de explicabilidad, facilitando su análisis y comparación. A continuacion, se describen las propiedades para evaluar la calidad y utilidad de los métodos en Inteligencia Artificial Explicable. Posteriormente, se introduce la explicabilidad causal, destacando su relevancia frente a otros enfoques tradicionales al permitir un entendimiento más profundo de las relaciones causales en los modelos. Finalmente, la sección se enfoca en la explicabilidad específica para MCD, resaltando técnicas basadas en teoría de grafos y análisis dinámico que contribuyen a interpretar estos modelos. .

2.3.1 Introducción

En las últimas décadas, la IA ha transformado profundamente la manera en que interactuamos con los sistemas tecnológicos. Desde vehículos autónomos hasta aplicaciones predictivas en salud, seguridad o justicia, los sistemas inteligentes se están implementando en escenarios de gran impacto social [69], [70]. No obstante, esta expansión plantea un problema crítico: la creciente complejidad de los modelos hace que sus decisiones sean cada vez más opacas o difíciles de interpretar [1], [71]. A medida que estos sistemas se vuelven más autónomos, comprender cómo funcionan deja de ser una opción técnica para convertirse en un imperativo ético, legal y funcional [71], [72].

La explicabilidad es fundamental para promover un uso confiable y responsable de la IA, especialmente en ámbitos sensibles como la salud, la justicia o las finanzas, donde decisiones automatizadas pueden tener consecuencias directas e irreversibles [2], [3]. Cuando los usuarios no comprenden las decisiones de un sistema, es probable que disminuya su confianza, se generen malentendidos y aumente el rechazo, incluso si el modelo tiene un alto rendimiento técnico [73]-[75]. Además, la falta de explicabilidad dificulta tareas clave como la validación, la auditoría o la detección de sesgos, lo que compromete la equidad y justicia en las decisiones automatizadas [76]. Existen numerosos ejemplos documentados de riesgos asociados a modelos opacos, como sistemas de reconocimiento facial con tasas de error significativamente mayores en

personas de piel oscura [77], o algoritmos de crédito que perjudican a minorías raciales [78]-[80]. Estos problemas suelen tener origen en datos de entrenamiento que reflejan desigualdades sociales históricas, las cuales los modelos pueden perpetuar o amplificar. La explicabilidad también se alinea con principios del diseño centrado en el ser humano, al favorecer la comprensión, previsibilidad y control sobre los sistemas inteligentes. Ofrecer explicaciones claras empodera a los usuarios, mejora la supervisión y fortalece su confianza frente a decisiones automatizadas que afectan sus vidas.

Un reto particular lo plantean los modelos generativos, como los grandes modelos de lenguaje. Su naturaleza probabilística y sensible al contexto dificulta trazar con precisión el razonamiento detrás de sus respuestas, lo que introduce nuevos desafíos explicativos [1], [81]. Esta dificultad se suma a la conocida tensión entre precisión y comprensibilidad, conocida como el trade-off entre precisión e interpretabilidad: los modelos más precisos suelen ser complejos, mientras que los interpretables tienden a sacrificar parte de su rendimiento. Para abordar esta tensión, han surgido estrategias como los modelos intrínsecamente interpretables (e.g., árboles de decisión) y técnicas post-hoc (e.g., saliencias o explicaciones locales) [3], [4]. No obstante, estas últimas no siempre reflejan fielmente el razonamiento interno, lo que limita su confiabilidad.

A nivel regulatorio, la creciente preocupación por la opacidad ha motivado iniciativas como el Reglamento General de Protección de Datos (RGPD), que reconoce el derecho a recibir explicaciones sobre decisiones automatizadas [82], [83], o la Ley de Responsabilidad Algorítmica (LRA) y el Reglamento de Inteligencia Artificial de la Unión Europea (RIA) europeos, que exigen mecanismos de transparencia y auditoría para sistemas de alto riesgo [73], [84], [85].

Frente a estos desafíos, surge la disciplina de la IAE, que busca no solo facilitar la comprensión del funcionamiento interno de los modelos, sino también proporcionar herramientas prácticas para evaluar, auditar y mejorar su desempeño de manera transparente, y promueve la transparencia como pilar fundamental. La IAE juega un papel clave en la construcción de sistemas responsables y éticos, donde las decisiones automatizadas puedan ser verificadas y validadas, contribuyendo a la confianza social y al cumplimiento de normativas emergentes en distintos dominios.

2.3.2 Taxonomías generales de la explicabilidad en IA

Las taxonomías son sistemas de clasificación que organizan los métodos de explicabilidad en inteligencia artificial según criterios específicos, permitiendo analizar sus características, diferencias y relaciones de manera ordenada. Estas clasificaciones ayudan a entender desde qué enfoques se aborda la explicabilidad, facilitando una visión estructurada y comprensible del campo, aunque en ocasiones distintos criterios pueden solaparse o superponerse. A continuación, se presentan las taxonomías más conocidas, basadas en el estudio del estado del arte realizado en [86], que distingue entre: (i) funcional, (ii) basado en resultados, (iii) conceptual, y (iv) mixto.

2.3.2.1 Taxonomía Funcional

La taxonomía funcional clasifica los métodos de explicabilidad según el mecanismo mediante el cual extraen y procesan la información del modelo de AA. Este enfoque se focaliza en cómo los métodos acceden a los datos internos o externos del modelo para generar explicaciones, impactando directamente en la precisión y relevancia de las interpretaciones producidas. En [87] se identifican tres categorías principales:

- **Perturbaciones locales**: Estos métodos modifican levemente las entradas para medir la influencia de cada característica en la predicción de un caso específico. Un ejemplo representativo es *LIME*.

Su principal ventaja es que son agnósticos al modelo y fáciles de aplicar en distintos contextos, aunque pueden ser sensibles a la elección de la vecindad o al ruido, lo que afecta la estabilidad de la explicación.

- **Aprovechamiento de la estructura interna**: Se basan en propiedades internas del modelo, como gradientes en redes neuronales, para determinar la importancia de las entradas. Ejemplos destacados incluyen *Grad-CAM* y *DeepLIFT*. Estos métodos ofrecen explicaciones más fieles al funcionamiento real del modelo, pero requieren acceso a su arquitectura y parámetros, limitando su aplicabilidad a modelos específicos.
- **Metaexplicaciones**: No operan directamente sobre el modelo, sino que combinan o comparan explicaciones generadas por otros métodos para obtener interpretaciones más completas. Un ejemplo es *Aggregated Local Explanation (ALE)*.

Además, Arrieta y otros [71] proponen dos categorías adicionales que amplían esta clasificación:

- **Modificación de la arquitectura**: Consiste en simplificar modelos complejos mediante cambios estructurales para mejorar su interpretabilidad. Ejemplos de ello son la creación de modelos ante-hoc más simples o el uso de *Capsule Networks*. Esta estrategia prioriza la transparencia desde el diseño, pero puede implicar una reducción en el rendimiento predictivo.
- **Extracción de ejemplos**: Estos métodos explican el comportamiento del modelo mediante la presentación de ejemplos representativos o contraejemplos. Ejemplos conocidos son las *Prototype Explanations* y las *Counterfactual Explanations (CE)*. Son especialmente útiles para usuarios finales al facilitar la comprensión intuitiva, aunque no siempre capturan completamente la lógica interna del modelo.

La Figura 2.3 presenta la clasificación de los métodos de explicabilidad basada en las categorías comentadas en esta subsección.

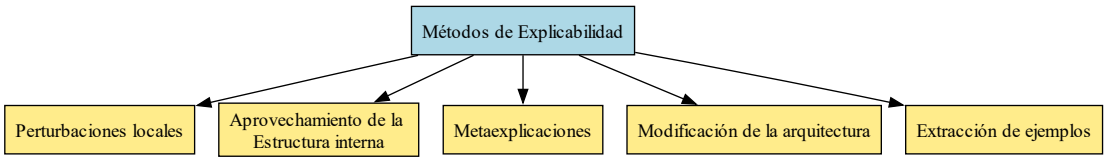

Figura 2.3: Taxonomía funcional de los métodos de explicabilidad.

2.3.2.2 Taxonomía Basada en Resultados

Este enfoque clasifica los métodos de explicabilidad según el tipo de resultado o salida que generan para el usuario. Tal categorización permite entender mejor cómo se presenta la información explicativa y qué tipo de comprensión facilita. Según [88], se distinguen tres categorías principales:

- **Importancia de características**: Estos métodos asignan un valor cuantitativo a cada característica de entrada, indicando su relevancia en la predicción realizada por el modelo. Son útiles para destacar cuáles variables influyen más en una decisión específica, facilitando la identificación de patrones y posibles sesgos. Ejemplos representativos incluyen *SHAP* y *Permutation Importance*.

Sin embargo, la interpretación de estos valores puede resultar compleja para usuarios no expertos, y la importancia asignada puede variar según el contexto o la instancia evaluada.

- **Modelos sustitutos**: Consisten en construir modelos interpretables y simples que approximan el comportamiento de un modelo complejo, permitiendo así un entendimiento global o local de sus decisiones. Los modelos sustitutos pueden ser árboles de decisión, reglas o regresiones lineales que imitan la salida del modelo original, como es el caso de *LIME*. Su principal ventaja es ofrecer explicaciones más accesibles, aunque su fidelidad puede ser limitada, especialmente en casos donde el modelo original es altamente no lineal o complejo.

- **Basada en Ejemplos**: Se basan en la presentación de ejemplos concretos, representativos o contraejemplos, para ilustrar y justificar las predicciones del modelo. Este enfoque es intuitivo, ya que se apoya en casos reales o hipotéticos para mostrar cómo pequeñas modificaciones pueden alterar la decisión. Los métodos más conocidos dentro de esta categoría incluyen las *CE*, que presentan escenarios alternativos que habrían cambiado el resultado, y técnicas basadas en *k-nearest neighbors*. La limitación principal radica en la selección y calidad de los ejemplos, que pueden no ser siempre representativos o suficientemente explicativos para todos los usuarios.

La Figura 2.4 muestra la clasificación de los métodos de explicabilidad de la taxonomía basada en los resultados, según las categorías explicadas en esta subsección.

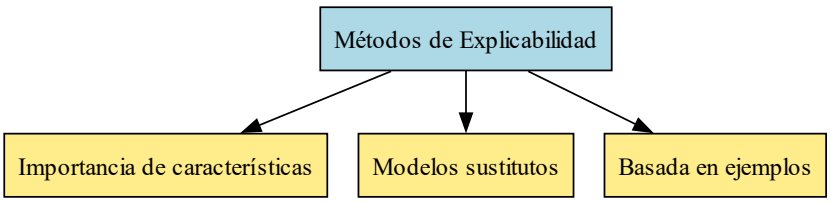

Figura 2.4: Taxonomía basada en resultados de los métodos de explicabilidad

2.3.2.3 Taxonomía conceptual

La taxonomía conceptual clasifica los métodos de explicabilidad tomando como base diferentes criterios o dimensiones teóricas que describen sus características fundamentales. Estas dimensiones sirven para examinar los métodos desde distintos ángulos, lo que ayuda a entender mejor cómo y por qué funcionan, así como a compararlos de forma más organizada y completa. Las principales categorías propuestas son:

- **Según la Etapa**: Distingue entre métodos *ante-hoc*, que son interpretables por diseño (como árboles de decisión o regresiones lineales simples), y métodos *post-hoc*, que generan explicaciones tras el entrenamiento de modelos complejos (como *LIME* o *SHAP*). Los métodos ante-hoc facilitan una interpretación directa y sencilla, aunque a veces sacrifican precisión, mientras que los post-hoc permiten explicar modelos más complejos sin modificar su arquitectura, aunque sus explicaciones pueden ser menos fieles o confiables.

- **Según la Aplicabilidad**: Distingue entre métodos *agnósticos al modelo*, que son aplicables a cualquier tipo de modelo (por ejemplo, *SHAP*), y métodos *específicos del modelo*, diseñados para arquitecturas concretas (como *Grad-CAM* en redes convolucionales). Los primeros destacan por su

versatilidad, aunque pueden sacrificar precisión o detalle, mientras que los segundos aprovechan mejor las particularidades internas del modelo, pero su uso está limitado a ciertos tipos específicos.
.

- **Según el Alcance**: Define si la explicación se centra en una predicción específica (*local*, como *LIME*) o en el comportamiento general del modelo (*global*, como reglas extraídas de un árbol de decisión). Las explicaciones locales facilitan la interpretación de casos individuales, mientras que las globales ofrecen una visión amplia del modelo. Sin embargo, las globales pueden resultar complejas o imprecisas para ciertos casos, y las locales no reflejan el funcionamiento completo del sistema.

Además, otros autores como [89] y [90] han propuesto dimensiones adicionales para enriquecer esta taxonomía:

- **Granularidad**: Considera niveles intermedios entre las explicaciones locales y globales, como las explicaciones a nivel de *cohortes* o subgrupos de datos. Estas proporcionan un equilibrio entre detalle y generalidad, permitiendo identificar patrones específicos en subpoblaciones, facilitando análisis más precisos. No obstante, requieren una segmentación adecuada de los datos, lo cual puede resultar complejo.
- **Detalle de aplicabilidad**: Considera niveles intermedios entre métodos completamente agnósticos y específicos, incluyendo aquellos diseñados para clases particulares de modelos.
- **Formato de salida**: Clasifica los métodos según el tipo de salida que generan: *numérica* (por ejemplo, importancia de características), *reglas* (explicaciones lógicas o simbólicas), *textual* (lenguaje natural), *visual* (mapas de calor o gráficos) o formatos mixtos. Por ejemplo, *DeepLIFT* produce salidas visuales.
- **Tipo de problema**: Distingue los métodos según la tarea específica a la que se aplican, como *clasificación* (por ejemplo, árboles de decisión) o *regresión* (por ejemplo, regresión lineal interpretable), adaptando la explicabilidad a las particularidades de cada problema.

En la Figura 2.5 se presenta la taxonomía conceptual de los métodos de explicabilidad, organizada según los resultados que generan y conforme a las categorías descritas en esta subsección.

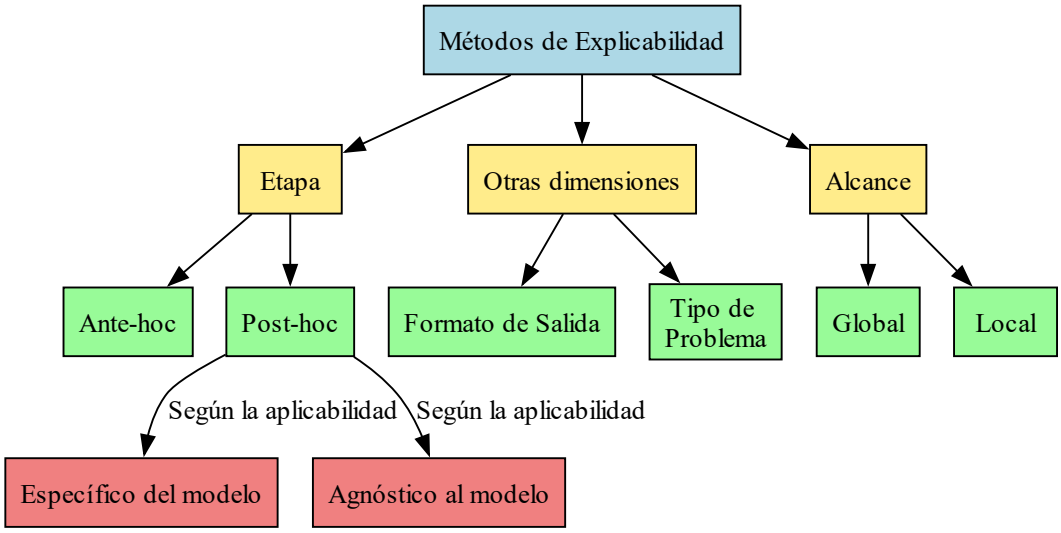

Figura 2.5: Taxonomía conceptual de los métodos de explicabilidad

2.3.2.4 Taxonomía Mixta

La taxonomía mixta surge como una propuesta integradora que combina las categorias de las taxonomías funcional, basada en resultados y conceptual, con el objetivo de ofrecer una visión más holística y estructurada de la explicabilidad en IA. Esta perspectiva reconoce que ningún enfoque, por sí solo, logra capturar toda la complejidad asociada a la interpretación de modelos, por lo que articula distintos criterios complementarios. En sus niveles superiores, esta taxonomía incorpora distinciones claves provenientes del enfoque conceptual, organizadas en torno a dos ejes fundamentales: por un lado, según la etapa en que se aplica la explicabilidad, diferenciando entre métodos *ante-hoc* y *post-hoc*; y por otro, según la aplicabilidad del método, distinguiendo entre enfoques *agnósticos* y *específicos*. Arrieta y otros [71] también identifican otras categorias dentro de los métodos de explicabilidad específicos del modelo, las cuales son explicaciones locales, visuales, basadas en la arquitectura y otras técnicas. A partir de esta integración, se distinguen cuatro categorías principales que agrupan las formas más habituales y fundamentales mediante las cuales los métodos de explicabilidad generan sus explicaciones:

- **Explicación por simplificación:** Aproxima el comportamiento de un modelo complejo mediante uno más simple e interpretable.
- **Explicación por relevancia de características:** Asigna puntuaciones o pesos a las variables de entrada en función de su influencia sobre la predicción.
- **Explicación visual:** Emplea representaciones gráficas que traducen el funcionamiento interno del modelo en elementos visuales comprensibles para el ser humano. Estas representaciones permiten identificar qué regiones, atributos o componentes de la entrada han tenido mayor influencia en la decisión del modelo, facilitando así la interpretación, especialmente en tareas donde la información es inherentemente visual o espacial.
- **Explicación local:** Se centra en explicar predicciones individuales, analizando cómo pequeñas perturbaciones en los datos o la comparación con ejemplos similares afectan la salida del modelo.

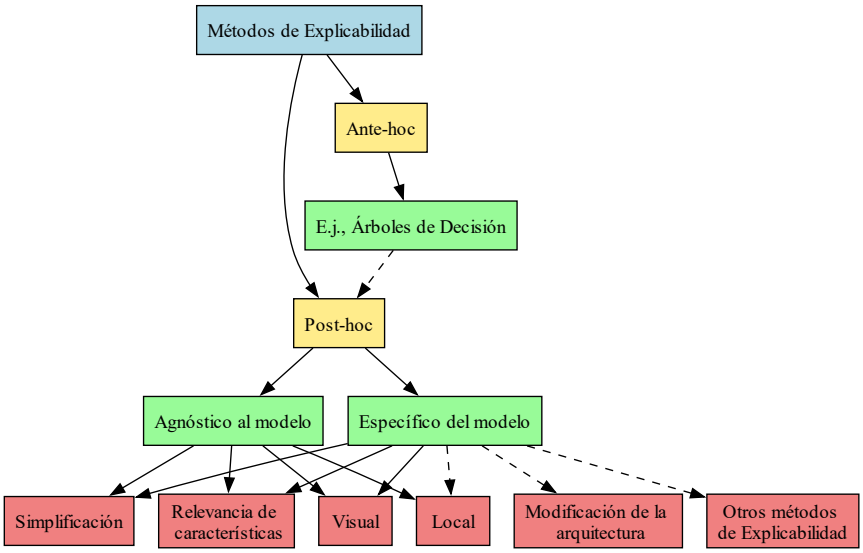

Figura 2.6: Taxonomía mixta de los métodos de explicabilidad

Aunque existen casos intermedios, la mayoría puede clasificarse en estas categorías. En la Figura 2.6 se muestra la taxonomía mixta de los métodos de explicabilidad, organizada según las categorías descritas en esta subsección. Las categorias de Arrieta y otros estan representadas mediante líneas discontinuas

2.3.3 Propiedades en Inteligencia Artificial Explicable

Para que un método de IAE sea considerado robusto en dominios sensibles como salud, finanzas, sistemas autónomos o legales, debe cumplir con un conjunto de propiedades clave que permitan evaluar la calidad, utilidad y confiabilidad de las explicaciones generadas [91]. Estas propiedades se agrupan en dos grandes enfoques: evaluación centrada en humanos y evaluación centrada en computadora. El enfoque centrado en humanos analiza cómo las explicaciones generadas por el sistema IAE satisfacen las necesidades cognitivas y prácticas de los usuarios. En contraste, el enfoque centrado en computadora emplea métricas objetivas y cuantificables, independientes del juicio humano.

En este trabajo se utiliza principalmente el enfoque centrado en computadora para evaluar el método de explicabilidad propuesto. A continuación, se describen las propiedades que componen este enfoque.

2.3.3.1 Fidelidad

La fidelidad mide el grado de correspondencia entre la explicación generada y el comportamiento real del modelo. Una alta fidelidad implica que la explicación refleja con precisión el razonamiento interno del modelo. Se calcula comparando las salidas del modelo ante la entrada original x y ante entradas perturbadas x'_i:

$$S = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\|Y(x_i) - Y(x'_i)\|}{\|Y(x_i)\|},$$

donde n es el número total de instancias evaluadas, x_i es la entrada original, y $x'_i \in X'$ representa una versión perturbada de la misma. Las funciones $Y(x_i)$ y $Y(x'_i)$ son las salidas del modelo para la entrada

original y perturbada, respectivamente. Un valor de S cercano a 1 indica que las perturbaciones apenas afectan la salida, evidenciando una alta fidelidad en la explicación.

2.3.3.2 Consistencia

La consistencia se refiere a la estabilidad y coherencia de las explicaciones generadas por el sistema cuando se utiliza la misma entrada en diferentes ejecuciones. Esto garantiza que el método produzca resultados similares en condiciones idénticas, fortaleciendo la confianza en el sistema. Una métrica para evaluar la consistencia es la *estabilidad*, que se cuantifica mediante la varianza entre las explicaciones obtenidas en múltiples ejecuciones con la misma entrada:

$$\sigma_{exp}^2 = \frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^2,$$

donde e_i es la explicación en la i-ésima ejecución, $\bar{e}$ es el promedio de todas las explicaciones, y N es el número de ejecuciones. Una varianza baja implica explicaciones muy similares, indicando alta estabilidad.

Además, la *uniformidad* evalúa cómo se distribuyen las relevancias entre las características de la entrada. Esta métrica determina si las relevancias están repartidas equilibradamente o concentradas en pocos atributos, lo que afecta la interpretabilidad. Se calcula como:

$$U = 1 - \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(r_n - \frac{1}{N}\right)^2},$$

donde r_n es la relevancia asignada a la n-ésima característica y N el número total de características. Un valor de U próximo a 1 indica una distribución uniforme de las relevancias, mientras que valores menores reflejan concentración desigual.

Juntas, estabilidad y uniformidad, permiten evaluar la consistencia del sistema, asegurando explicaciones coherentes, reproducibles y confiables.

2.3.3.3 Robustez

La robustez evalúa la capacidad de las explicaciones para mantenerse fiables y coherentes frente a pequeñas modificaciones en la entrada o cambios en el modelo, incluyendo actualizaciones y posibles ataques adversariales. Esto es crítico en dominios sensibles donde la inestabilidad puede conducir a desconfianza o errores. Además, considera si el método sigue funcionando adecuadamente cuando se implementa en distintas plataformas o cuando el modelo subyacente es actualizado.

Se mide comparando las explicaciones generadas para la entrada original y para versiones ligeramente perturbadas de esta:

$$R = 1 - \frac{1}{N}\sum_{i=1}^{N}\|\exp(x_i) - \exp(x_i')\|,$$

donde $\exp(x)$ es la explicación para la entrada x, x_i' es una perturbación leve de la entrada original x_i, y N es el número de perturbaciones evaluadas. Un valor de R cercano a 1 indica alta resistencia de las explicaciones ante cambios en la entrada.

2.3.3.4 Eficiencia

La eficiencia del método de evaluación se refiere a la capacidad computacional y los recursos necesarios para generar las explicaciones, así como al tiempo empleado en el proceso. Es fundamental que el método sea escalable, capaz de manejar grandes volúmenes de datos sin degradar su rendimiento ni incrementar excesivamente su costo computacional. La velocidad computacional, que indica la rapidez con que un sistema IAE genera explicaciones, se expresa mediante la fórmula:

$$C_s = \frac{1}{T \times R},$$

donde T es el tiempo requerido para generar una explicación y R los recursos computacionales utilizados (memoria, ciclos de CPU, etc.). Un valor bajo de C_s indica mayor eficiencia, reflejando un menor uso de tiempo y recursos para obtener explicaciones.

2.3.4 Explicabilidad Causal

La mayoría de los enfoques actuales de explicabilidad se basan en relaciones correlacionales entre las entradas y salidas de un modelo. Métodos como *LIME* [3], *SHAP* [4] o los *mapas de saliencia* estiman la importancia de las variables observando cómo varía la predicción ante cambios en los atributos de entrada. Aunque útiles para obtener explicaciones locales, estos métodos presentan limitaciones en entornos complejos y sensibles [1], [92]. Al estar basados en información observacional, estos enfoques no capturan relaciones causales reales, lo que restringe su capacidad para responder preguntas contrafactuales o identificar causas subyacentes [93]. Problemas como la *multicolinealidad* pueden hacer que la atribución de importancia sea ambigua, y la presencia de variables *confusoras* puede introducir asociaciones espurias [94]-[96]. Esto hace que las explicaciones basadas en correlación sean potencialmente inestables o engañosas desde una perspectiva causal.

Ante estas limitaciones, han emergido enfoques de *explicabilidad causal* que incorporan nociones de causa y efecto mediante marcos como los *SCMs*, los *grafos acíclicos dirigidos (DAGs)* y el *razonamiento contrafactual* [97], [98]. Estos métodos permiten responder preguntas del tipo "¿Qué habría pasado si la variable X hubiera tomado otro valor?", proporcionando explicaciones más robustas y accionables [99]. En contextos como la medicina, por ejemplo, este enfoque permite distinguir entre un síntoma que causa un deterioro y otro que simplemente está asociado, mejorando la toma de decisiones clínicas [96], [100]. A diferencia de los métodos tradicionales, los modelos causales pueden controlar explícitamente variables confusoras y estimar efectos directos, indirectos o colaterales [97], [101]. Esto es fundamental en dominios de alto impacto como el derecho, la medicina o las políticas públicas, donde confundir correlación con causalidad puede tener consecuencias serias [92]. Además, dado que las funciones objetivo de los modelos de AA suelen capturar correlaciones en lugar de verdaderas relaciones causales, estos pueden fallar ante cambios en la distribución de los datos o cuando enfrentan situaciones no observadas previamente [98].

En este contexto, la investigación en explicabilidad causal puede agruparse en cuatro grandes áreas: (i) análisis causal de componentes del modelo, (ii) generación de explicaciones contrafactuales, (iii) relación entre causalidad e imparcialidad, y (iv) verificación de relaciones causales a través de la interpretabilidad.

2.3.5 Explicabilidad en Mapas Cognitivos Difusos

En este contexto, los (MCDs) surgen como una herramienta particularmente relevante para abordar la explicabilidad desde una perspectiva causal. Su estructura basada en grafos dirigidos y ponderados

permite representar explícitamente relaciones causa-efecto entre conceptos, lo que los posiciona como un marco natural para el modelado causal interpretable. No obstante, al igual que los métodos analizados previamente, muchos enfoques de explicabilidad en MCDs han tendido a centrarse en representaciones estáticas o estructurales, dejando de lado la dinámica inherente del sistema, fundamental para comprender su comportamiento a lo largo del tiempo. A continuación, se exploran en detalle las principales estrategias utilizadas para dotar de explicabilidad a los MCDs, así como sus limitaciones, y el potencial de enfoques dinámicos para superar dichas barreras.

Existen numerosas investigaciones que utilizan los (MCDs) tanto para dotar de explicabilidad a sistemas de IA como para desarrollar métodos específicos de explicabilidad basados en esta técnica. Principalmente, las investigaciones orientadas a proporcionar explicabilidad a los sistemas suelen combinar técnicas de IA con modelos de MCDs para ofrecer interpretaciones del funcionamiento del modelo [102], [103]. Sin embargo, este enfoque presenta limitaciones. En muchos casos, solo se considera la imagen final del modelo, sin tener en cuenta las propiedades dinámicas y los estados ocultos que emergen de la interacción entre las condiciones iniciales, la matriz de pesos y la función de activación. Estas dinámicas internas son fundamentales para una comprensión profunda y explicativa. Permiten capturar el comportamiento temporal y la evolución del sistema, aspectos que un análisis estático no puede revelar.

En [11] se identifican dos estrategias de análisis estructural comúnmente aplicadas en los MCDs para determinar la relevancia de los conceptos: (i) *medidas de centralidad basadas en teoría de grafos* y (ii) *reducción de la red de conceptos*. Aunque estas técnicas no son métodos de explicabilidad en sentido estricto, contribuyen a la interpretabilidad al identificar los nodos más influyentes del sistema. Una revisión reciente de la literatura amplía esta clasificación e incorpora un enfoque explícito orientado a la explicabilidad en MCDs, que se organiza en una tercera categoría: (iii) *análisis de la dinámica del sistema*. A continuación, se describen estas tres categorías, comenzando por las métricas estructurales.

2.3.5.1 Medidas de Centralidad en Teoría de Grafos

Una forma directa de obtener explicaciones en MCDs consiste en analizar su estructura estática mediante técnicas de teoría de grafos. En este enfoque, el modelo se representa como un grafo dirigido y ponderado, donde los nodos son conceptos y las aristas indican relaciones causales con pesos asociados. Las métricas de centralidad identifican los conceptos más influyentes según su conectividad estructural.

Estas métricas se aplican sobre la representación del grafo resultante tras el proceso de inferencia. De este modo, proporcionan una caracterización estática que, aunque no refleja el comportamiento dinámico, resulta útil para interpretar la importancia relativa de los conceptos. Entre las métricas más usadas en la literatura [12], [13], [104] se encuentran:

- **Grado de entrada** $d_{in}(v)$: en un grafo dirigido ponderado, el grado de entrada de un nodo se define como la suma de los valores absolutos de los pesos de las aristas que llegan a dicho nodo.

$$d_{in}(v) = \sum_{u \in V} |w_{uv}| \tag{2.7}$$

  donde w_{uv} es el peso de la arista desde el nodo u hacia v.

- **Grado de salida** $d_{out}(v)$: en un grafo dirigido ponderado, el grado de salida de un nodo es la suma de los valores absolutos de los pesos de las aristas que salen del nodo hacia otros nodos.

$$d_{out}(v) = \sum_{u \in V} |w_{vu}| \tag{2.8}$$

- **Grado total** $d(v)$: suma del grado de entrada y de salida, que da una medida global de la conectividad del nodo en la red:

$$d(v) = d_{in}(v) + d_{out}(v) \tag{2.9}$$

- **Intermediación** $B(v)$: mide la importancia de un nodo como intermediario en la transmisión de información dentro del grafo. Se calcula como la proporción de caminos más cortos entre pares de nodos que pasan por el nodo v:

$$B(v) = \sum_{\substack{s,t \in V \\ s \neq v \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.10}$$

  donde σ_{st} es el número total de caminos más cortos entre los nodos s y t, y $\sigma_{st}(v)$ es la cantidad de esos caminos que atraviesan v. Un valor alto indica que el nodo actúa como un puente clave en la red.

- **PageRank** $PR(v)$: mide la importancia de un nodo no solo por el número de conexiones entrantes, sino también considerando la importancia de los nodos que lo enlazan. Así, un nodo conectado a otros nodos importantes recibe una puntuación mayor.

$$PR(v) = \frac{1-d}{N} + d \sum_{u \in In(v)} \frac{PR(u)}{d_{out}(u)} \tag{2.11}$$

  donde d es el factor de amortiguamiento (generalmente 0.85), N es el total de nodos, $In(v)$ es el conjunto de nodos con aristas que apuntan a v, y $d_{out}(u)$ es el grado de salida del nodo u. Esta métrica refleja la importancia global de un nodo dentro de la red, considerando la calidad y cantidad de sus conexiones entrantes.

Estas métricas aportan información sobre la estructura y relevancia relativa de los conceptos dentro del grafo. Sin embargo, ninguna considera los valores de activación de los nodos ni la función de activación usada en el proceso de inferencia de los MCD. Por ello, estas medidas no capturan el comportamiento dinámico del sistema, que es una característica fundamental y distintiva de los modelos basados en MCD.

2.3.5.2 Reducción de la Red de Conceptos

Otra forma común de mejorar la interpretabilidad de MCDs es simplificar la estructura del modelo mediante técnicas de reducción de la red. Estos métodos eliminan conceptos redundantes, fusionan nodos con comportamientos similares, y conservan solo los conceptos más relevantes. El objetivo es obtener versiones compactas y manejables del modelo que mantengan su capacidad de representación y predicción, sin afectar la precisión ni la coherencia del sistema.

En este contexto, se desarrollaron diversos enfoques para equilibrar complejidad y precisión. Por ejemplo, un método reduce el número de conceptos agrupando aquellos similares en clústeres, y luego optimiza los parámetros de transformación para conservar el comportamiento dinámico del sistema [105]. Otro enfoque usa técnicas de agrupamiento como K-Means y Fuzzy C-Means para simplificar modelos complejos y simples. Este método logra que el modelo reducido mantenga un comportamiento fiel al original y supere en fidelidad a métodos previos [106]. Además, se propone una reducción basada en relaciones de tolerancia difusas que facilita modelos más transparentes y accesibles para los responsables de la toma de decisiones [107].

Aunque estas técnicas ayudan a manejar la complejidad y aumentan la interpretabilidad, no bastan para explicar completamente el modelo. La reducción puede conllevar pérdida de información importante

y dificultar la comprensión profunda de las relaciones causales y dinámicas entre conceptos. Además, la simplificación estructural no siempre refleja las sutilezas del razonamiento humano. Tampoco captura adecuadamente las condiciones contextuales que influyen en la toma de decisiones. Por ello, se necesitan enfoques complementarios que conserven tanto el significado semántico como el dinámico del sistema original.

2.3.5.3 Dinámica

La dinámica analiza cómo evoluciona el sistema representado por el MCD a lo largo del tiempo. En lugar de limitarse a la estructura estática del grafo, considera el comportamiento temporal de las activaciones y las influencias causales entre conceptos durante el proceso de inferencia. Por ejemplo, Tyrovola y otros [108] desarrollan un enfoque dinámico basado en teoría de grafos que calcula de forma eficiente el efecto causal total entre conceptos en MCDs. Este método no solo evalúa la estructura estática, sino que también considera la propagación acumulativa de influencias a través de múltiples caminos causales y pasos temporales. De este modo, captura la dinámica de interacción entre conceptos.

Por otro lado, Napoles y otros [109] presentan un método basado en valores SHAP (SHapley Additive exPlanations) que calcula la atribución de conceptos usando como entradas los valores iniciales de activación y como salidas los estados ocultos generados durante el razonamiento recurrente. Finalmente, un enfoque basado en el análisis de flujo de información identifica automáticamente relaciones causales verdaderas a partir de datos, y las impone como restricciones en el aprendizaje del modelo. Esta técnica evita la captura de correlaciones espurias y mejora la precisión, interpretabilidad y capacidad dinámica del MCD [110].

Dado que las técnicas actuales para la explicabilidad en MCDs suelen centrarse en aspectos estructurales o simplificaciones estáticas, y considerando la complejidad inherente de los comportamientos dinámicos y estados ocultos que surgen durante la inferencia, se hace evidente la necesidad de enfoques que integren explícitamente la dinámica del sistema para mejorar la interpretabilidad. En este sentido, **el presente trabajo propone un método de explicabilidad dinámico para MCDs que permite capturar de forma transparente las relaciones causales entre los conceptos y su evolución temporal a lo largo del proceso de razonamiento**, superando así las limitaciones de los métodos existentes al posibilitar una comprensión más profunda y precisa del modelo.

La propuesta destaca por incorporar de manera continua los valores de activación de los nodos durante todo el proceso de inferencia, junto con las relaciones causales entre conceptos. Esta integración permite representar con mayor exactitud la evolución interna del sistema y genera explicaciones más detalladas y contextuales, que reflejan tanto la interacción temporal como las influencias causales entre los conceptos, a diferencia de enfoques que se restringen a análisis estáticos o a estados finales del modelo.

.

Capítulo 3

Metodología

Para la realización de este trabajo, se adopta la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) [14], un marco estructurado ampliamente utilizado que guía el ciclo de vida de proyectos de análisis y minería de datos. Esta metodología establece un conjunto claro de fases, que van desde la comprensión del negocio hasta la implantación de resultados, facilitando así un desarrollo ordenado, sistemático y replicable. Dado que el enfoque de este estudio está orientado al desarrollo de un método de explicabilidad para modelos de clasificación basados en MCDs, se realizó una adaptación de la estructura tradicional de CRISP-DM para ajustarla a las particularidades y objetivos específicos del proyecto. A continuación, se describen las fases que conforman la metodología y las acciones llevadas a cabo en cada una de ellas.

3.1 Comprensión del negocio

La fase inicial de la metodología CRISP-DM tiene como objetivo entender los requerimientos del proyecto desde una perspectiva de negocio, con el fin de definir los objetivos del análisis y transformar ese conocimiento en un plan técnico. En el contexto de este trabajo, esto implicó comprender las necesidades en torno a la explicabilidad de modelos de IA basados en MCDs, y el valor que aportaría el desarrollo de un nuevo enfoque en este ámbito.

Se realizo un estudio teórico exhaustivo, presentado en detalle en la sección 2, que abarcó desde la revisión del estado del arte de los métodos de explicabilidad existentes, hasta detallar las métricas y propiedades de robustez que deben cumplir dichos métodos. Durante este proceso, se identificaron tanto los enfoques más usados como los desarrollos recientes, además de las taxonomías empleadas para su clasificación. Paralelamente, se revisaron mejoras y adaptaciones de métodos previos, orientadas a corregir deficiencias o a ajustarlas para uso en ámbitos poco explorados.

También se profundizó en la teoría que sustenta los MCDs, analizando su funcionamiento, orígenes, avances y principales líneas de investigación. Dado que estos modelos se fundamentan en principios de causalidad, se abordó igualmente el estudio de la explicabilidad causal en IA, analizando los métodos existentes aplicados a modelos basados en MCDs. Este análisis permitió confirmar la originalidad del método propuesto, al no encontrarse alternativas similares en la literatura.

3.2 Desarrollo del método de explicabilidad para MCD

Una vez finalizado el estudio teórico, se inició el diseño y desarrollo del método de explicabilidad propuesto, con el objetivo principal de construir un enfoque capaz de generar explicaciones que reflejen el comportamiento dinámico subyacente de los MCDs. Para ello, se partió de una idea conceptual basada en las propiedades estructurales y temporales características de los MCDs, que se fue consolidando a través de un análisis técnico exhaustivo, en el que se evaluaron aspectos como su viabilidad computacional, las posibles limitaciones, y los requisitos necesarios para asegurar una correcta aplicación del método.

A partir de este análisis, se definió formalmente el método: se estableció su modelo matemático, los algoritmos requeridos para su funcionamiento, y las condiciones bajo las cuales resulta aplicable de manera efectiva. Este diseño consideró tanto la interpretabilidad como la coherencia causal de las explicaciones generadas, asegurando que el enfoque fuera tanto claro como técnicamente sólido.

Finalmente, se implementó el método y se integró en modelos de clasificación construidos a partir de MCDs, lo que permitió su validación en diferentes contextos experimentales. El desarrollo completo y detallado del método propuesto, incluyendo sus fundamentos, algoritmos y ejemplos de aplicación, se presenta en la sección 4. Es importante resaltar que este enfoque novedoso incorpora explícitamente la dinámica del modelo en el proceso explicativo, constituyendo una contribución original al campo de la IAE.

3.3 Entendimiento de los datos

Esta etapa comprendió la recopilación, exploración y análisis de los conjuntos de datos que serían utilizados para la construcción de los modelos de clasificación basados en MCDs (descritos en la sección 5.1.1), sobre los cuales se hará posteriormente la aplicación del método de explicabilidad propuesto en este trabajo. La selección de los conjuntos se realizó atendiendo a dos criterios fundamentales: que pertenecieran a dominios críticos donde la explicabilidad aporte un valor añadido significativo, y que hubieran sido previamente utilizados en la literatura científica, lo cual facilita tanto la comparación de resultados como la validación del enfoque propuesto. Como resultado, se seleccionaron conjuntos de datos correspondientes a casos de dengue, COVID-19, diabetes y diagnóstico de fallos en vehículos submarinos.

3.4 Preparación de los datos

Una vez seleccionados los conjuntos de datos, se realizó el preprocesamiento, que incluye tareas de limpieza, transformación y selección de las variables relevantes. Estas operaciones se describen con mayor detalle en la sección 5.1.2. Para cada conjunto de datos, se realizó un análisis exploratorio inicial para comprender la estructura general y las características principales. Posteriormente, se realizó un análisis de correlación utilizando métricas adecuadas según el tipo de variable, como el coeficiente de Pearson para variables numéricas y Cramér's V para variables categóricas, y se evaluó la colinealidad para detectar posibles problemas de multicolinealidad entre las variables explicativas.

Además del procedimiento general de preprocesamiento descrito, cada conjunto de datos presentó particularidades que requirieron tratamientos específicos. A continuación, se comentan las acciones realizadas en cada caso, que son detalladas en la sección 5.1.2:

- **Dengue:** se eliminó una variable redundante cuyo valor permanecía constante en todas las instancias, ya que no aportaba información útil al modelo.

- **COVID-19:** se construyó una nueva variable, sobre la cual se aplicó un algoritmo de detección y eliminación de instancias contradictorias, mejorando así la coherencia del conjunto. Posteriormente, se aplicaron técnicas de sobremuestreo como Synthetic Minority Over-sampling Technique (SMOTE) para balancear las clases.
- **Diabetes** y **Diagnóstico de Fallos en Vehículos Submarinos Autónomos:** dado que ambos conjuntos están compuestos exclusivamente por variables numéricas, se aplicó un mismo proceso de preprocesamiento que incluyó:
  - Análisis de la distribución de las variables mediante histogramas, utilizando la regla de Freedman-Diaconis para determinar el número de intervalos.
  - Evaluación de la normalidad mediante *Q-Q plots*.
  - Detección de valores atípicos con los métodos del rango intercuartílico (IQR) y *Z-score*.
  - Normalización de las variables usando el método *Min-Max*.
  - Balanceo de clases usando SMOTE.
  - Análisis de la significancia de las variables predictoras respecto a la variable objetivo mediante Analysis of Variance (ANOVA).

La razón de estos pasos extras en el caso de estos dos conjuntos de datos se debe a que están compuestos exclusivamente por variables numéricas, cuya distribución, diferencias de rangos de valores y presencia de valores atípicos pueden influir de forma significativa en el rendimiento de los modelos basados en técnicas de IA, haciendo necesario un análisis estadístico riguroso y una adecuada transformación de los datos.

En el conjunto de **diabetes** se identificaron valores atípicos relevantes, los cuales fueron corregidos mediante una técnica de imputación basada en vecinos más cercanos (*k-nearest neighbors*), lo que permitió mejorar la distribución de las variables. Por otro lado, en el conjunto de **diagnóstico de fallos** se detectaron posibles valores atípicos, aunque su impacto se consideró poco significativo, por lo que no se aplicaron métodos de corrección. No obstante, se construyó una nueva variable que sintetiza el comportamiento conjunto de las señales de los motores.

3.5 Modelado

El proceso incluyó la definición de la arquitectura de los modelos de MCDs, la selección de parámetros adecuados y su construcción efectiva, tal como se detalla en la sección 5.3.1. Para validar su capacidad predictiva, se evaluó el desempeño de estos modelos mediante métricas específicas (ver dichas métricas en la sección 5.2). Además, con el fin de asegurar que los modelos basados en MCDs ofreciesen un rendimiento competitivo, sus resultados se compararon con los obtenidos mediante otras técnicas clásicas de IAl, cuya construcción se describe en la sección 5.3.2.

Dado que el objetivo principal es desarrollar un nuevo método de explicabilidad para MCDs, una vez construidos los modelos de clasificación, se compararon las explicaciones generadas por el método propuesto con las obtenidas mediante otros enfoques de explicabilidad existentes. Para ello, se consideraron dos métodos clásicos de explicabilidad ampliamente usados en la literatura, concretamente SHAP y FP. De esta manera, se realizó un análisis comparativo de los distintos métodos de explicabilidad en los diferentes conjuntos de datos. A su vez, se emplearon medidas de centralidad de la teoría de grafos utilizadas en la literatura para evaluar la explicabilidad en los modelos de MCDs. Esta doble comparación permitió evaluar la calidad y consistencia de las explicaciones desde diferentes perspectivas, combinando

una visión específica y adaptada a la estructura causal de los MCDs con un marco general que facilita contrastar el nuevo enfoque con métodos consolidados de IAE. Los resultados de todo este análisis de explicabilidad se presentan en la sección 5.5.2.

3.6 Evaluación

La calidad y efectividad de los resultados del método de explicabilidad propuesto se evaluaron mediante dos enfoques complementarios que abordan tanto aspectos cuantitativos como cualitativos del desempeño explicativo. En primer lugar, se utilizó la técnica RemOve And Retrain (ROAR) para medir el impacto que tienen las variables declaradas como importantes por el método propuesto sobre el rendimiento predictivo del modelo de clasificación, comparándolo con los resultados obtenidos mediante los métodos SHAP y FP. El desarrollo y los resultados de este análisis cuantitativo se presentan con detalle en la sección 5.5.3.

En segundo lugar, se realizó una evaluación basada en el cumplimiento de un conjunto de propiedades deseables en los métodos de explicabilidad, tales como robustez, fidelidad, consistencia y utilidad para el usuario final. Estas propiedades, definidas previamente en la sección 2.3.3, ofrecen una valoración más teórica y cualitativa del método, complementando así los resultados cuantitativos, proporcionando una visión integral sobre su comportamiento y aplicabilidad práctica.

3.7 Implantación

Esta fase no aplica directamente en nuestro trabajo. Ahora bien, los resultados obtenidos durante el desarrollo y evaluación del método fueron formalizados y difundidos mediante la elaboración de informes técnicos y la redacción de artículos científicos. Este proceso asegura que los hallazgos se presenten con rigor, facilitando tanto la validación independiente como la reproducibilidad del método por parte de otros investigadores y profesionales.

Adicionalmente, se desarrollaron materiales y recursos complementarios como código fuente documentado, conjuntos de datos procesados disponibles en repositorios, y guías de uso para facilitar el uso del método en proyectos reales y su aplicación práctica. Esta fase también incluyó la identificación de posibles limitaciones y recomendaciones para su implementación en distintos entornos, estableciendo un marco claro para futuras mejoras y adaptaciones.

Capítulo 4

Desarrollo del Método de Explicabilidad

En esta sección se presenta un método dinámico de explicabilidad para interpretar la importancia causal de los conceptos en un modelo MCD. En primer lugar, se describe el enfoque propuesto y, a continuación, se muestra un ejemplo que ilustra su aplicación. Todo lo presentado en esta sección constituye la contribución de esta investigación, completamente original. No se trata de una adaptación ni de una extensión de propuestas previas, sino la definición de un enfoque íntegramente novedoso, diseñado específicamente para capturar la evolución dinámica y causal de los conceptos en modelos basados en MCDs, para usarlos en sus análisis de explicabilidad. Según nuestro conocimiento, en la literatura científica no se ha reportado un método de explicabilidad que integre las características aquí planteadas en el contexto de los MCDs, lo que refuerza el carácter pionero y el potencial impacto de esta propuesta.

4.1 Especificación de Nuestro Enfoque

El método de explicabilidad está concebido para su aplicación en modelos de clasificación basados en MCDs. La Figura 4.1 presenta de forma esquemática las fases que componen el método de explicabilidad dinámico propuesto. Dicho enfoque permite analizar la influencia que cada concepto ejerce dentro del sistema respecto a un concepto clase.

En primer lugar, se establecen los **requisitos previos** para aplicar el método: se requiere que el modelo converja y que se disponga de la matriz de pesos junto con las activaciones dinámicas obtenidas tras el proceso de inferencia. A partir de esta base, el método se desarrolla en cuatro fases secuenciales. En la fase **(i)** se identifican los caminos causales, tanto directos como indirectos, que conectan los conceptos en el grafo del modelo, permitiendo mapear cómo fluye la influencia de los conceptos a través de la red. Luego, en la fase **(ii)** se calcula la influencia que un concepto ejerce sobre otro, considerando los pesos causales, las activaciones temporales, y una penalización que reduce progresivamente la contribución de los tramos más alejados. A continuación, en la fase **(iii)** se integran estas influencias directa e indirectas en una única medida que captura el impacto global de un concepto sobre otro. Finalmente, en la fase **(iv)** se construye un ranking de conceptos en función de su importancia relativa, lo cual permite identificar cuáles son los más influyentes en el modelo, y facilita su interpretación desde una perspectiva explicativa. A continuación, se describen en detalle cada una de estas fases.

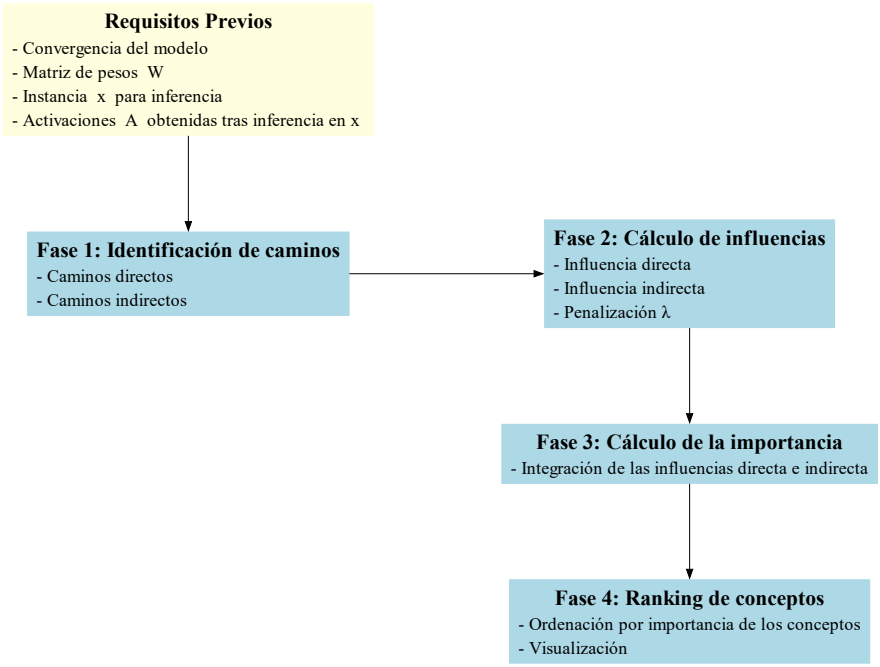

Figura 4.1: Diagrama de flujo del Método de Explicabilidad

4.1.1 Requisitos para la Aplicación del Método de Explicabilidad

Como se definió en la sección 2.2, un MCD se representa mediante una 4-tupla (C, W, A, f). Como condición fundamental para el correcto funcionamiento del método de explicabilidad, se asume que el sistema construido converge a un estado estable, es decir, que las activaciones A alcanzan un punto fijo tras iteraciones sucesivas. Esta convergencia es necesaria para garantizar la estabilidad y la interpretabilidad de los resultados obtenidos mediante el método propuesto. Además, una vez que se dispone del modelo de MCD, es necesario contar con la matriz de pesos W, la cual representa las relaciones causales directas entre los conceptos del MCD. Finalmente, para hacer el análisis de explicabilidad, dada una instancia x, al inferir en el modelo se obtiene una serie de activaciones A, donde cada $A_i^{(t)}$ representa el nivel de activación del concepto c_i en la iteración t. Esa matriz A la requiere el método de explicabilidad.

4.1.2 Fase 1: Identificación de Caminos.

Siendo c_i el concepto clase sobre el cual se desea calcular la importancia, el método utiliza caminos directos e indirectos dentro del MCD para evaluar la influencia que cada concepto ejerce en el modelo sobre c_i. Los caminos de influencia directa son aquellos que conectan directamente a un concepto c_j con el concepto de interés c_i, es decir, relaciones inmediatas en la estructura causal del modelo. Por otro lado, los caminos de influencia indirecta involucran secuencias de conceptos intermedios que conectan c_j con c_i a través de varios pasos en la red causal. Eventualmente, esto permite captar cómo conceptos más alejados en la red pueden influir en c_i.

Para definir la influencia indirecta, consideramos el conjunto $R_{j \to i}$ de todos los caminos causales indirectos que conectan el concepto c_j con el concepto c_i. Estos caminos son rutas simples, es decir, secuencias de nodos sin repeticiones, salvo en el caso especial cuando $c_i = c_j$. En este caso particular, se

permite que el nodo origen aparezca dos veces, una al principio y otra al final del camino, para poder representar ciclos simples donde un concepto se influye a sí mismo.

Dado que la cantidad total de caminos entre dos conceptos puede crecer exponencialmente en grafos grandes o densos, y para mantener el cálculo computacionalmente viable, se limita la evaluación a los x caminos simples más cortos. Estos caminos son generados mediante un algoritmo iterativo conocido como *algoritmo de caminos simples más cortos* [111], que produce los caminos en orden creciente de longitud. Primero devuelve el camino más corto; luego, de forma sucesiva, explora caminos más largos sin repetir nodos, salvo la excepción mencionada para ciclos autorreferenciales, eliminando aristas o nodos temporalmente para evitar ciclos más complejos. Por ejemplo, si tenemos un grafo con nodos A, B, C, D y queremos encontrar caminos de A a D, el algoritmo devolverá primero el camino $A \to D$ (si existe), luego $A \to B \to D$, luego $A \to C \to D$, y así sucesivamente, siempre sin repetir nodos en un mismo camino, excepto la posible segunda aparición del nodo origen en casos de ciclos autorreferenciales simples.

4.1.3 Fase 2: Cálculo de Influencias Directas e Indirectas.

La **influencia directa** de un concepto c_j sobre otro c_i se define como el promedio temporal de la influencia ejercida a través de una conexión directa en el grafo causal durante el proceso de inferencia del MCD, y se expresa mediante la siguiente ecuación:

$$I_{\text{dir}}(c_j, c_i) = \frac{1}{T} \sum_{t=1}^{T} w_{j,i} \cdot A_j^{(t)} \tag{4.1}$$

donde $w_{j,i} \in \mathbb{R}$ representa el elemento de la matriz de pesos W que indica la influencia directa (positiva o negativa) del concepto c_j sobre c_i; $A_j^{(t)} \in [-1, 1]$ corresponde al nivel de activación del concepto c_j en la iteración t; y $T \in \mathbb{N}$ representa el número máximo de iteraciones definidas durante la creación del modelo para el proceso de inferencia. Se asume que el modelo converge antes o al alcanzar T iteraciones; en caso contrario, T funciona como un límite para evitar ciclos infinitos y asegurar la finalización del cálculo.

Esta expresión representa la *influencia acumulada que el concepto c_j ejerce sobre c_i* a través de la conexión directa entre ambos conceptos. Se ponderan el peso de la relación directa ($w_{j,i}$) y la activación temporal ($A_j^{(t)}$) del concepto origen en cada iteración del proceso de inferencia. Así, se captura el impacto inmediato y puntual que un concepto tiene sobre otro en la dinámica del modelo.

La **influencia indirecta** de un concepto c_j sobre otro concepto c_i mide la influencia que ejerce c_j sobre c_i a través de caminos causales de longitud mayor a uno en el grafo del MCD. Esta influencia se calcula considerando todos los caminos causales R que conectan c_j con c_i mediante rutas indirectas. Para cada camino $r = 1, \ldots, R$, se toma en cuenta su longitud n_r, que representa el número total de nodos del camino, es decir, la cantidad de conceptos consecutivos conectados. Cada camino se define como una secuencia ordenada de conceptos:

$$(c_{p_0^{(r)}}, c_{p_1^{(r)}}, \ldots, c_{p_{n_r-1}^{(r)}}) \tag{4.2}$$

donde $c_{p_0^{(r)}} = c_j$ es el nodo origen y $c_{p_{n_r-1}^{(r)}} = c_i$ es el nodo destino.

La matriz de pesos W está definida tal que el elemento $w_{j,i}$ representa la influencia **del concepto c_j sobre el concepto c_i**; es decir, el primer subíndice indica el nodo origen y el segundo el nodo destino de la relación causal. Por lo tanto, para cada tramo $k = 0, \ldots, n_r - 2$ del camino r, el peso correspondiente es

$$w_{p_k^{(r)},p_{k+1}^{(r)}},$$

donde $p_k^{(r)}$ es el índice del nodo origen y $p_{k+1}^{(r)}$ el índice del nodo destino para ese tramo.

Con el fin de reducir progresivamente la influencia de los conceptos según su distancia al concepto destino c_i, se introduce un hiperparámetro $\lambda \in [0,1]$. Este parámetro ajusta el peso que cada tramo causal aporta a la importancia total, de modo que los tramos más cercanos al destino tienen una mayor contribución, mientras que los más alejados son penalizados y afectan menos la influencia acumulada. Así, λ actúa como un factor de decaimiento que modula la relevancia de la información transmitida a través de caminos causales de distintas longitudes:

$$f(k,r) = \begin{cases} n_r - k - 1 & \text{si } k < n_r - 1, \\ 0 & \text{si } k = n_r - 1. \end{cases} \tag{4.3}$$

De esta forma, los tramos más alejados del concepto destino son penalizados más fuertemente mediante un factor $\lambda^{f(k,r)}$, que disminuye la influencia acumulada conforme aumenta la distancia en el camino causal. El valor de λ determina la intensidad de esta penalización: cuando λ se aproxima a 1, la penalización es débil; en cambio, si λ tiende a 0, la penalización es fuerte. Esto lo que indica en los caminos indirectos es que las conexiones lejanas tienden a tener una influencia debil, prácticamente eliminandola en caminos largos, mientras que la influencia de las conexiones cercanas al concepto de origen tienden a ser altas. La elección adecuada de λ depende del comportamiento esperado del modelo y de la naturaleza del sistema causal, permitiendo equilibrar la inclusión de influencias lejanas con la simplicidad y pertinencia de las relaciones inmediatas.

Para ilustrar el funcionamiento de $f(k,r)$, considérese un camino causal de longitud $n_r = 4$. En este caso, el tramo más cercano al concepto destino tendrá $f(k,r) = 0$, y su penalización será $\lambda^0 = 1$; el siguiente tramo tendrá $f(k,r) = 1$ y será penalizado por λ^1, luego $f(k,r) = 2$ con penalización λ^2, y así sucesivamente. De esta manera, se garantiza que los tramos más alejados del destino contribuyan menos a la influencia total.

Finalmente, para calcular la influencia indirecta, se incorpora también el nivel de activación $A^{(t)}_{p_k^{(r)}} \in [-1,1]$ del concepto en la posición k del camino r durante la iteración t. Este nivel de activación refleja la intensidad con la que un concepto participa en la inferencia en cada momento del proceso dinámico. Así, la influencia indirecta se define como el promedio sobre todos los caminos R y todas las iteraciones T de la suma de las influencias ponderadas en cada tramo, dada por:

$$I_{\text{ind}}(c_j, c_i) = \frac{1}{R} \sum_{r=1}^{R} \sum_{k=0}^{n_r-2} \left[\frac{1}{T} \sum_{t=1}^{T} \left(w_{p_k^{(r)},p_{k+1}^{(r)}} \cdot A^{(t)}_{p_k^{(r)}} \cdot \lambda^{f(k,r)} \right) \right]. \tag{4.4}$$

Esta expresión representa la *influencia acumulada que el concepto c_j ejerce sobre c_i* a través de todos los caminos indirectos posibles en el grafo causal. En ella, se ponderan tres factores fundamentales: la fuerza de conexión entre conceptos dada por los pesos w, el nivel de activación temporal de cada concepto A en las distintas iteraciones del proceso de inferencia, y una penalización exponencial $\lambda^{f(k,r)}$ que disminuye el impacto de los tramos más alejados del concepto destino. De esta manera, se captura de forma detallada y dinámica el impacto indirecto que un concepto puede tener sobre otro, considerando la estructura y evolución del sistema causal a lo largo del tiempo.

4.1.4 Fase 3: Cálculo de la Importancia Total.

Se define la importancia que un concepto c_j tiene sobre otro concepto c_i en un MCD como la suma de la influencia directa e indirecta que ejerce c_j sobre c_i. Esta medida refleja el impacto global que un concepto ejerce sobre otro a lo largo del proceso dinámico de inferencia, y se expresa mediante la ecuación (4.5):

$$I_{\text{total}}(c_j, c_i) = I_{\text{dir}}(c_j, c_i) + I_{\text{ind}}(c_j, c_i). \tag{4.5}$$

4.1.5 Fase 4: Ranking de Conceptos.

La importancia total $I_{\text{total}}(c_j, c_i)$ combina la influencia directa e indirecta que el concepto c_j tiene sobre c_i. De esta forma, se obtiene una medida completa del grado de impacto que un concepto ejerce sobre otro, considerando tanto las conexiones inmediatas como las mediadas por otros conceptos en la red. Esta métrica permite evaluar la relevancia estructural y dinámica de las relaciones en el modelo.

En esta fase, se ordenan los conceptos según el valor de su importancia $I_{\text{total}}(c_j, c_i)$ respecto al concepto c_i sobjetivo, en forma descendente. El ranking también puede representarse gráficamente para facilitar su interpretación visual y comunicar claramente cuáles son los conceptos más relevantes del modelo.

A continuación, se presenta el algoritmo para calcular la importancia total que un conjunto de conceptos c_j ejerce sobre un concepto objetivo c_i en un MCD (véase el Algoritmo 1).

Según el Algoritmo 1, para cada concepto c_j se identifica primero la conexión directa con c_i, en caso de existir, y se obtienen todos los caminos simples indirectos $R_{j \to i}$ que lo conectan con c_i, permitiendo la inclusión de ciclos autorreferenciales simples. Si el número de caminos indirectos excede un umbral máximo x, se seleccionan únicamente los x caminos más cortos. Estos caminos representan las rutas causales, tanto directas como indirectas, a través de las cuales se transmite la influencia dentro del modelo. Finalmente, la importancia total se obtiene como la suma de las influencias directa e indirecta para cada par (c_j, c_i), integrando así el impacto global del concepto origen sobre el destino. Para concluir, se construye un ranking ordenando los conceptos según su importancia total, lo que permite identificar los nodos más relevantes dentro del sistema.

Data: Matriz de pesos W, activaciones temporales $A^{(1)}, \ldots, A^{(T)}$, penalización λ, número máximo de caminos indirectos x, concepto objetivo c_i
Result: Matriz de importancia total $I_{\text{total}}(c_j, c_i)$, ranking de conceptos según influencia sobre c_i

```
// Fase 1: Identificación de caminos
```
for *cada concepto* c_j **do**
- `// Obtener el camino directo` $c_j \to c_i$`, si existe`
- $\text{CaminoDirecto}_{j \to i} \leftarrow$ camino directo $c_j \to c_i$ o vacío
- `// Obtener todos los caminos indirectos simples desde` c_j `hasta` c_i
- $R_{j \to i} \leftarrow$ conjunto de todos los caminos indirectos desde c_j hasta c_i
- `// Si hay más de` x `caminos indirectos, seleccionar los` x `más cortos`
- **if** $|R_{j \to i}| > x$ **then**
  - $R_{j \to i} \leftarrow$ los x caminos indirectos más cortos

```
// Fase 2: Cálculo de influencias
```
for *cada concepto* c_j **do**
- `// Calcular importancia directa si existe conexión directa`
- **if** $\textit{CaminoDirecto}_{j \to i} \neq \emptyset$ **then**
  - $I_{\text{dir}}(c_j, c_i) \leftarrow \frac{1}{T} \sum_{t=1}^{T} w_{j,i} \cdot A_j^{(t)}$
- **else**
  - $I_{\text{dir}}(c_j, c_i) \leftarrow 0$
- `// Calcular importancia indirecta sumando influencias ponderadas sobre caminos indirectos`
- $I_{\text{ind}}(c_j, c_i) \leftarrow 0$
- **for** *cada camino* $r \in R_{j \to i}$ **do**
  - $n_r \leftarrow$ longitud del camino r
  - **for** *cada tramo* $k = 0$ *hasta* $n_r - 2$ **do**
    - $f(k, r) \leftarrow n_r - k - 1$ `// Penalización por distancia al destino`
    - **for** *cada tiempo* $t = 1$ *hasta* T **do**
      - $I_{\text{ind}}(c_j, c_i) \mathrel{+}= w_{p_k^{(r)}, p_{k+1}^{(r)}} \cdot A_{p_k^{(r)}}^{(t)} \cdot \lambda^{f(k,r)}$
- `// Promediar influencia indirecta según número de caminos e iteraciones`
- **if** $|R_{j \to i}| > 0$ **then**
  - $I_{\text{ind}}(c_j, c_i) \leftarrow \frac{1}{|R_{j \to i}| \cdot T} \cdot I_{\text{ind}}(c_j, c_i)$
- **else**
  - $I_{\text{ind}}(c_j, c_i) \leftarrow 0$
- `// Fase 3: Importancia total`
- $I_{\text{total}}(c_j, c_i) \leftarrow I_{\text{dir}}(c_j, c_i) + I_{\text{ind}}(c_j, c_i)$

```
// Fase 4: Ranking de conceptos
```
Ranking $\leftarrow$ Ordenar conceptos c_j por $I_{\text{total}}(c_j, c_i)$ descendente
return Ranking

Algoritmo 1: Cálculo completo de la importancia de cada concepto sobre un nodo objetivo c_i combinando influencias directas e indirectas a lo largo del tiempo.

4.2 Ejemplo Ilustrativo del Funcionamiento del Método

La Figura 4.2 muestra el grafo causal correspondiente al sistema analizado, donde c_4 es la variable de clasificación sobre la cual se desea calcular la importancia del resto de los conceptos. El análisis parte de la siguiente instancia inicial del sistema:

$$A^{(0)} = [0{,}5,\ 0{,}056,\ 0{,}509,\ 0{,}5,\ c_4]$$

A partir de esta configuración, y utilizando la dinámica del modelo definida por la matriz de pesos W, se obtienen los vectores de activación correspondientes a las $T = 3$ iteraciones del sistema:

$$W = \begin{bmatrix} 0 & 0{,}6 & 0{,}4 & 0 & 0 \\ 0 & 0 & 0{,}3 & 0{,}5 & 0 \\ 0 & 0 & 0 & 0{,}7 & 0 \\ 0 & 0 & 0 & 0 & 0{,}6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{aligned} A^{(1)} &= [0{,}8,\ 0{,}5,\ 0{,}3,\ 0{,}2,\ 0{,}0] \\ A^{(2)} &= [0{,}6,\ 0{,}6,\ 0{,}5,\ 0{,}4,\ 0{,}1] \\ A^{(3)} &= [0{,}9,\ 0{,}7,\ 0{,}6,\ 0{,}5,\ 0{,}2] \end{aligned}$$

A continuación, se desarrollan de manera secuencial todas las fases que componen el método de explicabilidad propuesto, aplicadas sobre el ejemplo ilustrativo.

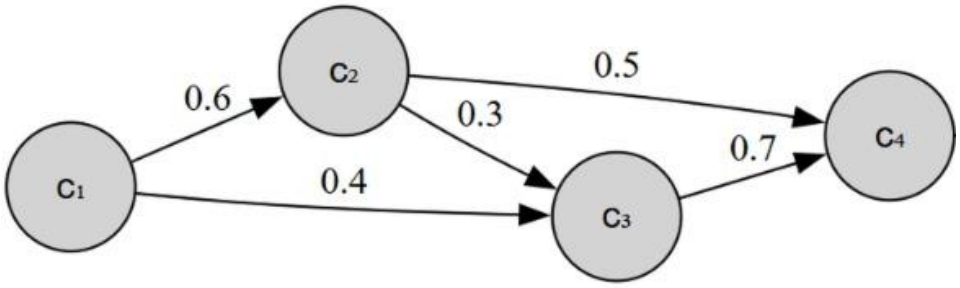

Figura 4.2: Ejemplo de grafo causal en un MCD

4.2.1 Fase 1: Identificación de Caminos

Para cada concepto dentro del MCD, se identifican los caminos directos e indirectos que conducen al concepto c_4. Los caminos directos hacia c_4 provienen de los conceptos c_2 y c_3. Por otro lado, los caminos indirectos hacia c_4 se originan en los conceptos c_1 y c_2. Desde c_1 se identifican tres caminos indirectos que conducen a c_4:

$$\begin{cases} \text{Camino 1:} & c_1 \to c_2 \to c_4, \quad (n_r = 2) \\ \text{Camino 2:} & c_1 \to c_3 \to c_4, \quad (n_r = 2) \\ \text{Camino 3:} & c_1 \to c_2 \to c_3 \to c_4, \quad (n_r = 3) \end{cases}$$

Mientras que desde c_2 existe un único camino indirecto hacia c_4, que es:

$$\begin{cases} \text{Camino 1:} & c_2 \to c_3 \to c_4, \quad (n_r = 2) \end{cases}$$

4.2.2 Fase 2: Cálculo de Influencias Directas e Indirectas

Se calcula la influencia directa que recibe el concepto c_4. Como se mencionó en la Fase 1, los conceptos c_2 y c_3 tienen conexiones directas hacia c_4. La influencia directa que un concepto c_j ejerce sobre otro concepto c_i está definida por la Ecuación (4.1).

Desde c_2 hacia c_4: La conexión desde c_2 hacia c_4 tiene un peso $w_{2,4} = 0{,}5$. Las activaciones del concepto c_2 durante las tres iteraciones $A_2^{(1)} = 0{,}5$, $A_2^{(2)} = 0{,}6$ y $A_2^{(3)} = 0{,}7$. Sustituyendo en la ecuación correspondiente, se obtiene:

$$I_{\text{dir}}(2,4) = \frac{1}{3}\left(0{,}5 \times 0{,}5 + 0{,}5 \times 0{,}6 + 0{,}5 \times 0{,}7\right) = \frac{1}{3}(0{,}25 + 0{,}30 + 0{,}35) = 0{,}3$$

Desde c_3 hacia c_4: Para la conexión desde c_3 hacia c_4, el peso es $w_{3,4} = 0{,}7$ y las activaciones de c_3 son: $A_3^{(1)} = 0{,}3$, $A_3^{(2)} = 0{,}5$ y $A_3^{(3)} = 0{,}6$. El cálculo de la influencia directa es:

$$I_{\text{dir}}(3,4) = \frac{1}{3}\left(0{,}7 \times 0{,}3 + 0{,}7 \times 0{,}5 + 0{,}7 \times 0{,}6\right) = \frac{1}{3}(0{,}21 + 0{,}35 + 0{,}42) = 0{,}3267$$

Se observa que, aunque en el grafo causal la relación entre c_2 y c_4 tiene un peso de 0.5, al aplicar el método y considerar la dinámica del sistema, dicha influencia efectiva se reduce a 0.3. De forma análoga, la relación entre c_3 y c_4 presenta un peso de 0.7 en el grafo, pero la influencia calculada mediante el método disminuye a 0.3267.

4.2.2.0.1 Influencia indirecta de c_1 En este apartado, se calcula la influencia indirecta que recibe el concepto c_4 a partir del resto de los nodos del grafo. Para ello, se consideran aquellos caminos que, sin ser conexiones directas, conducen a c_4 a través de secuencias de conceptos intermedios. En este ejemplo, los conceptos que presentan caminos indirectos hacia c_4 son c_1 y c_2. A continuación, se analiza detalladamente la contribución indirecta de cada uno de ellos.

Influencia indirecta a través del Camino 1: $c_1 \to c_2 \to c_4$ El peso de la relación desde c_1 hacia c_2 es $w_{1,2} = 0{,}6$, mientras que el de c_2 hacia c_4 es $w_{2,4} = 0{,}5$. De acuerdo con la función de penalización definida en la ecuación (4.3), y considerando que el camino tiene longitud $n_r = 2$, se tiene que $f(0,2) = 1$ y $f(1,2) = 0$. Aplicando el valor $\lambda = 0{,}9$, se obtiene:

$$\lambda^{f(0,2)} = 0{,}9, \quad \lambda^{f(1,2)} = 1$$

Sustituyendo en la ecuación de influencia indirecta (4.4), la expresión para este camino queda:

$$I_1^{(t)} = 0{,}6 \times A_1^{(t)} \times 0{,}9 + 0{,}5 \times A_2^{(t)}.$$

La tabla 4.1 muestra los cálculos para cada iteración:

t	$A_1^{(t)}$	$A_2^{(t)}$	$I_1^{(t)}$
1	0.8	0.5	$0{,}6 \times 0{,}8 \times 0{,}9 + 0{,}5 \times 0{,}5 = 0{,}73$
2	0.6	0.6	$0{,}6 \times 0{,}6 \times 0{,}9 + 0{,}5 \times 0{,}6 = 0{,}66$
3	0.9	0.7	$0{,}6 \times 0{,}9 \times 0{,}9 + 0{,}5 \times 0{,}7 = 0{,}83$

Tabla 4.1: Evolución de la influencia indirecta sobre c_4 a través del Camino 1 ($c_1 \to c_2 \to c_4$) con penalización dinámica durante las iteraciones

Una vez calculada la influencia para cada iteración, se obtiene el valor promedio:

$$\bar{I}_1 = \frac{0{,}73 + 0{,}66 + 0{,}83}{3} = 0{,}74.$$

Influencia indirecta a través del Camino 2: $c_1 \to c_3 \to c_4$ El peso de la relación desde c_1 hacia c_3 es $w_{1,3} = 0{,}4$, mientras que el de c_3 hacia c_4 es $w_{3,4} = 0{,}7$. De acuerdo con la función de penalización definida en la ecuación (4.3), se tiene que $f(0,2) = 1$ y $f(1,2) = 0$.

Sustituyendo en la ecuación de influencia indirecta (4.4), la expresión para este camino queda:

$$I_2^{(t)} = 0{,}4 \times A_1^{(t)} \times 0{,}9 + 0{,}7 \times A_3^{(t)}.$$

La tabla 4.2 muestra la evolución de la influencia indirecta para cada iteración:

t	$A_1^{(t)}$	$A_3^{(t)}$	$I_2^{(t)}$
1	0.8	0.3	$0{,}4 \times 0{,}8 \times 0{,}9 + 0{,}7 \times 0{,}3 = 0{,}768$
2	0.6	0.5	$0{,}4 \times 0{,}6 \times 0{,}9 + 0{,}7 \times 0{,}5 = 0{,}58$
3	0.9	0.6	$0{,}4 \times 0{,}9 \times 0{,}9 + 0{,}7 \times 0{,}6 = 0{,}766$

Tabla 4.2: Evolución de la influencia indirecta sobre c_4 a través del Camino 2 ($c_1 \to c_3 \to c_4$) durante las iteraciones

El valor promedio es:

$$\bar{I}_2 = \frac{0{,}768 + 0{,}58 + 0{,}766}{3} = 0{,}705.$$

Influencia indirecta a través del Camino 3: $c_1 \to c_2 \to c_3 \to c_4$ Los pesos de las relaciones son $w_{1,2} = 0{,}6$, $w_{2,3} = 0{,}3$ y $w_{3,4} = 0{,}7$. Según la función de penalización de la ecuación (4.3), se tienen los valores $f(0,3) = 2$, $f(1,3) = 1$ y $f(2,3) = 0$. Aplicando la ecuación de influencia indirecta (4.4) con el factor $\lambda = 0{,}9$, la influencia indirecta en la iteración t es:

$$I_3^{(t)} = 0{,}6 \times A_1^{(t)} \times 0{,}9^2 + 0{,}3 \times A_2^{(t)} \times 0{,}9 + 0{,}7 \times A_3^{(t)}.$$

La tabla 4.3 presenta los cálculos para cada iteración:

t	$A_1^{(t)}$	$A_2^{(t)}$	$A_3^{(t)}$	$I_3^{(t)}$
1	0.8	0.5	0.3	$0{,}6 \times 0{,}8 \times 0{,}9^2 + 0{,}3 \times 0{,}5 \times 0{,}9 + 0{,}7 \times 0{,}3 = 0{,}734$
2	0.6	0.6	0.5	$0{,}6 \times 0{,}6 \times 0{,}9^2 + 0{,}3 \times 0{,}6 \times 0{,}9 + 0{,}7 \times 0{,}5 = 0{,}804$
3	0.9	0.7	0.6	$0{,}6 \times 0{,}9 \times 0{,}9^2 + 0{,}3 \times 0{,}7 \times 0{,}9 + 0{,}7 \times 0{,}6 = 1{,}046$

Tabla 4.3: Evolución de la influencia indirecta sobre c_4 a través del Camino 3 ($c_1 \to c_2 \to c_3 \to c_4$) durante las iteraciones

El promedio de la influencia es:

$$\bar{I}_3 = \frac{0{,}734 + 0{,}804 + 1{,}046}{3} = 0{,}861.$$

Influencia total indirecta de c_1 sobre c_4 La influencia indirecta total que recibe c_4 desde c_1 se calcula como el promedio de las influencias a través de cada camino:

$$\bar{I}_{\text{total}} = \frac{\bar{I}_1 + \bar{I}_2 + \bar{I}_3}{3} = \frac{0{,}74 + 0{,}705 + 0{,}861}{3} = 0{,}769.$$

4.2.2.0.2 Influencia indirecta de c_2

Influencia indirecta a través del Camino 1: $c_2 \to c_3 \to c_4$ El peso de la relación desde c_2 hacia c_3 es $w_{2,3} = 0{,}3$, mientras que el de c_3 hacia c_4 es $w_{3,4} = 0{,}7$. De acuerdo con la función de penalización definida en la ecuación (4.3), y considerando que el camino tiene longitud $n_r = 2$, se tiene que:

$$f(0, 2) = 1, \quad f(1, 2) = 0.$$

Aplicando el valor $\lambda = 0{,}9$, se obtiene:

$$\lambda^{f(0,2)} = 0{,}9, \quad \lambda^{f(1,2)} = 1.$$

Sustituyendo en la ecuación de influencia indirecta (4.4), la expresión para este camino queda:

$$I_1^{(t)} = 0{,}3 \times A_2^{(t)} \times 0{,}9 + 0{,}7 \times A_3^{(t)}.$$

La tabla 4.4 muestra los cálculos para cada iteración:

t	$A_2^{(t)}$	$A_3^{(t)}$	$I_1^{(t)}$
1	0.5	0.3	$0{,}3 \times 0{,}5 \times 0{,}9 + 0{,}7 \times 0{,}3 = 0{,}345$
2	0.6	0.5	$0{,}3 \times 0{,}6 \times 0{,}9 + 0{,}7 \times 0{,}5 = 0{,}512$
3	0.7	0.6	$0{,}3 \times 0{,}7 \times 0{,}9 + 0{,}7 \times 0{,}6 = 0{,}609$

Tabla 4.4: Evolución de la influencia indirecta sobre c_4 a través del Camino 1 ($c_2 \to c_3 \to c_4$) con penalización dinámica durante las iteraciones

Una vez calculada la influencia para cada iteración, se obtiene el valor promedio:

$$\bar{I}_1 = \frac{0{,}345 + 0{,}512 + 0{,}609}{3} = 0{,}489.$$

Influencia total indirecta de c_2 sobre c_4 Dado que hay un único camino, la influencia indirecta total es simplemente:

$$\bar{I}_{\text{total}}^{(c_2)} = 0{,}489.$$

4.2.3 Fase 3: Cálculo de la importancia Total.

En esta sección se presenta el cálculo de la influencia total que cada concepto ejerce sobre el concepto c_4, la cual se obtiene como la suma de su influencia directa e indirecta, según lo establecido en la ecuación (4.5). La Tabla 4.5 resume los valores de importancia calculados para cada concepto en relación con c_4.

Concepto	Influencia Directa	Influencia Indirecta	Importancia Total
c_1	0.000	0.769	0.769
c_2	0.300	0.489	0.789
c_3	0.327	0.000	0.327

Tabla 4.5: Influencia directa, indirecta e importancia total de cada concepto

Se observa que el componente c_2 presenta la mayor influencia total en el sistema, con un valor de 0.789. Esto se debe a que combina una influencia directa moderada (0.300) con una influencia indirecta significativa (0.489), lo que indica que su efecto se propaga notablemente a través de otros nodos. En segundo lugar se encuentra c_1, que aunque no presenta influencia directa, tiene una influencia indirecta alta (0.769), sugiriendo que su impacto se manifiesta a través de caminos intermedios. Finalmente, c_3 muestra una influencia total menor (0.327), compuesta únicamente por una influencia directa.

Este análisis permite contrastar con la percepción inicial basada en el grafo causal. Por ejemplo, aunque c_3 tiene una relación causal directa hacia c_4 con un peso elevado (0.7), los resultados muestran que su influencia global es la más limitada, ya que no se amplifica indirectamente. En cambio, c_1 y c_2 poseen una mayor relevancia estructural en el sistema al considerar tanto las influencias directas como las indirectas.

4.2.4 Fase 4: Ranking de Conceptos.

Se construye el ranking de los conceptos en función de su importancia relativa respecto al concepto c_4. La Tabla 4.6 presenta dicho ranking, con los conceptos ordenados de mayor a menor según su impacto global sobre c_4, donde se observa que el concepto c_2 es el más importante, seguido por c_1 y, en último lugar, c_3.

Ranking	Concepto	Importancia
1	c_2	0.789
2	c_1	0.769
3	c_3	0.327

Tabla 4.6: Ranking de conceptos según su importancia total respecto a c_4

De forma complementaria, y para facilitar la interpretación visual, se propone una representación gráfica. La Figura 4.3 muestra esta representación de la importancia de los conceptos respecto a c_4, en la cual se observa el orden de importancia de cada concepto. Además, aunque en este caso todos

los conceptos contribuyen de forma positiva, la representación admite también valores negativos, que reflejarían influencias contrarias, lo que no ocurre en este ejemplo.

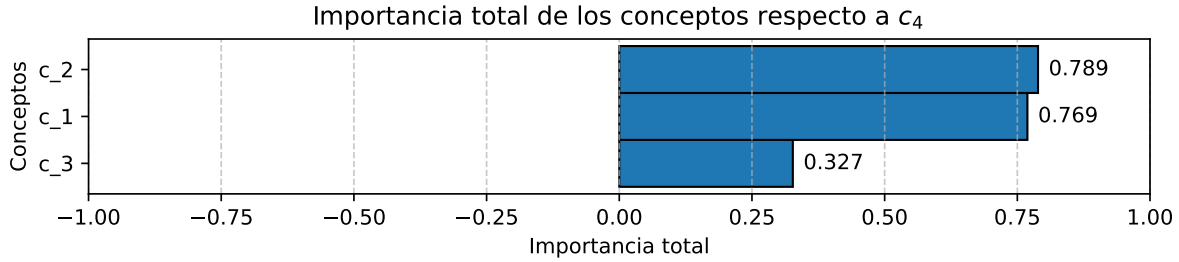

Figura 4.3: Representación gráfica de la importancia total de los conceptos respecto al concepto c_4.

Capítulo 5

Experimentos

Una vez presentado el método de explicabilidad propuesto en este trabajo, en esta sección se describen los conjuntos de datos empleados en los experimentos, incluyendo el proceso de preparación aplicado. A continuación se especifican las métricas de rendimiento utilizadas para evaluar los modelos de aprendizaje construidos, así como aquellas empleadas para evaluar la explicabilidad aportada por los distintos métodos usados. Seguidamente se expone el proceso de modelado usando los MCD y las tecnicas de AA usadas en este trabajo, para despues hacer una evaluacion de la calidad de dichos modelos. Finalmente, se hace un analisis de explicabilidad basadas en los rankings de variables establecidos por los métodos de explicabilidad (incluyendo nuestro métodos), y se realiza una comparación con otros métodos de explicabilidad basada en la degradación de los modelos según las variables relevantes definidas por cada método. Es bueno acotar que todos los experimentos realizados son para modelos de clasificación.

5.1 Datasets

En esta sección se describen los conjuntos de datos empleados en los experimentos, su procedencia, las variables que los conforman y el proceso de preparación aplicado.

5.1.1 Descripción

A continuación se describen los datasets empleados en la costrucción de los modelos de clasificación de MCDs.

5.1.1.1 Conjunto de Datos de Dengue

El dataset de dengue fue presentado en [49] y fue recopilado a partir de 52.051 pacientes que acudieron a Instituciones Prestadoras de Servicios de Salud (IPS) con diagnóstico de dengue, reportados al Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) de Colombia [112] durante el período comprendido entre 2008 y 2018 en Medellín, Colombia.

El dataset final utilizado, tras el proceso de preparación aplicado, consta de 19 variables dicotómicas que representan la presencia o ausencia de síntomas en los pacientes, codificadas como 0 o 1. También incluye una variable dicotómica que indica si el paciente pertenece a un grupo etario de riesgo, es decir, si tiene más de 60 años o menos de 5 años, así como una variable objetivo.

Nombre	Descripción
Edad	Indica pertenencia a grupos de riesgo (mayores de 60 años o menores de 5 años).
Cefalea	Dolor de cabeza.
Dolor retroocular	Dolor localizado detrás de los ojos.
Mialgias	Dolor o molestias musculares.
Artralgia	Dolor en las articulaciones.
Erupción	Presencia de lesiones cutáneas o sarpullido.
Dolor abdominal	Dolor localizado en la zona abdominal.
Vómito	Expulsión forzada del contenido gástrico por la boca.
Somnolencia	Estado de sueño o adormecimiento anormal.
Hipotensión	Presión arterial baja.
Hepatomegalia	Agrandamiento del hígado.
Hemorragias en mucosas	Sangrado visible en encías, nariz u otras mucosas.
Hipotermia	Disminución anormal de la temperatura corporal.
Aumento de hematocrito	Incremento en la concentración de glóbulos rojos en sangre.
Caída de plaquetas	Disminución en el número de plaquetas en sangre.
Acumulación de líquidos	Presencia anormal de líquido en cavidades corporales.
Extravasación	Fuga de líquido desde vasos sanguíneos hacia tejidos.
Hemorragias hemáticas	Sangrados internos o externos en tejidos o cavidades.
Shock	Insuficiencia circulatoria crítica.
Daño orgánico	Disfunción o lesión de órganos vitales.
Severidad	Nivel general de gravedad del dengue en el paciente.

Tabla 5.1: Descripción de las variables del conjunto de datos de dengue.

Este conjunto se emplea en tareas de clasificación y predicción. En la predicción, la variable objetivo presenta tres estados posibles: *DWS-negativo*, *DWS-positivo* y *Dengue severo*. Para esta tarea, se calcula la Probabilidad de Severidad del Dengue (**PDS**) mediante la siguiente función:

$$PDS(S_1) = \begin{cases} 0, & \text{si } S_1 \leq 0{,}5 \\ \left(\dfrac{S_1 - 0{,}5}{0{,}5}\right) \times 100\,\%, & \text{si } S_1 > 0{,}5 \end{cases}$$

Donde S_1 representa la probabilidad predicha por el modelo para la severidad del dengue.

El valor de *PDS* varía entre 0 y 100, clasificando la severidad del caso según el *PDS*: valores menores a 20 indican *DWS-negativo*; entre 20 y 60, *DWS-positivo*; y superiores a 60, *Dengue severo*. En las tareas de clasificación, el modelo predice directamente una de las tres categorías posibles de la variable objetivo. El dataset final consta de 32.559 registros, distribuidos de forma balanceada entre las tres clases, con un 34,5 % para *DWS-negativo*, 34,3 % para *DWS-positivo* y 31,4 % para *Dengue severo*. Las variables consideradas en este conjunto se detallan en la tabla 5.1.

5.1.1.2 Conjunto de Datos de COVID-19

El conjunto de datos de COVID-19 empleado en este estudio fue presentado en [113]. Se trata de un recurso clínico estandarizado, amplio e internacional que recopila información de pacientes hospitalizados con sospecha o diagnóstico confirmado de infección por COVID-19. La recolección de datos se realizó de forma

prospectiva en los centros de salud participantes, mediante observación directa o revisión de historias clínicas y registros médicos electrónicos, utilizando los formularios de reporte de caso desarrollados por el Consorcio Internacional para Infecciones Respiratorias Agudas Graves y Emergentes (CIIRAGE) y la Organización Mundial de la Salud (OMS).

Tras el proceso de preparación, el conjunto final incluye un total de diez variables. Cinco de ellas son dicotómicas e indican la presencia (valor 1) o ausencia (valor 0) de determinados síntomas clínicos. Además, se incluye una variable dicotómica que identifica si el paciente pertenece al grupo de riesgo (60 años o más), junto con otra que representa el sexo del paciente. Otras dos variables están asociadas al motivo por el cual se realizó la prueba diagnóstica: una indica exposición a un caso confirmado de COVID-19, y la otra, si el paciente se sometió a la prueba por haber llegado del extranjero y estar sujeto a un requisito sanitario obligatorio. Ambas se codifican de forma binaria (1 para presencia de la condición, 0 para su ausencia). Finalmente, se contempla una variable que refleja el resultado de la prueba diagnóstica para la detección del COVID-19, codificada como positiva o negativa. La Tabla 5.2 presenta una descripción detallada de las variables incluidas en el conjunto, que consta de aproximadamente 100,000 registros, balanceados en una proporción cercana al 50 % entre resultados positivos y negativos.

Nombre	Descripción
Tos	Expulsión involuntaria y repentina de aire desde los pulmones.
Fiebre	Temperatura corporal por encima del rango normal.
Dolor de garganta	Sensación dolorosa o irritativa en la mucosa faríngea.
Dificultad respiratoria	Sensación de falta de aire o dificultad para respirar.
Dolor de cabeza	Dolor localizado en la región cefálica.
Edad 60 o más	Indica pertenencia al grupo de riesgo (mayores de 60 años).
Género	Sexo del paciente (masculino o femenino).
Contacto con infectado	Prueba realizada por contacto con caso confirmado.
Procedencia del extranjero	Prueba obligatoria realizada por haber llegado del extranjero.
Resultado de la prueba	Resultado del test de COVID-19

Tabla 5.2: Descripción detallada de las variables del conjunto de datos de COVID-19.

5.1.1.3 Conjunto de Datos de Diabetes

El conjunto de datos de Diabetes utilizado en este estudio es un conjunto de referencia ampliamente conocido en la literatura científica. Fue originalmente introducido por el Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de los Estados Unidos, y está disponible públicamente a través del repositorio de aprendizaje automático de la Universidad de California en Irvine [114]. Este conjunto contiene información médica de mujeres de origen pima, un grupo indígena del sur de Arizona, mayores de 21 años. Su objetivo es predecir la aparición de diabetes tipo 2 basándose en ciertas mediciones diagnósticas comunes. Después del proceso de preparación, el conjunto incluye un total de 9 variables, todas numéricas y continuas, que incluyen mediciones relacionadas con el paciente. La tabla 5.3 presenta una descripción detallada de las variables incluidas, que consta de 1000 instancias, balanceadas aproximadamente en un 50 % entre los resultados positivos y negativos de diagnóstico.

Nombre	Descripción
Número de embarazos	Número de embarazos que ha tenido la paciente.
Glucosa	Concentración de glucosa en plasma (mg/dL).
Presión sanguínea	Presión arterial diastólica (mm Hg).
Grosor cutáneo	Espesor del pliegue cutáneo del tríceps (mm).
Insulina	Nivel de insulina en suero (mu U/mL).
Índice Masa Corporal (BMI)	Calculado como peso dividido por estatura al cuadrado (kg/m^2).
DPF (Pedigree Diabetes)	Medida del riesgo genético familiar de diabetes.
Edad	Edad de la paciente (años).
Diagnóstico	Resultado del test de diabetes (positivo o negativo).

Tabla 5.3: Descripción detallada de las variables del conjunto de datos de diabetes.

Este conjunto ha sido ampliamente utilizado como *benchmark* en tareas de clasificación, especialmente en investigaciones relacionadas con modelos de predicción médica y evaluación de algoritmos de AA [115]-[118].

5.1.1.4 Conjunto de Datos de Diagnóstico de Fallos en Vehículos Submarinos Autónomos

Este conjunto de datos fue generado empleando el vehículo submarino autónomo *Haizhe*, desarrollado en laboratorio [119]. Para su recopilación, se realizaron múltiples pruebas en las que el Vehículo Submarinos Autónomo (VAS) ejecutó un programa de navegación bajo el agua mientras se inducían fallos. Durante cada prueba, se registraron automáticamente los datos de estado del vehículo, incluyendo lecturas de sensores y variables de control, sin intervención humana. El conjunto final, tras el preprocesamiento, contiene 7 variables: 6 de ellas son continuas, y una variable objetivo denominada *Diagnóstico*, que indica el estado del submarino, especificando si se encuentra funcionando correctamente o si se ha detectado algún fallo. La tabla 5.4 presenta la descripción de las variables utilizadas. El dataset cuenta con 5000 instancias, balanceadas equitativamente entre ambos estados (funcionamiento correcto y con fallo).

Nombre	Descripción
pwm	Señal de modulación por ancho de pulso enviada a los actuadores.
voltaje	Voltaje eléctrico medido en voltios (V).
presión	Presión medida en pascales (Pa).
ángulo inclinación	Ángulo de inclinación en grados (°).
profundidad	Profundidad medida en metros (m).
ángulo rodar	Ángulo de rodar en grados (°).
velocidad angular de guiñada	Velocidad angular de guiñada en grados por segundo (°/s).

Tabla 5.4: Descripción detallada de las variables del conjunto de datos de diagnóstico de fallos en vehículos submarinos.

5.1.2 Preparación

En esta sección se detallan las operaciones de preprocesamiento realizadas sobre los distintos conjuntos de datos utilizados en el presente estudio. El propósito de este proceso es asegurar que los datos se encuentren en un formato adecuado y cuenten con la calidad necesaria para la construcción de modelos de aprendizaje automático robustos, precisos y fiables. Las tareas de preparación incluyeron, entre otras, la limpieza de datos (eliminación de registros incompletos o erróneos), la transformación de variables (como la normalización o estandarización de variables numéricas), la codificación de variables categóricas, la detección y tratamiento de valores atípicos, así como la selección de características relevantes. Estas acciones resultan esenciales para minimizar el ruido en los datos, mejorar el rendimiento predictivo de los modelos y mitigar riesgos como el sobreajuste.

Adicionalmente, en función de las particularidades de cada conjunto de datos, se llevaron a cabo ajustes específicos, tales como la creación de nuevas variables derivadas o la reestructuración del formato original de los datos. Es importante señalar que, en esta sección, se ha sintetizado la explicación del proceso general de análisis y preprocesamiento. La exposición se limita a las operaciones efectivamente realizadas y a los hallazgos más relevantes. No se abordan en profundidad los aspectos teóricos ni estadísticos del análisis exploratorio, ya que este no constituye el objeto principal del estudio.

5.1.2.1 Conjunto de Datos de Dengue

Se llevó a cabo un análisis exploratorio preliminar con el objetivo de comprender la estructura del conjunto de datos. Este conjunto está constituido principalmente por variables de tipo dicotómico, con excepción de la variable objetivo, *Severidad*, que presenta tres clases, tal como se describió previamente. En primer lugar, se verificó el tipo de dato de cada variable y, dado que la mayoría son categóricas, se evaluó el número de valores únicos presentes en cada una. Adicionalmente, se calcularon estadísticas descriptivas básicas, como la media y la desviación estándar, que, aunque aplicadas a variables dicotómicas, proporcionan información relevante sobre la distribución de los datos.

Se confirmó la ausencia de valores nulos en las variables, por lo que no fue necesario realizar labores adicionales de preprocesamiento. La Figura 5.1 presenta el gráfico de sectores correspondiente a las variables, en el que se representa el porcentaje relativo de cada categoría dentro de las variables analizadas. Se identificó que la variable *Fiebre* presentaba un único valor en todas las instancias, por lo que se consideró una constante. Dado que una variable constante no aporta variabilidad ni información útil para el análisis, se procedió a eliminarla del conjunto de datos.

Se calculó la correlación de Cramér, una medida especialmente adecuada para variables categóricas binarias, ya que evalúa la asociación entre variables cualitativas sin asumir un orden o una escala numérica. Esta métrica se basa en la tabla de contingencia y permite cuantificar la fuerza de la relación entre dos variables categóricas. La Figura 5.2 presenta los resultados de la correlación, donde se destaca que la variable *Dolor Abdominal* es la que presenta mayor correlación con la variable objetivo (0.62), seguida por *Caída de Plaquetas* (0.56) y *Hemorragias Hemáticas* (0.53). Esta matriz de correlación se usa posteriormente en el análisis de explicabilidad

Se calculó el Factor de Inflación de la Varianza (VIF), que mide el grado de colinealidad entre las variables predictoras. Los resultados indicaron que el valor máximo de VIF correspondió a la variable Dolor Abdominal con un valor de 1.70. Dado que estos valores son inferiores a los umbrales críticos comúnmente aceptados (generalmente 5 o 10) [120]-[122], no se consideró necesario eliminar ninguna variable por multicolinealidad.

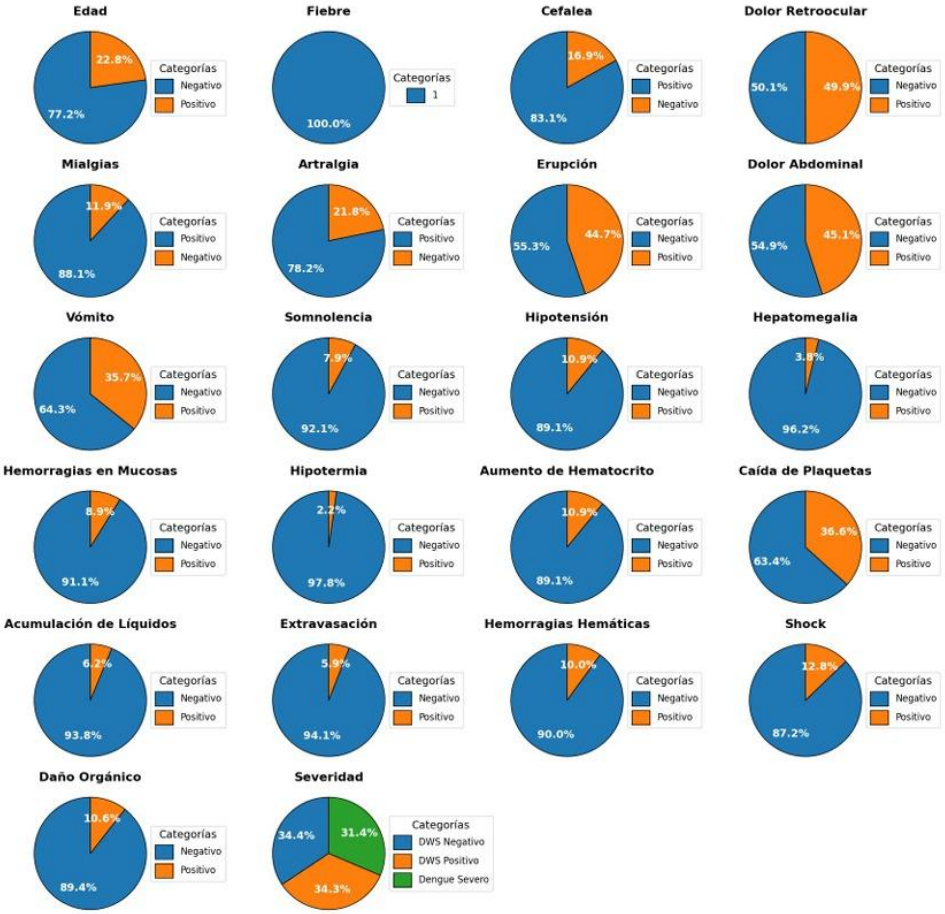

Figura 5.1: Gráfico de sectores del conjunto de datos de Dengue.

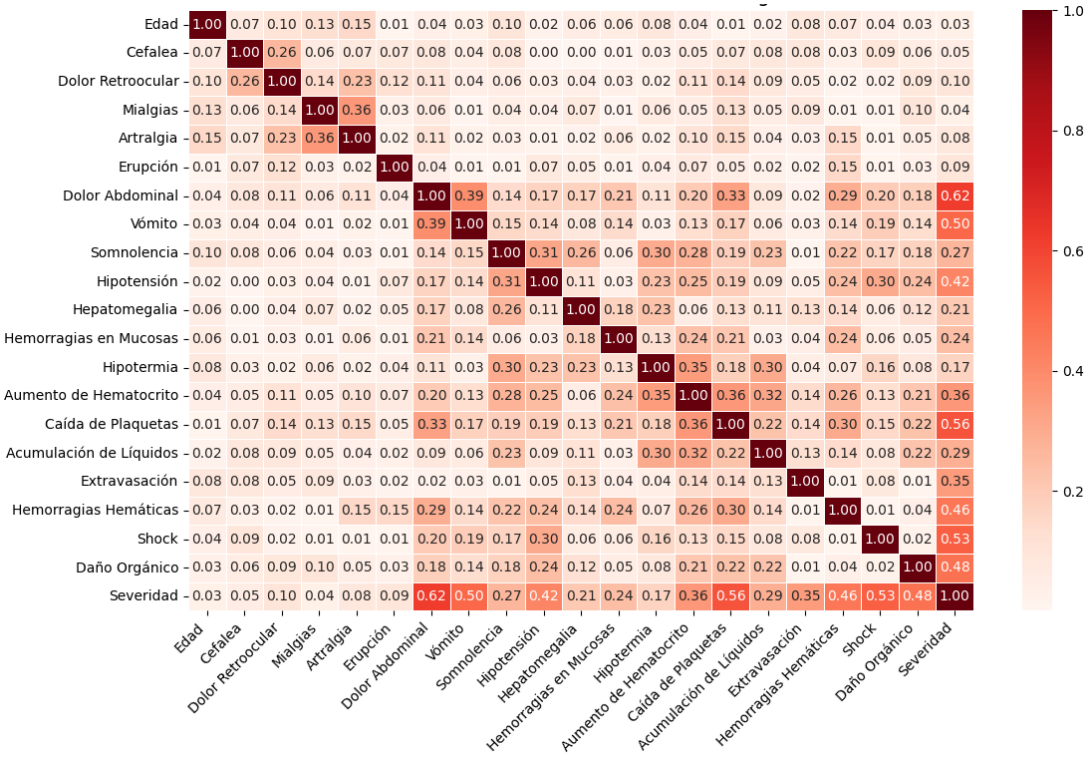

Figura 5.2: Correlación de Cramér para el conjunto de datos de Dengue.

5.1.2.2 Conjunto de Datos de COVID-19

Se verificó el número de valores únicos de cada variable. Además, se calcularon estadísticas básicas, como la media y la varianza, que resultan relevantes pese a tratarse de variables categóricas. Se analizó la proporción de valores positivos y negativos mediante gráficos de sectores. En la Figura 5.3 se observa que la variable objetivo *Resultado de Prueba* presenta un marcado desequilibrio, con una cantidad significativamente mayor de casos negativos respecto a los positivos. Para corregir esta desproporción, se aplicó la técnica de sobremuestreo sintético SMOTE[123] , que genera nuevas muestras sintéticas de la clase minoritaria con el propósito de equilibrar la distribución de clases y mejorar el rendimiento de los modelos predictivos.

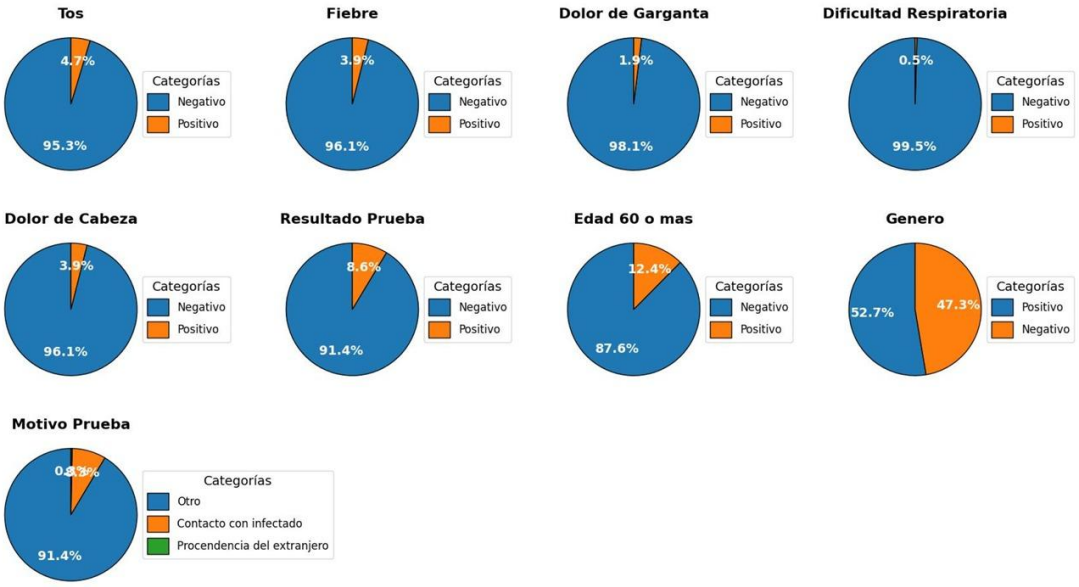

Figura 5.3: Gráfico de sectores del conjunto de datos de COVID-19.

Dado que la variable *Motivo de Prueba* cuenta con tres clases distintas, esta fue dividida en dos variables binarias: *Procedencia del Extranjero* y *Contacto con Infectado*, ya que esta transformación permitió obtener mejores resultados en la construcción de los modelos.

Se identificaron instancias con características idénticas pero con valores diferentes en la variable objetivo *Resultado de Prueba*, lo que puede generar inconsistencias durante el entrenamiento de modelos debido a la información contradictoria. Para mitigar este problema, se definió una variable auxiliar denominada *Coeficiente de Riesgo*, que cuantifica la probabilidad o nivel de riesgo asociado a cada instancia. El *Coeficiente de Riesgo* se estableció como una combinación lineal ponderada de variables clínicas y demográficas relevantes, expresada mediante la siguiente fórmula:

$$\begin{aligned}\text{Coeficiente de Riesgo} = {} & 0{,}1 \times \text{Tos} + 0{,}2 \times \text{Fiebre} \\ & + 0{,}1 \times \text{Dolor de Garganta} + 0{,}1 \times \text{Dificultad Respiratoria} \\ & + 0{,}1 \times \text{Dolor de Cabeza} + 0{,}1 \times \text{Edad 60 o más} \\ & + 0{,}1 \times \text{Motivo Procedencia del Extranjero} \\ & + 0{,}2 \times \text{Motivo Contacto con Infectado}\end{aligned}$$

Los pesos asignados a cada variable fueron obtenidos a partir de la literatura [124]-[126]. Esta definición facilitó la identificación y eliminación de instancias contradictorias, contribuyendo a mejorar la calidad del conjunto de datos para el desarrollo de modelos de clasificación.

Con base en este coeficiente, se aplicaron criterios de filtrado para eliminar observaciones inconsistentes: se descartaron aquellas en las que el coeficiente de riesgo es elevado pero el resultado de la prueba es negativo, y viceversa. El Algoritmo 2 muestra el procedimiento para identificar y eliminar instancias contradictorias en el conjunto de datos. Para cada observación, calcula el Coeficiente de Riesgo, que representa una estimación del nivel de gravedad asociado. El algoritmo elimina aquellas instancias donde existe discrepancia entre este coeficiente y el valor real de la variable objetivo: se descartan los casos con un coeficiente alto pero un resultado negativo, y también los que presentan un coeficiente bajo pero un resultado positivo. Este filtrado busca mejorar la calidad, coherencia y robustez del conjunto de datos, evitando que datos inconsistentes afecten negativamente el entrenamiento y desempeño de los modelos predictivos.

Data: Conjunto de datos con variables predictoras y la variable objetivo *Resultado de Prueba*
Result: Conjunto de datos sin instancias contradictorias
`// Calcular el Coeficiente de Riesgo para cada instancia`
Para cada instancia $x \in D$, calcular $CR(x) \in [0, 1]$, que representa el riesgo estimado
`// Filtrar instancias contradictorias`
for *cada instancia* $x \in D$ **do**
  if $CR(x) \geq 0{,}6$ ***y*** $Resultado(x) = 0$ **then**
    Eliminar x del conjunto de datos
  else if $CR(x) \leq 0{,}3$ ***y*** $Resultado(x) = 1$ **then**
    Eliminar x del conjunto de datos
return *Conjunto de datos filtrado* D

Algoritmo 2: Filtrado de instancias contradictorias utilizando el Coeficiente de Riesgo.

La Figura 5.4 presenta el gráfico de sectores correspondiente al conjunto de datos de COVID-19 tras las operaciones de preprocesamiento descritas anteriormente. Se observa que, en comparación con el estado inicial, la variable objetivo *Resultado de Prueba* presenta un balance adecuado entre clases. Además, se evidencia un incremento en la proporción de casos positivos asociados a cada uno de los síntomas analizados.

Se aplicó la correlación de Cramér para evaluar la asociación entre las variables categóricas y la variable objetivo. Se observó que la variable *Motivo Contacto con Infectado* presenta una correlación de 0.60 con el *Resultado de Prueba*, seguida por *Motivo Procedencia del Extranjero* con 0.60, *Fiebre* con 0.54 y *Tos* con 0.51, siendo estas las variables con mayor relación destacada. Esta matriz de correlación se usa posteriormente en el análisis de explicabilidad

Se realizó la prueba de significación chi-cuadrado para evaluar la independencia entre cada variable y la variable objetivo. En todos los casos, se rechazó la hipótesis nula de independencia, lo que indica que todas las variables analizadas son estadísticamente significativas respecto al *Resultado de Prueba*. Finalmente, se llevó a cabo un análisis de multicolinealidad mediante el VIF. Los resultados indicaron que la variable *Resultado de Prueba* presenta el valor máximo de VIF con 3.64, seguida por *Fiebre* con 1.59. Dado que estos valores son inferiores a umbrales críticos, no se consideró necesario eliminar ninguna variable por multicolinealidad.

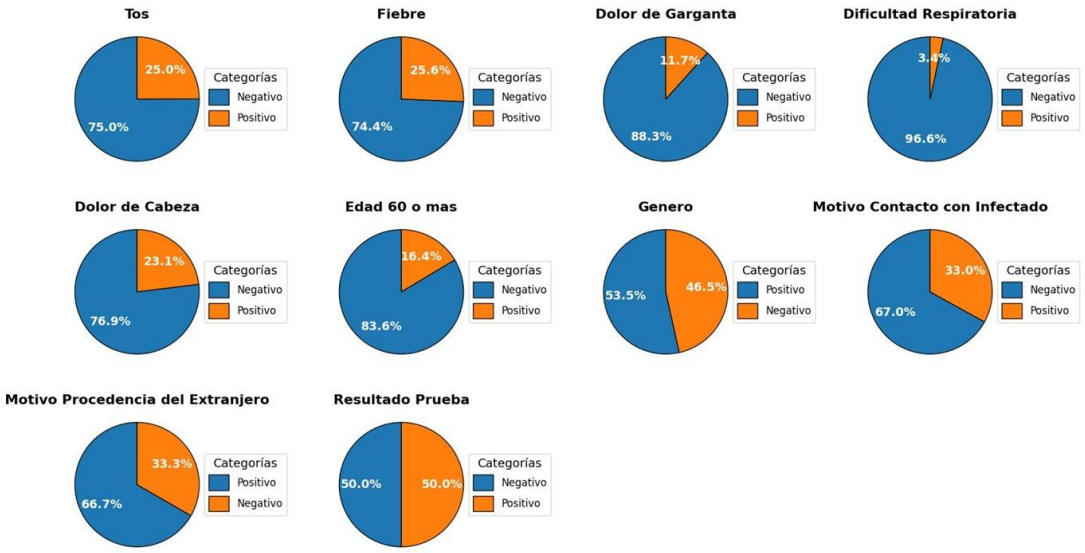

Figura 5.4: Gráfico de sectores del conjunto de datos de COVID-19 tras el preprocesamiento.

5.1.2.3 Conjunto de Datos de Diabetes

El conjunto de datos está compuesto por variables numéricas, excepto la variable objetivo *Diagnóstico*, la cual es categórica binaria, con valores 0 (*negativo*) y 1 (*positivo*). Se llevó a cabo un análisis exploratorio en el que se examinaron estadísticas básicas y la posible presencia de valores nulos. No se detectaron datos faltantes. Posteriormente, se analizó la distribución de las variables numéricas mediante histogramas. La Figura 5.5 muestra los histogramas correspondientes a las variables continuas del conjunto de datos. Para determinar el número de intervalos (*bins*) en cada histograma, se aplicó la regla de Freedman–Diaconis [127], que ajusta el ancho de los intervalos en función de la dispersión de los datos:

$$\text{Ancho del bin} = 2 \times \frac{IQR}{\sqrt[3]{n}}$$

donde IQR es el rango intercuartílico y n el número de observaciones. En los histogramas se observa que variables como *Número de Embarazos*, Grosor Cutáneo, *Insulina* y *Edad* no presentan una distribución normal, lo cual es relevante para el modelado posterior.

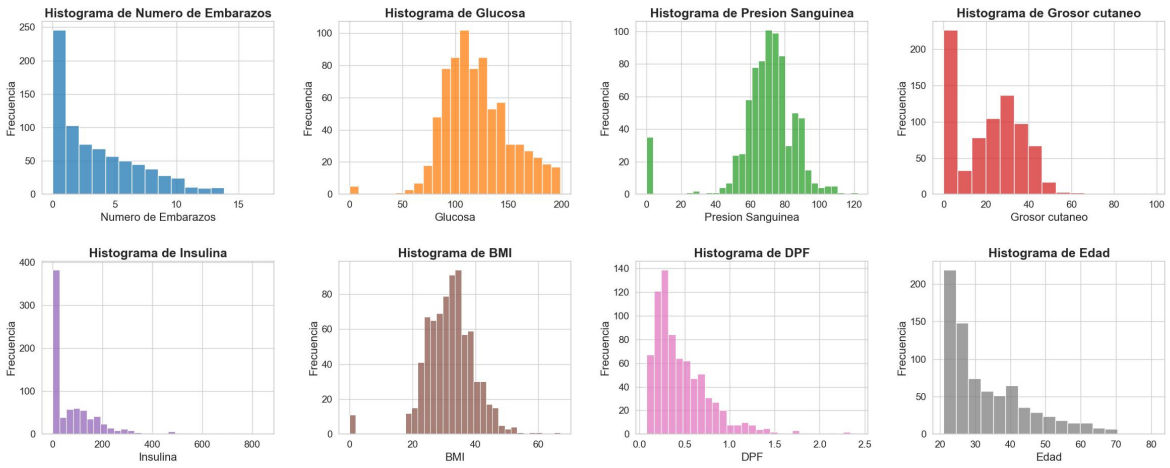

Figura 5.5: Histogramas de variables numéricas del conjunto de datos de diabetes.

Para complementar este análisis, se emplearon *Q-Q plots* (Quantile-Quantile plots) [128], [129], que permiten evaluar visualmente la aproximación de una variable a una distribución normal mediante la comparación de cuantiles teóricos y observados. La Figura 5.6 presenta los *Q-Q plots* de las variables continuas. Se observa que la variable *Número de Embarazos* exhibe una estructura escalonada debido a su naturaleza discreta, con múltiples observaciones compartiendo valores idénticos. Variables como *Glucosa*, *Presión Sanguínea*, *Grosor Cutáneo*, *Insulina* e *BMI* presentan una concentración notable de valores iguales a cero, lo que podría reflejar registros erróneos o datos faltantes codificados inapropiadamente. Además, en las variables *DPF*, *Insulina* y *Edad* se aprecia una separación en la cola superior respecto a la línea de referencia, sugiriendo distribuciones asimétricas o con colas pesadas.

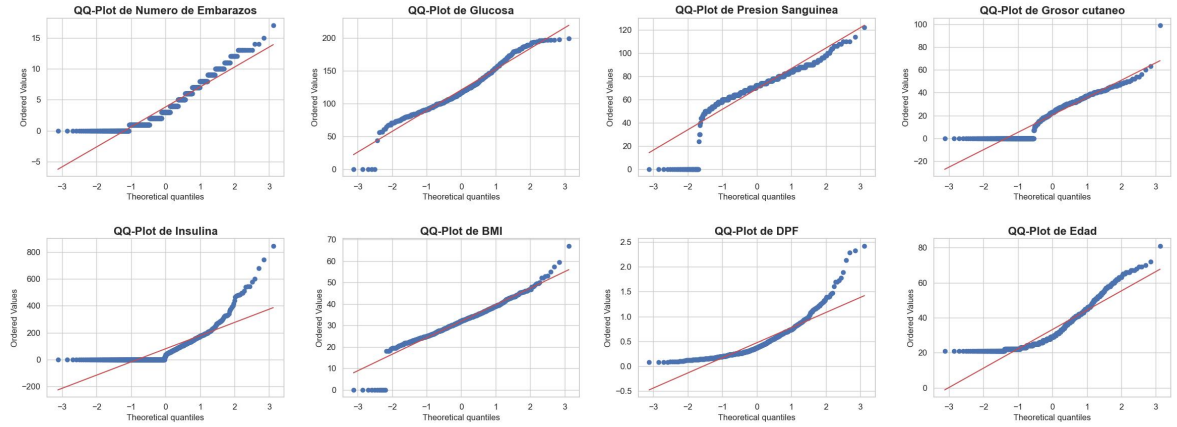

Figura 5.6: *Q-Q plots* de las variables numéricas del conjunto de datos de diabetes.

A partir de lo observado en los *Q-Q plots*, donde se identificaron patrones como la alta concentración de valores en cero, estructuras escalonadas, y desviaciones en las colas de algunas distribuciones, se consideró necesario verificar la presencia de valores atípicos (*outliers*) en las variables analizadas. Para ello, se emplearon dos métodos estadísticos: el rango intercuartílico (IQR) y el puntaje estándar (Z-score).

En el método del rango intercuartílico (IQR), se consideraron atípicos aquellos valores situados fuera del rango $[Q1 - 1{,}5 \times IQR, Q3 + 1{,}5 \times IQR]$, donde $Q1$ y $Q3$ son los cuartiles primero y tercero, respectivamente, e $IQR = Q3 - Q1$ representa el rango intercuartílico. En el método del *Z-score*, se consideraron atípicos aquellos valores cuya puntuación estándar se encontraba fuera del rango $[-3, 3]$. Esta puntuación se calcula mediante la fórmula $Z = \frac{x-\mu}{\sigma}$, donde x es el valor de la observación, μ la media de la variable y σ su desviación estándar. De esta forma, se identificaron como valores atípicos las observaciones que se alejaban más de tres desviaciones estándar respecto a la media, ya que tales casos son poco probables bajo una distribución normal.

Los resultados para cada variable son:

- *Número de Embarazos*: 4 outliers (0.52 %) detectados por ambos métodos.
- *Glucosa*: 5 outliers (0.65 %) identificados por ambos métodos.
- *Presión Sanguínea*: 45 outliers (5.86 %) por IQR y 35 (4.56 %) por *Z-score*.
- *Grosor Cutáneo*: 1 outlier (0.13 %) según ambos métodos.
- *Insulina*: 34 outliers (4.43 %) con IQR y 18 (2.34 %) con *Z-score*.

- *Índice de Masa Corporal (BMI)*: 19 outliers (2.47 %) por IQR y 14 (1.82 %) por *Z-score*.
- *DPF*: 29 outliers (3.78 %) detectados por IQR y 11 (1.43 %) por *Z-score*.
- *Edad*: 9 outliers (1.17 %) por IQR y 5 (0.65 %) por *Z-score*.

La Figura 5.7 presenta los diagramas de caja y bigotes que corroboran la detección de outliers mediante el método IQR.

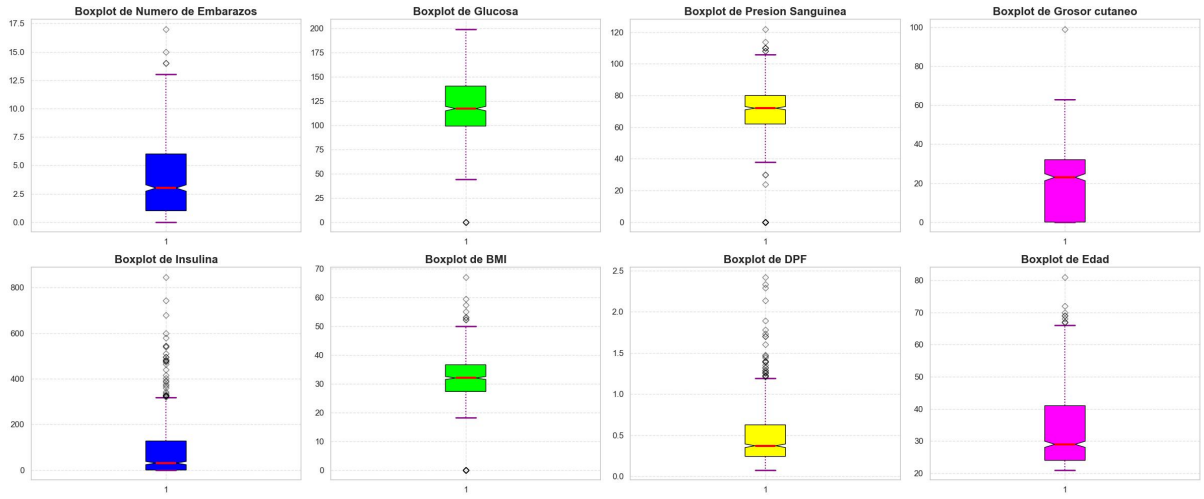

Figura 5.7: Diagramas de caja y bigotes para las variables numéricas del conjunto de datos de diabetes.

Para mitigar el impacto de los valores cero anómalos, presentes en variables como *Glucosa*, *Presión Sanguínea*, *Grosor Cutáneo*, *Insulina* y *Índice de Masa Corporal*, se aplicó una técnica de imputación basada en vecinos más cercanos (*K-Nearest Neighbors Imputation*). Esta técnica considera dichos valores cero como faltantes y estima su valor a partir de la media de los $k = 5$ vecinos más cercanos, lo que permite preservar la coherencia de la información y la estructura de los datos. La Figura 5.8 muestra los histogramas posteriores a la imputación, en los cuales se observa un aumento en el número de bins y una distribución más cercana a la normalidad para las variables que inicialmente presentaban distribuciones atípicas. Cabe destacar que esta imputación se aplicó únicamente a las variables mencionadas, mientras que otras como el número de embarazos y la edad mantuvieron su distribución original, ya que los valores cero en ellas no se consideraron anómalos.

Posteriormente, se llevó a cabo un análisis ANOVA para evaluar la significancia estadística entre las variables numéricas y la variable objetivo *Diagnóstico*. El ANOVA es una técnica estadística que permite determinar si existen diferencias significativas en las medias de una variable numérica entre dos o más grupos definidos por una variable categórica. En este caso, se utilizó para comparar las medias de cada variable numérica entre los dos grupos de *Diagnóstico* (0: negativo, 1: positivo). Un resultado significativo indica que la variable numérica contribuye a diferenciar los grupos, justificando su inclusión en el modelo predictivo. Se comprobó que todas las variables numéricas presentan significancia estadística.

Finalmente, se calculó el VIF, obteniéndose valores entre 1.86 para *BMI*, seguido de *Glucosa* y 1 para *DPF*, con el resto de variables situadas en rangos intermedios. Dado que estos valores son bajos, no se procedió a eliminar ninguna variable. Se calcularon las correlaciones entre las variables numéricas utilizando los coeficientes de Pearson y Spearman. Se observó una correlación notable entre las variables *Número de embarazos* y *Edad*, con un valor de 0.54 según Pearson y 0.61 según Spearman, lo que indica una asociación positiva moderada entre ambas variables. No se detectaron otras correlaciones significativas

de mayor magnitud entre las variables numéricas del conjunto de datos. A las variables numéricas se le aplicó la normalización *Min-Max* para escalarlas, ya que muchos algoritmos de AA son sensibles a la escala de los datos. Cuando las variables tienen rangos muy diferentes, las de mayor magnitud pueden dominar el proceso de aprendizaje, afectando el rendimiento del modelo. Además, al escalar las variables a un rango común, se evita que alguna variable tenga más peso simplemente por su escala numérica. Finalmente, para balancear el conjunto de datos, se aplicó la técnica SMOTE.

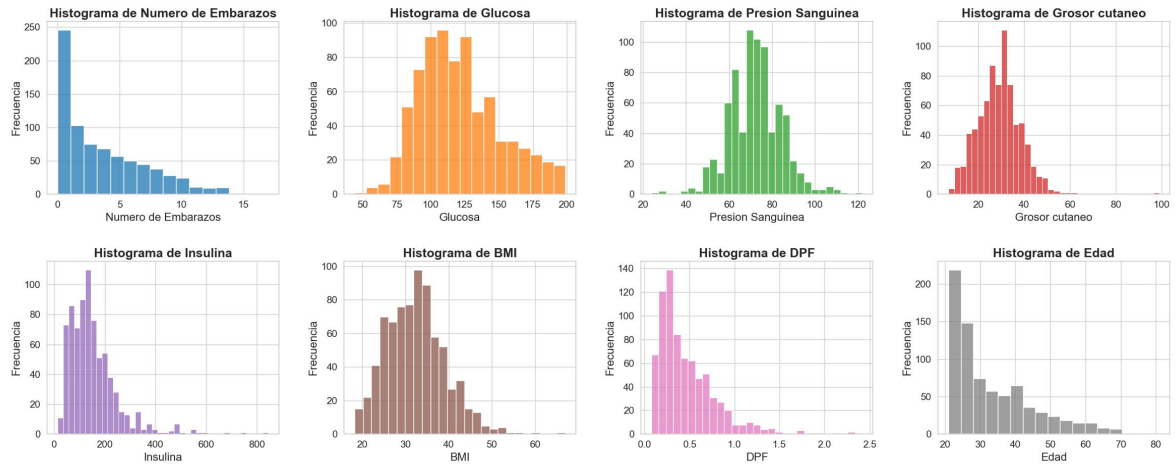

Figura 5.8: Histogramas de variables numéricas del conjunto de datos de diabetes despues de la imputación.

Tras realizar la imputación de los valores cero, se comprobó que el número de valores atípicos, determinado mediante los métodos de IQR y Z-score, disminuyó considerablemente, por lo que no fue necesario aplicar técnicas adicionales para su tratamiento.

5.1.2.4 Conjunto de Datos de Diagnóstico de Fallos en Vehículos Submarinos Autónomos

En el conjunto de datos correspondiente al diagnóstico de fallos en vehículos submarinos, todas las variables son numéricas, a excepción de la variable objetivo *estado*, la cual es binaria: el valor 0 indica que el vehículo se encuentra en buen estado, mientras que el valor 1 indica la presencia de una avería. Se realizó un análisis exploratorio inicial que incluyó la visualización de estadísticas descriptivas básicas, sin detectarse valores faltantes en ninguna de las variables. Además, se examinó la distribución de las variables numéricas mediante histogramas, los cuales se presentan en la Figura 5.9.

Al igual que en el conjunto de datos de diabetes, se empleó la regla de Freedman–Diaconis para determinar el número de intervalos (*bins*) en cada histograma. A partir de los resultados, se observa que la mayoría de las variables presentan distribuciones cercanas a la uniforme, con excepción de aquellas asociadas a las señales *PWM* de los motores y la variable *velocidad angular guinada*.

Para complementar este análisis, se utilizaron gráficos *Q-Q plots*, los cuales permitieron evaluar la normalidad de las distribuciones. Como se muestra en la Figura 5.10, la mayoría de las variables siguen una distribución aproximadamente normal, aunque con ligeras desviaciones en las colas. La variable *velocidad angular guinada* destaca especialmente por sus desviaciones respecto a la línea de referencia, lo que evidencia una mayor discrepancia con la normalidad teórica.

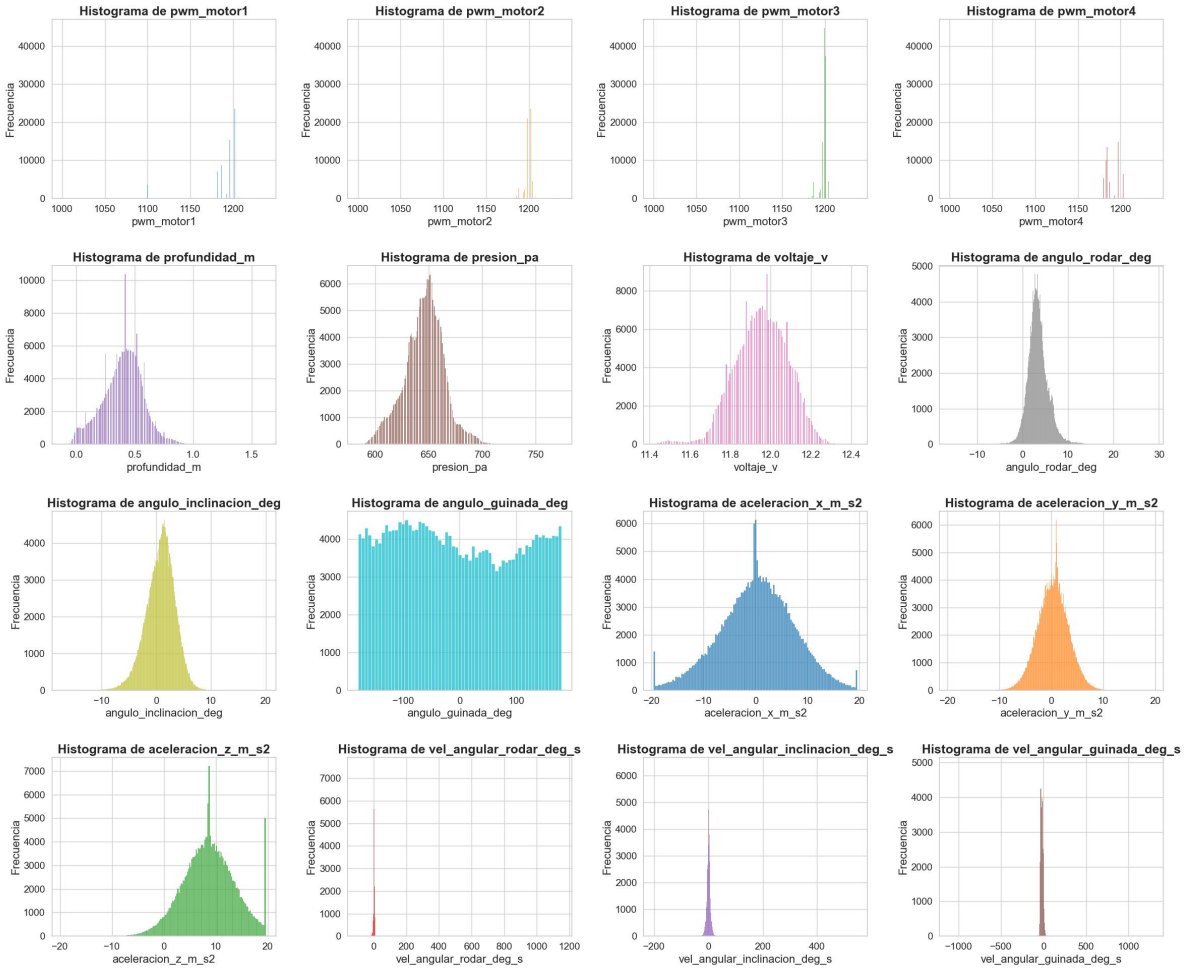

Figura 5.9: Histogramas de variables numéricas del conjunto de datos de diagnóstico de fallos en vehículos submarinos.

Se llevó a cabo un análisis de detección de valores atípicos (*outliers*) utilizando tanto el método del IQR como el método del Z-score. En general, no se identificaron anomalías relevantes, a excepción de las variables correspondientes a las señales *PWM* de los motores, en las cuales se observó una proporción cercana al 4 % de valores atípicos en ambos métodos.

Dado que cada uno de los cuatro motores del vehículo recibe una señal *PWM* independiente, y considerando que dichas señales pueden diferir en magnitud o comportamiento, se optó por combinar la información en una única variable representativa. Para ello, se construyó una nueva variable calculando el producto de las señales *PWM* de los cuatro motores Esta operación permite capturar de forma compacta la interacción conjunta de las señales enviadas a los motores. Posteriormente, se eliminaron las columnas originales de las señales individuales, conservando únicamente la variable compuesta `pwm`, que se empleará en los análisis posteriores. A continuación, se calcularon las correlaciones entre las variables numéricas utilizando los coeficientes de Pearson y Spearman, sin observarse asociaciones de relevancia estadística.

También se aplicó un análisis de varianza (ANOVA) para evaluar la significancia estadística de las variables respecto a la variable objetivo. Los resultados indicaron que varias variables no presentaban diferencias significativas entre los grupos. A su vez, se realizó un análisis de colinealidad mediante el cálculo del VIF, sin detectarse valores que justificaran la eliminación de variables por multicolinealidad.

Con el objetivo de equilibrar las clases del conjunto de datos, se aplicó la técnica SMOTE, seguida de una normalización de las variables numéricas mediante escalado *Min-Max*.

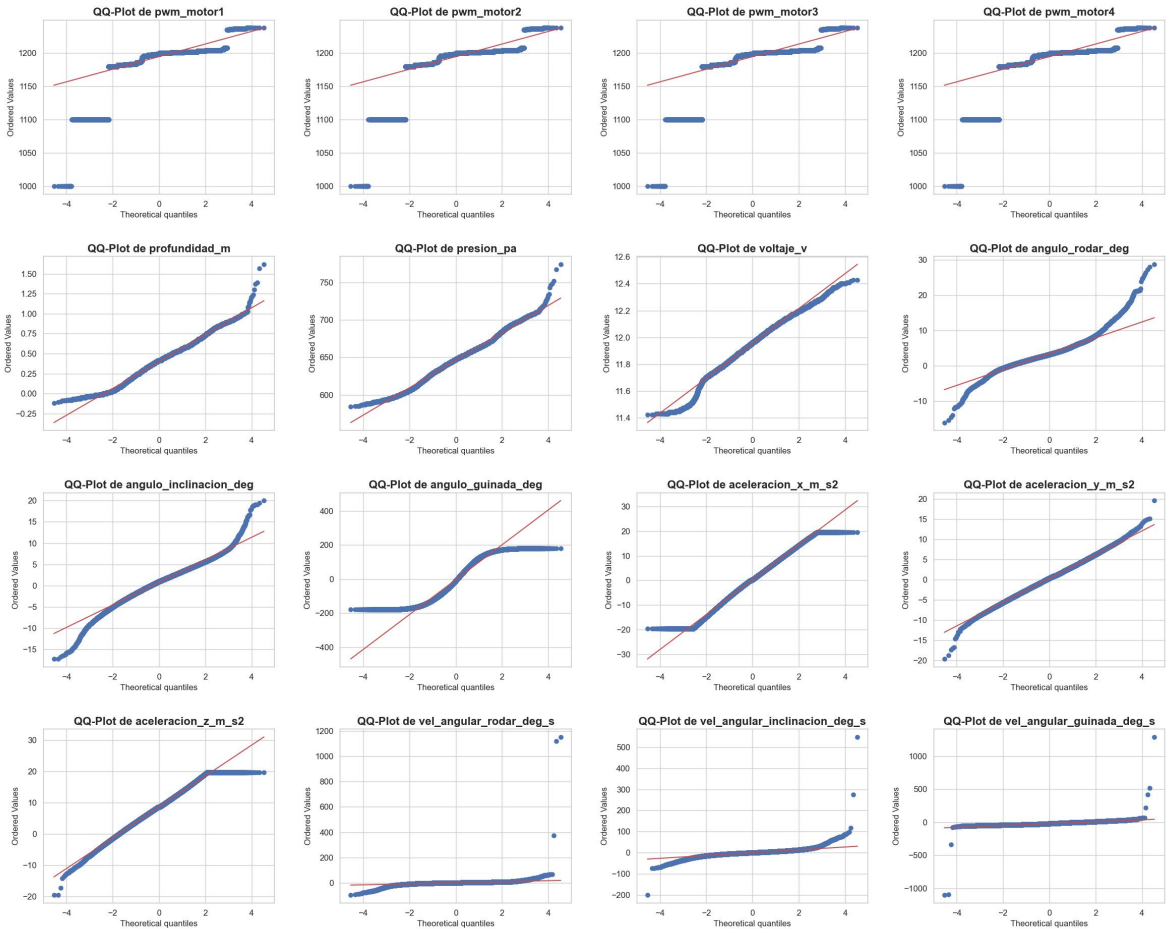

Figura 5.10: *Q-Q plots* de las variables numéricas del conjunto de datos de diagnóstico de fallos en vehículos submarinos.

Inicialmente, la variable objetivo *estado* presentaba cinco clases distintas. No obstante, al aplicar modelos de MCDs, se comprobó que el algoritmo no lograba converger debido a la escasez de muestras por clase. Esta falta de convergencia se observó tanto al incluir todas las variables, como al utilizar únicamente las seleccionadas por ANOVA como estadísticamente significativas.

Por tal motivo, se reformuló el problema como una tarea de clasificación binaria, donde el `estado 0` representa el funcionamiento normal y el `estado 1` agrupa las cuatro clases asociadas a fallos. Una vez redefinido el objetivo, se aplicó nuevamente la técnica SMOTE para abordar el desbalance de clases en esta nueva configuración. Finalmente, se construyeron modelos utilizando tanto el conjunto de variables significativas identificadas por ANOVA como el conjunto completo de variables. Los resultados mostraron un mejor rendimiento al excluir las variables no significativas, lo que respalda la utilidad del análisis ANOVA en la etapa de selección de características.

5.2 Métricas

Esta sección presenta las métricas empleadas para evaluar el **rendimiento** de los modelos desarrollados para tareas de clasificación. Se utilizan métricas clásicas de clasificación, como la matriz de confusión y sus métricas derivadas: *accuracy*, *precision*, *recall* y *F1-score*. Además, se analiza el área bajo la curva ROC (AUC-ROC) para medir la capacidad discriminativa de los modelos.

A continuación se describen las métricas de clasificación:

- **Matriz de confusión:** Es una tabla que resume el desempeño del modelo al clasificar las instancias en:
  - **TP (True Positives):** instancias positivas correctamente clasificadas.
  - **FP (False Positives):** instancias negativas clasificadas incorrectamente como positivas.
  - **TN (True Negatives):** instancias negativas correctamente clasificadas.
  - **FN (False Negatives):** instancias positivas clasificadas incorrectamente como negativas.

  Esta matriz permite calcular todas las métricas de evaluación posteriores.

- **Exactitud (Accuracy):** Mide la proporción de predicciones correctas sobre el total de predicciones.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

- **Precisión (Precision):** Indica cuántas de las instancias clasificadas como positivas son efectivamente positivas.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

- **Sensibilidad (Recall):** Mide la proporción de instancias positivas correctamente identificadas por el modelo.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

- **Puntuación F1 (F1-score):** Representa la media armónica entre precisión y sensibilidad, lo cual favorece un equilibrio entre ambas métricas.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

- **Curva ROC y AUC:** La curva ROC muestra la relación entre la Tasa de Verdaderos Positvos (TPR) y Tasa de Falsos Positvos (FPR) al variar el umbral de decisión. El área bajo esta curva (AUC) cuantifica la capacidad del modelo para distinguir entre clases.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{5.5}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5.6}$$

  Un AUC cercano a 1 indica un modelo con alta capacidad discriminativa, mientras que un valor cercano a 0.5 sugiere un modelo sin poder de discriminación.

Además de evaluar el rendimiento del modelo, se analiza la robustez de las **explicaciones** generadas por el método de explicabilidad propuesto. Para que un método explicativo se considere robusto, debe

cumplir con un conjunto de propiedades que permitan medir su **calidad**, **utilidad** y **confiabilidad**. Estas propiedades, descritas en la Sección 2.3.3, incluyen: *fidelidad*, *consistencia*, *robustez* y *eficiencia*. Además, las explicaciones generadas por el método propuesto se comparan con medidas de centralidad provenientes de la teoría de grafos, descritas en la Sección 2.3.5.1, con el objetivo de analizar la correspondencia entre dichas explicaciones y la relevancia estructural de los nodos en el grafo. Las medidas consideradas incluyen: *grado de entrada*, *grado de salida*, *grado total*, *PageRank* e *intermediación*.

Para realizar una **comparación cuantitativa de la calidad** entre distintos métodos de explicabilidad, se utiliza la técnica ROAR, que permite estimar la importancia real de las características seleccionadas. ROAR [130] elimina de forma progresiva las características consideradas más relevantes por un método de explicabilidad, reentrena el modelo desde cero en cada iteración y evalúa su rendimiento. Si el método es eficaz, la eliminación de características importantes debería provocar una caída notable en el rendimiento del modelo. El procedimiento se repite de forma iterativa: se elimina primero la característica más relevante y se mide el rendimiento; luego se eliminan las dos más relevantes y se reevalúa; posteriormente tres, y así sucesivamente. Este proceso permite medir de manera objetiva el impacto de las características seleccionadas sobre el rendimiento del modelo.

5.3 Modelado

En esta sección se presentan tanto el modelado usando los MCD como con las otras tecnicas de aprendizaje automátivo usadas en este trabajo.

5.3.1 Modelado Usando MCDs

Esta sección describe detalladamente el proceso seguido para implementar el enfoque basado en modelos MCD para tareas de clasificación. Se explica el uso del software *FCM Experts* para la construcción y entrenamiento de los modelos, así como la optimización de sus parámetros mediante algoritmos evolutivos. Además, se discuten las principales limitaciones encontradas en la definición estructural de los modelos.

El software *FCM Experts* [53] fue empleado para la construcción y entrenamiento de los modelos de clasificación basados en MCDs. La construcción y entrenamiento de los modelos de clasificación se fundamenta en algoritmos de optimización por poblaciones, como se explica en la Sección 2.2. En particular, se utilizó el algoritmo *Particle Swarm Optimization* (PSO) [131], y específicamente su variante *Global-best PSO* [132], para la optimización de la matriz de pesos W. *Global-best PSO* es un algoritmo diseñado para resolver problemas de optimización mediante un enfoque estocástico inspirado en el comportamiento colectivo de las poblaciones. En este contexto, un conjunto de partículas explora el espacio de soluciones, donde cada partícula representa una posible solución. Cada partícula se caracteriza por:

1. Una posición $\mathbf{X}_i$ en el espacio de búsqueda.
2. Una velocidad $\mathbf{V}_i$ que determina su desplazamiento.
3. Su mejor posición histórica pbest_i.

El algoritmo también mantiene gbest, que representa la mejor posición global hallada por el enjambre. Las partículas actualizan sus posiciones y velocidades en cada iteración usando las ecuaciones 5.7 y 5.8 :

$$\mathbf{X}_i^{t+1} = \mathbf{X}_i^t + \mathbf{V}_i^{t+1} \tag{5.7}$$

$$\mathbf{V}_i^{t+1} = w \cdot \mathbf{V}_i^t + c_1 \cdot r_1 \cdot (\text{pbest}_i - \mathbf{X}_i^t) + c_2 \cdot r_2 \cdot (\text{gbest} - \mathbf{X}_i^t) \tag{5.8}$$

donde w es el coeficiente de inercia, y c_1, c_2 son los coeficientes cognitivo y social, respectivamente. Los valores r_1 y r_2 son variables aleatorias uniformemente distribuidas en el intervalo $[0, 1]$.

Global-best PSO se emplea para construir, a partir de los datos, la matriz de pesos W del MCD. En este contexto, cada partícula del enjambre representa una posible solución, es decir, una matriz W candidata que define las relaciones causales entre los conceptos del modelo. Durante la búsqueda, el enjambre explora iterativamente el espacio de soluciones conformado por todas las configuraciones posibles de W, evaluando el desempeño de cada partícula mediante una función objetivo basada en la *exactitud* del modelo. Así, las partículas actualizan sus posiciones y velocidades guiadas tanto por su mejor experiencia individual (pbest_i) como por la mejor solución global encontrada (gbest), hasta que se cumple un criterio de parada predefinido, como alcanzar un número máximo de iteraciones o la estabilización del rendimiento del modelo.

La selección de los hiperparámetros (tamaño de población, número máximo de iteraciones, c_1, c_2) se realizó de forma empírica para cada conjunto de datos, empleando validación cruzada para garantizar una evaluación robusta del rendimiento. El trabajo de [133] destaca que coeficientes $c_1 = c_2 \approx 0{,}2$ proporcionan un equilibrio efectivo entre la exploración y explotación del espacio de búsqueda. Por otro lado, aunque [134] sugiere valores óptimos para el tamaño del enjambre en función de las características del conjunto de datos (específicamente, el tamaño de la población), en el presente estudio fue necesario ajustar dicho parámetro de manera específica, ya que las recomendaciones propuestas no proporcionaron resultados satisfactorios con los datos empleados. Sin embargo, las recomendaciones respecto a los valores de c_1 y c_2 se mantuvieron en 0.2. El proceso completo seguido para construir y validar cada modelo fue el siguiente:

1. **Validación cruzada:** Se llevaron a cabo diez validaciones cruzadas, dividiendo los datos en un 70 % para entrenamiento y un 30 % para prueba, con el fin de obtener una estimación robusta del rendimiento del modelo.

2. **Selección de hiperparámetros:** Se establecieron los valores de los hiperparámetros relevantes, tales como el tamaño de la población, el número máximo de iteraciones y los coeficientes cognitivo y social.

3. **Entrenamiento del modelo:** Se entrenó el modelo utilizando la configuración seleccionada, aplicando el algoritmo *Global-best PSO* para la optimización de la matriz de pesos.

4. **Evaluación del rendimiento:** El modelo fue evaluado utilizando las métricas definidas. En caso de obtener resultados insatisfactorios, se ajustaron los hiperparámetros y se repitió el proceso de entrenamiento, iterando hasta encontrar el modelo con el mejor rendimiento.

La herramienta *FCM Experts* permite construir la estructura de un MCD definiendo las relaciones entre conceptos de forma aleatoria. Esto se realiza mediante la selección de un porcentaje entre 0 y 100, que indica cuántas relaciones se desean establecer entre todas las posibles. Un valor de 0 genera un MCD sin relaciones entre conceptos, mientras que un valor de 100 produce un MCD completamente conectado, en el cual todos los elementos están relacionados entre sí, incluyendo las relaciones consigo mismos.

Este método presenta una limitación significativa, ya que la generación aleatoria no se fundamenta en métodos basados en evidencia ni en datos observados que justifiquen o validen la selección de las relaciones causales. Para mitigar esta limitación, se intentó aplicar algoritmos basados en poblaciones que, utilizando

los datos disponibles, pudieran inferir la estructura causal del MCD. Sin embargo, dichos métodos no lograron obtener una estructura adecuada ni confiable. Asimismo, se exploraron estrategias iterativas para crear y eliminar relaciones durante la ejecución del modelo, con el fin de optimizar su rendimiento y determinar la mejor estructura posible. Esta alternativa tampoco produjo resultados satisfactorios. Finalmente, se optó por utilizar MCD totalmente conectados. Esta solución, aunque práctica, genera un problema inherente: la propagación de relaciones causales a todos los nodos provoca una disminución en el rendimiento del modelo, tanto en términos computacionales como en la calidad de sus predicciones.

5.3.2 Modelado Basado en Otras Técnicas de IA

Además de los modelos basados en MCD, se emplearon otras técnicas de IA para construir modelos de clasificación, con el objetivo de realizar una comparación tanto en términos de rendimiento como de explicabilidad. Las técnicas de IA usandas fueron seleccionadas utilizando la biblioteca de Python *Optuna* [135], la cual esta especializada en la optimización automática de hiperparámetros, en particular, para algoritmos de aprendizaje automático. Usando a Optuna, el proceso de optimización se estructuró en las siguientes etapas:

1. **Selección de modelos candidatos:** Se evaluaron distintos algoritmos de clasificación mediante validación cruzada de 10 pliegues, {así como utilizando segmentaciones de los conjuntos de datos con proporciones de 20 % - 80 %, 30 % - 70 %, para analizar el desempeño en diferentes particiones de entrenamiento y prueba. Además, en esta primera etapa se utilizó Optuna para una primera búsqueda de los hiperparámetros de los modelos. Con base en su desempeño promedio, se seleccionaron los cinco modelos con mejores resultados para continuar con la optimización.

2. **Definición del espacio de búsqueda:** Para cada modelo seleccionado, se refinó el espacio de búsqueda inicialmente establecido en el paso 1, ajustando con precisión los rangos de los hiperparámetros mediante su ampliación o reducción según fuera necesario. De este modo, se identificaron los hiperparámetros clave y se definieron sus posibles valores, conformando el espacio de búsqueda para la siguiente etapa de optimización.

3. **Exploración del espacio:** A partir de la función objetivo basada en métricas de rendimiento, *Optuna* generó combinaciones de hiperparámetros. En lugar de explorar el espacio al azar, fue dirigiendo la búsqueda hacia aquellas combinaciones que mostraban mayor potencial, basándose en los resultados obtenidos en cada iteración.

4. **Modelado y refinamiento:** Conforme se evaluaban nuevas configuraciones, *Optuna* construía y actualizaba un modelo probabilístico del espacio de búsqueda. Este modelo facilitaba la identificación de combinaciones prometedoras, haciendo el proceso más eficiente y aumentando las probabilidades de encontrar una configuración cercana al óptimo.

El ciclo de optimización continuó hasta cumplir un criterio de parada predefinido, como alcanzar un número máximo de evaluaciones o estabilizar el rendimiento. Gracias a este enfoque, se obtuvieron modelos de referencia con configuraciones bien ajustadas, lo que permitió realizar una comparación justa y rigurosa frente a los modelos desarrollados con MCD.

Las técnicas finales seleccionadas para la comparación fueron las siguientes:

- **Clasificador de Árboles Extra (ETC)** [136]: Técnica de construcción de modelos basada en ensamblados de árboles de decisión, que introduce aleatoriedad en la selección de divisiones en cada nodo para mejorar la generalización.

- **Máquina de Vectores de Soporte (SVM)** [137]: Técnica que construye modelos clasificadores buscando un hiperplano óptimo que maximice el margen de separación entre clases en el espacio de características.
- **Perceptrón Multicapa (MLP)** [138]: Técnica de construcción de modelos basada en redes neuronales artificiales compuestas por múltiples capas de neuronas con funciones de activación no lineales, permitiendo capturar relaciones complejas en los datos.
- **Regresión Logística (LR)** [139]: Técnica estadística para construir modelos que estima la probabilidad de pertenencia a una clase mediante una función logística aplicada a una combinación lineal de variables predictoras.
- **K-Vecinos Más Cercanos (KNN)** [140]: Técnica basada en instancias que construye modelos asignando la clase de una muestra según la mayoría de las etiquetas de sus k vecinos más cercanos en el espacio de características.

5.4 Análisis de Resultados

En esta sección se presentan los resultados obtenidos en términos de rendimiento para los modelos MCD, comparados con aquellos generados mediante otras técnicas de AA, descritas previamente en la Sección 5.3.2.

5.4.1 Conjunto de Datos de Dengue

Para la construcción del modelo MCD sobre el conjunto de datos de dengue, se establecieron los siguientes hiperparámetros: número máximo de iteraciones 200, población de partículas de 65, y coeficientes cognitivo y social en 2.01. La Tabla 5.5 presenta el desempeño de diversos modelos de clasificación evaluados con este conjunto. Los modelos basados en técnicas tradicionales de AA mostraron un rendimiento sobresaliente, con métricas entre 0.9960 y 0.9999. En particular, el modelo ETC alcanzó los valores más altos en todas las métricas (0.9999), evidenciando una capacidad casi perfecta para identificar correctamente los casos. Los modelos LR, MLP y SVM presentaron resultados muy similares, con valores de 0.9998 en todas las métricas. El modelo KNN tuvo un rendimiento ligeramente inferior, pero competitivo, con precisión de 0.9959 y exactitud de 0.9960, demostrando buena capacidad de generalización. En contraste, el modelo MCD obtuvo métricas inferiores a las de los modelos clásicos, posiblemente debido a que los nodos del grafo están completamente conectados, lo que puede limitar la captura de patrones relevantes (ver Sección 5.3.1).

Modelo	Exactitud	Precision	Sensibilidad	Puntuación F1
ETC	0.9999	0.9999	0.9999	0.9999
KNN	0.9960	0.9959	0.9961	0.9960
LR	0.9998	0.9998	0.9998	0.9998
MLP	0.9998	0.9998	0.9998	0.9998
SVM	0.9998	0.9998	0.9998	0.9998
MCD	0.8487	0.8580	0.8470	0.8503

Tabla 5.5: Métricas de rendimiento de los modelos evaluados sobre el conjunto de datos de dengue

La Figura 5.14 muestra las curvas ROC y los valores de Are Under Curve (AUC) obtenidos. Los modelos ETC, LR, SVM y MLP alcanzaron un AUC de 1.0000, reflejando discriminación perfecta. KNN obtuvo un AUC de 0.9988 y MCD 0.8852, lo que confirma la diferencia en desempeño.

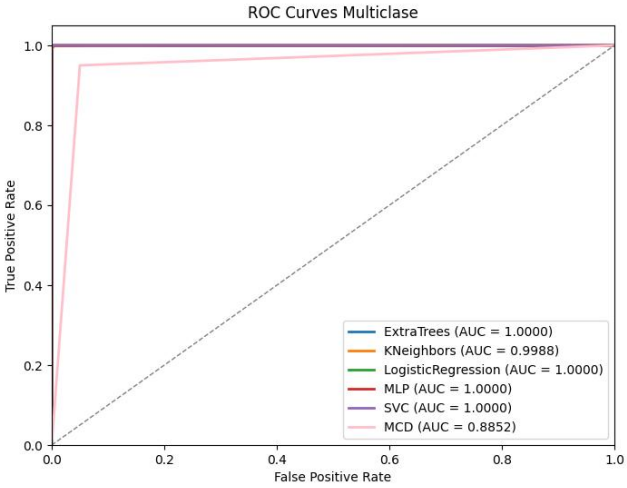

Figura 5.11: Curva ROC para el conjunto de datos de dengue

5.4.2 Conjunto de Datos de COVID-19

Para el modelo MCD construido sobre el conjunto de datos de COVID-19, se configuraron los siguientes hiperparámetros: número máximo de iteraciones en 100, población de partículas en 30, y coeficientes cognitivo y social en 2.01. La Tabla 5.6 presenta el desempeño de diversos modelos de clasificación aplicados a este conjunto. En general, los modelos convencionales de AA muestran un rendimiento homogéneo y elevado, con métricas cercanas a 0.978 en todas las evaluaciones. Los modelos ETC, LR, MLP y SVM alcanzan valores aproximados de 0.9786 en todas las métrica, indicando capacidad consistente para clasificar correctamente los casos. El modelo KNN exhibe un rendimiento ligeramente inferior (aproximadamente 0.9783), aunque la diferencia no es sustancial. Finalmente, el modelo MCD presenta una capacidad de clasificaciónn menor en comparación con las técnicas tradicionales, con una *exactitud* de 0.9357, una *puntuación F1* y *sensibilidad* de 0.9305

Modelo	Exactitud	Precision	Sensibilidad	Puntuación F1
ETC	0.9786	0.9795	0.9786	0.9786
KNN	0.9783	0.9791	0.9783	0.9783
LR	0.9786	0.9795	0.9786	0.9786
MLP	0.9786	0.9795	0.9786	0.9786
SVM	0.9786	0.9795	0.9786	0.9786
MCD	0.9357	0.9450	0.9305	0.9305

Tabla 5.6: Métricas de rendimiento de los modelos evaluados sobre el conjunto de datos de COVID-19

La Figura 5.12 muestra las curvas ROC y valores de AUC correspondientes. Los modelos tradicionales presentan curvas y valores muy similares, reflejando alto rendimiento. El modelo MCD se encuentra ligeramente por debajo, indicando menor capacidad discriminativa.

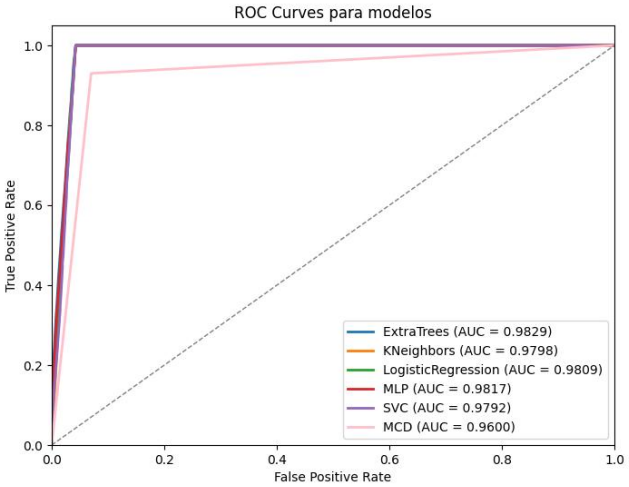

Figura 5.12: Curva ROC sobre el conjunto de datos de COVID-19

5.4.3 Conjunto de Datos de Diabetes

Para este conjunto, el modelo MCD se configuró con un límite máximo de 100 iteraciones y población de partículas de tamaño 40, con coeficientes cognitivo y social en 2.01. La Tabla 5.7 presenta las métricas obtenidas. Los modelos ETC y KNN mostraron mejor desempeño. Los modelos MLP, SVM y LR presentan rendimiento intermedio, mientras que MCD evidenció desempeño menor, con una *exactitud* de 0.7433 y valores similares en las demás métricas.

Modelo	Exactitud	Precision	Sensibilidad	Puntuación F1
ETC	0.8567	0.8653	0.8567	0.8558
KNN	0.8167	0.8349	0.8167	0.8141
RL	0.7900	0.8160	0.7900	0.7856
MLP	0.8300	0.8354	0.8300	0.8293
SVM	0.8433	0.8489	0.8433	0.8427
MCD	0.7433	0.7450	0.7450	0.7400

Tabla 5.7: Métricas de rendimiento de los modelos evaluados sobre el conjunto de datos de diabetes

La Figura 5.13 muestra las curvas ROC y los valores de AUC correspondientes a los modelos evaluados sobre el conjunto de datos de diabetes. Se observa que los modelos ETC y KNN presentan curvas muy similares, con valores de AUC de 0.9020 y 0.9058, respectivamente, reflejando una alta capacidad para distinguir entre clases. Los modelos MLP y SVM muestran un desempeño ligeramente inferior, con AUC cercanos a 0.8899 y 0.8876, respectivamente. Por otro lado, LR alcanza un valor de AUC de 0.8484, mientras que el modelo MCD presenta el menor valor, aproximadamente 0.8075.

Cabe destacar que las curvas ROC exhiben algunos escalones, característica típica cuando se trabaja con conjuntos de datos de tamaño reducido, ya que el número limitado de muestras afecta la cantidad de posibles puntos de corte para calcular las tasas de verdaderos y falsos positivos. Esta particularidad puede generar una apariencia discontinua en las curvas, sin que ello afecte la validez de la evaluación comparativa entre los modelos.

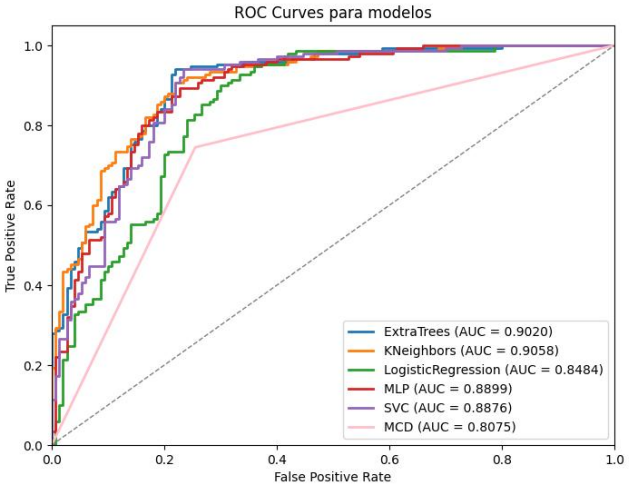

Figura 5.13: Curva ROC dengue sobre el conjunto de datos de diabetes

5.4.4 Conjunto de Datos de Diagnóstico de Vehículos Submarinos

Para el conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos, se emplearon los mismos parámetros del modelo MCD que en los experimentos previos, con 150 iteraciones máximas, población de partículas de tamaño 50, y coeficientes cognitivo y social de 2.01. La Tabla 5.8 presenta las métricas de rendimiento de los modelos evaluados. El modelo ETC alcanzó la mayor exactitud con un valor de 0.9229, mostrando, además, valores idénticos en *precision*, *sensibilidad* y *puntuación F1*, lo que indica un comportamiento equilibrado en todas las métricas. El modelo KNN obtuvo un rendimiento intermedio con una exactitud de 0.8119 y métricas similares. Por otro lado, los modelos MLP y SVM mostraron resultados moderados, con exactitudes de 0.8105 y 0.7633, respectivamente. El modelo LR presentó una eficacia menor, con una *exactitud* de 0.7295. Finalmente, el modelo MCD registró resultados comparables a los de SVM, con una *exactitud* de 0.7610 y el resto de métricas con valores cercanos.

Modelo	Exactitud	Precision	Sensibilidad	Puntuación F1
ETC	0.9229	0.9229	0.9229	0.9229
KNN	0.8119	0.8174	0.8126	0.8113
RL	0.7295	0.7297	0.7297	0.7295
MLP	0.8105	0.8116	0.8108	0.8104
SVM	0.7633	0.7649	0.7637	0.7631
MCD	0.7610	0.7700	0.7600	0.7600

Tabla 5.8: Métricas de rendimiento de los modelos evaluados sobre el conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos

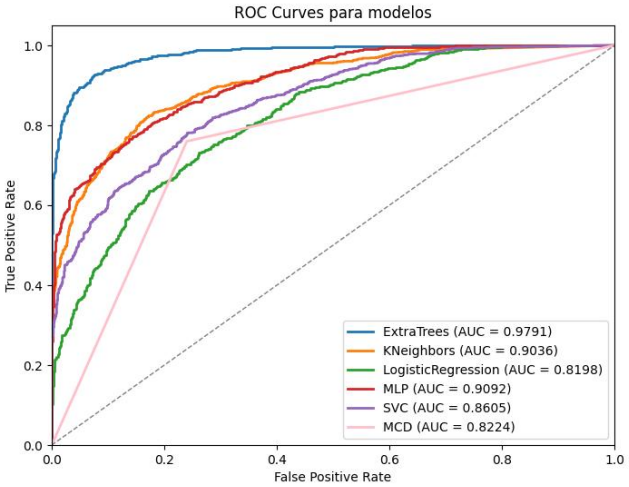

Figura 5.14: Curva ROC dengue sobre el conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos

La Figura 5.14 muestra las curvas ROC y los valores de AUC para los distintos modelos aplicados al conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos. Se observan algunos escalones en las curvas, característicos de conjuntos de datos con tamaño limitado, debido a la cantidad restringida de muestras y posibles puntos de corte.

En cuanto a la capacidad discriminativa, el modelo ETC presenta el valor más alto de AUC con 0.9791, seguido por MLP con 0.9092, KNN con 0.9026, y SVM con 0.8605. Los modelos LR y MCD alcanzan valores de AUC de 0.8198 y 0.8242, respectivamente, reflejando una capacidad moderada para distinguir entre clases en este escenario.

5.5 Analisis de Explicabilidad

Esta sección hace un análisis de la calidad de la explicabilidad aportada por nuestro método. Para ello, se definirán otros métodos de explicabilidad para comparar, se hará un análisis del ranking de variables aportada por cada método, un análisis de sensibilidad de la degradación del rendimiento de los modelos según las variables relevantes de cada método, y finalmente, para nuestro método de explicabilidad, se determina su comportamiento en las propiedades de explicabilidad comentadas en las secciones anteriores.

5.5.1 Métodos de Explicabilidad de Referencia

Los resultados de nuestro método de explicabilidad se contrastaron con los obtenidos mediante dos métodos de explicabilidad clasicos que pertenecen a diferentes categorias, **SHAP** y **FP**, los cuales se describen brevemente a continuación:

- **SHAP**: Este método se basa en los valores de Shapley de la teoría de juegos cooperativos. Asigna a cada característica una contribución al resultado del modelo considerando todas las combinaciones posibles de variables. Fue introducido por Lundberg y Lee [4] como un enfoque unificado para interpretar las predicciones de modelos complejos, y se ha consolidado como una de las técnicas más populares para la explicabilidad local.

- **FP**: Método que evalúa la importancia de una característica midiendo el aumento en el error del modelo cuando se permutan aleatoriamente sus valores. Si dicha permutación afecta considerablemente el rendimiento, se considera que la característica tiene un alto impacto. Esta técnica fue propuesta por Altmann y otros [141] como una alternativa eficaz de modelo-agnóstico para estimar la relevancia de variables.

5.5.2 Resultados de Explicabilidad

En esta sección se realiza un análisis comparativo entre los tres métodos de explicabilidad usando las medidas de centralidad de grafos presentadas en la sección 2.3.5.1 y el ranking de relevancia propuesto por cada método. Se realizaron n simulaciones para el método propuesto y el método SHAP, ambos de naturaleza local y dependientes de instancias específicas de entrada. El objetivo fue recopilar una cantidad suficiente de resultados que permitiera comparar de manera fiable su comportamiento promedio bajo condiciones controladas y equitativas. Se fijó el valor de n en 20 simulaciones, lo que se consideró adecuado para obtener resultados representativos y estables. Para cada simulación, se utilizaron diferentes instancias de entrada, y posteriormente se promediaron los resultados obtenidos con cada método, lo que permitió una comparación objetiva y cuantitativa de su desempeño relativo. El análisis agregado a partir de múltiples ejecuciones permite obtener una estimación útil del comportamiento general de cada enfoque. Por otro lado, en el caso de la FP, que es un método global, no fue necesario realizar múltiples simulaciones, ya que este enfoque proporciona resultados globales por definición y no depende de instancias específicas de entrada.

5.5.2.1 Conjunto de Datos de Dengue

La Tabla 5.9 muestra la correspondencia entre el nombre de cada variable y su concepto C dentro del MCD asociado, con el objetivo de facilitar la interpretación de los resultados obtenidos. Los resultados del método propuesto para MCD se muestran en la Figura 5.15, como también, los derivados de las medidas de centralidad de grafos, representados mediante gráficos de araña, donde la variable más relevante se indica en azul, y las siguientes en orden decreciente en sentido contrario a las agujas de reloj. Se observa que nuestro método asigna mayor importancia a los conceptos *C2*, seguido de *C9* y *C3* (ver Dynax-FCM en Fig. 5.15). Las medidas de centralidad de grado de entrada, grado total y PageRank coinciden en que el concepto más relevante es *C1*. En el caso del grado de entrada y grado total, el segundo lugar lo ocupa *C13*, mientras que en grado de salidad el valor más alto corresponde a *C16*. Por su parte, la medida de intermediación ofrece resultados distintos, destacando como más relevantes los conceptos *C14*, y posteriormente *C20*. Se observa que, en todas las medidas excepto en *PageRank*, la importancia de los conceptos decrece de manera sostenida. En contraste, en PageRank la relevancia se concentra principalmente en *C1*, mientras que el resto de los conceptos presentan valores considerablemente inferiores.

Se concluye que las distintas métricas de centralidad basadas en grafos no coinciden plenamente en la identificación de los conceptos más relevantes dentro del MCD. Mientras algunas destacan a *C1*, otras asignan mayor importancia a conceptos como *C13*, *C14* o *C20*, lo que evidencia que cada métrica capta diferentes aspectos de la estructura del sistema. Esta variabilidad impide establecer una jerarquía única basada únicamente en estas medidas, lo que refuerza la pertinencia de emplear enfoques complementarios, como el método propuesto en este trabajo.

Variable	Nombre	Variable	Nombre	Variable	Nombre	Variable	Nombre
C1	Edad	C6	Erupción	C11	Hepatomegalia	C16	Acumulación de Líquidos
C2	Cefalea	C7	Dolor Abdominal	C12	Hemorragias en Mucosas	C17	Extravasación
C3	Dolor Retroocular	C8	Vómito	C13	Hipotermia	C18	Hemorragias Hemáticas
C4	Mialgias	C9	Somnolencia	C14	Aumento de Hematocrito	C19	Shock
C5	Artralgia	C10	Hipotensión	C15	Caída de Plaquetas	C20	Daño Orgánico

Tabla 5.9: Relación entre concepto y nombre de concepto en el conjunto de datos de COVID-19

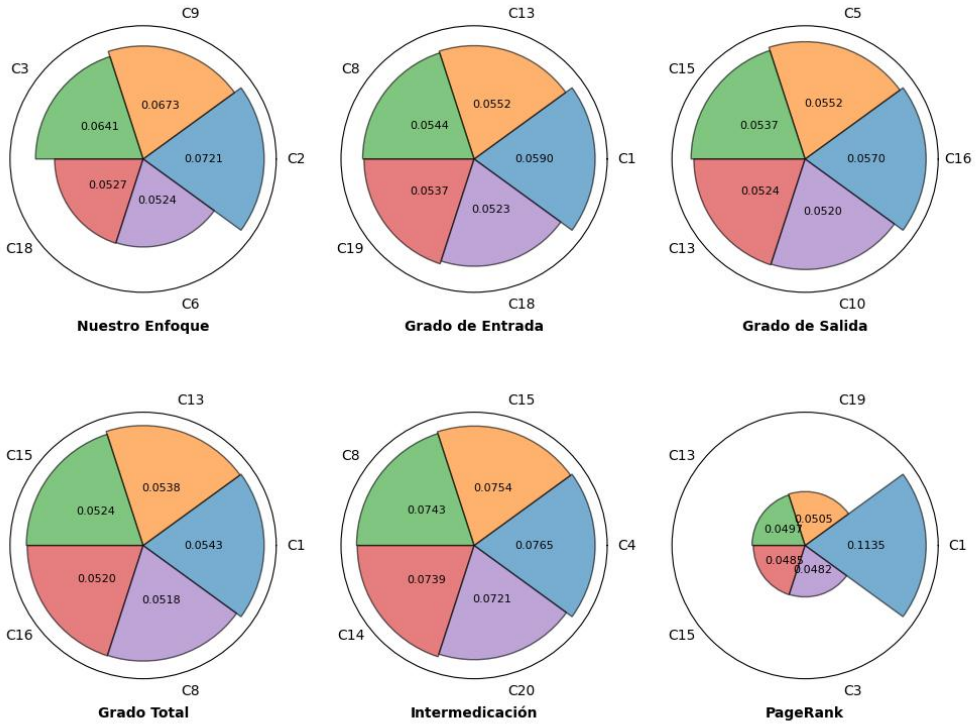

Figura 5.15: Importancia del método propuesto y medidas de centralidad de grafos en el conjunto de datos dengue

La Figura 5.16 presenta el ranking de relevancia propuesto por cada método de explicabilidad. Se observa que los modelos basados en SHAP asignan la mayor relevancia al concepto *C7*, correspondiente a la variable *Dolor Abdominal*. Le siguen, en distintas posiciones según el modelo, los conceptos *C8*, *C15* y *C12*. Un patrón similar se observa en el método FP, que también sitúa a *C7* como la variable más relevante, seguida por *C8*, *C15* y *C12*, aunque con diferencias en el orden de importancia entre modelos.

En contraste, el método propuesto establece un patrón de relevancia distinto, identificando como conceptos más importantes a *C2*, *C9* y *C3*, correspondientes a *Cefalea*, *Somnolencia* y *Dolor retroocular*, respectivamente. Esta divergencia sugiere que nuestro método es capaz de capturar dinámicas internas y relaciones causales desapercibidas para métodos de explicabilidad como SHAP o FP.

Como se detalló en la Sección 5.1.2.1, tanto la Figura 5.2, que presenta las correlaciones de Cramér, como el análisis de VIF, identifican a la variable *Dolor Abdominal*, correspondiente a *C7*, como relevante pero propensa a alta colinealidad. El hecho de que los métodos SHAP y FP sitúen esta variable como la más relevante, podría una limitación por la presencia de *multicolinealidad*. En presencia de *multicolinealidad*, las variables correlacionadas contienen información redundante que contribuye de manera similar a la predicción del modelo. Esta redundancia dificulta que los métodos de interpretación

cuantifiquen de forma aislada la contribución individual de cada variable, ya que sus efectos se solapan en el espacio de características. Por lo tanto, en lugar de distribuir correctamente la importancia entre todas las variables correlacionadas, dichos métodos tienden a asignar una proporción desproporcionada de la importancia a una sola variable, lo que conduce a una sobreestimación de su influencia real. En consecuencia, esta asignación sesgada distorsiona la evaluación de la relevancia de las variables, generando la falsa impresión de que una variable tiene un impacto mayor del que posee en realidad, cuando en efecto la contribución relevante está compartida entre múltiples variables interrelacionadas.

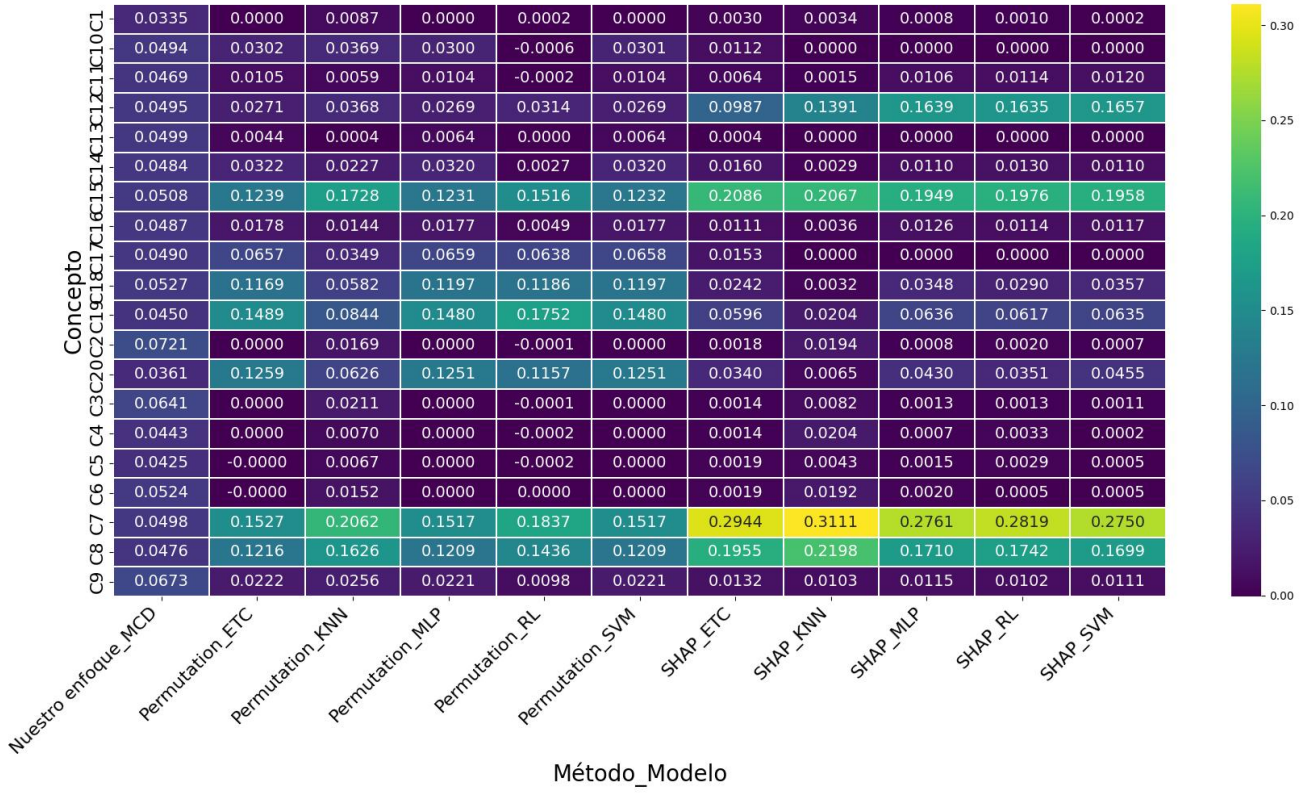

Figura 5.16: Importancia del método propuesto y métodos SHAP y FP en el conjunto de datos dengue

5.5.2.2 Conjunto de Datos de COVID-19

La Tabla 5.10 muestra la correspondencia entre el nombre de cada variable y su concepto C en los modelos construidos con el conjunto de datos de COVID-19, con el objetivo de facilitar la interpretación de los resultados obtenidos. La Figura 5.17 presenta una comparación entre el método propuesto y las medidas de centralidad basadas en grafos. Se observa que el método propuesto asigna mayor importancia al concepto *C4*. En cuanto a las medidas de centralidad, todas, excepto *grado de salida*, identifican al concepto *C1* como el más relevante. No obstante, el orden de importancia del resto de los conceptos varía entre métodos, aunque los conceptos *C3* y *C9* aparecen de forma recurrente en los distintos rankings, incluido el método propuesto. Por su parte, la medida de *grado de salida* difiere del resto al situar como más relevante al concepto *C2*.

En conclusión, aunque el método propuesto y las medidas de centralidad basadas en grafos coinciden parcialmente en la identificación de conceptos clave, como *C1*, *C3* y *C9*, también evidencian diferencias significativas en la jerarquía de importancia asignada a cada concepto.

Concepto	Nombre
C1	Tos
C2	Fiebre
C3	Dolor de garganta
C4	Dificultad para respirar
C5	Dolor de cabeza
C6	Edad 60 o más
C7	Género
C8	Motivo contacto con infectado
C9	Motivo viaje al extranjero

Tabla 5.10: Correspondencia entre códigos de concepto y nombres clínicos en el conjunto de datos de COVID-19.

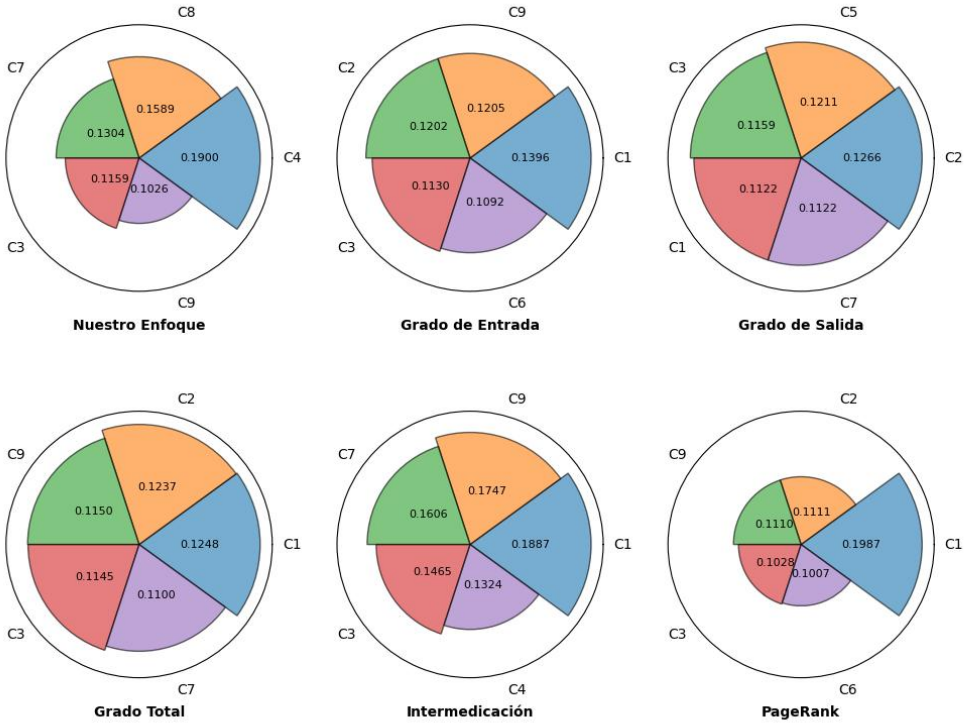

Figura 5.17: Comparación de la importancia de los conceptos según el método propuesto y distintas medidas de centralidad en grafos en el conjunto de datos de COVID-19.

La Figura 5.18 presenta el ranking de relevancia propuesto por cada método de explicabilidad. En relación con las explicaciones generadas por SHAP y FP, se observa que los modelos basados en SHAP identifican como variable más importante a *C8*, seguida de *C2* o *C5*, dependiendo del modelo. En el caso de FP, la variable más relevante corresponde a *C2*, seguida por *C5* o *C8*, según el modelo considerado.

Por su parte, nuestro enfoque propuesto destaca a los conceptos *C4*, *C7* y *C8*, siendo únicamente *C8* considerada importante por los otros métodos. Esto sugiere que ni SHAP ni FP son capaces de capturar adecuadamente las relaciones de causalidad al calcular la importancia de las variables.

Cabe destacar que tanto *C8* como *C2*, correspondientes a las variables *Fiebre* y *Motivo: contacto con infectado*, respectivamente, fueron identificadas en la sección de preparación del conjunto de datos como variables con alta correlación según el coeficiente de Cramér y con presencia de multicolinealidad. Este hecho pone nuevamente de manifiesto una limitación inherente a los métodos SHAP y FP, que tienden a sobreestimar la importancia de variables cuando existe multicolinealidad en el conjunto de datos.

Figura 5.18: Commportancia del método propuesto y métodos SHAP y FP en el conjunto de datos COVID-19

5.5.2.3 Conjunto de Datos de Diabetes

La Tabla 5.11 presenta la correspondencia entre el nombre de cada variable y su concepto C en los modelos construidos con el conjunto de datos diabetes. La figura 5.19 presenta los resultados de la comparación entre el método propuesto y las distintas medidas de centralidad basadas en grafos. Se observa que el método propuesto identifica al concepto *C6* como el más relevante, seguido por *C2* y *C8*. De forma consistente, las medidas de *grado de salida*, *grado total* e *intermediación* también consideran a *C6* como el concepto más importante. En cuanto al segundo puesto, tanto el *grado de salida* como el *grado total* coinciden con el método propuesto al destacar a *C2*, mientras que la medida de *intermediación* difiere, situando a *C1* como la segunda más relevante y relegando a *C2* al quinto lugar. Por otro lado, las medidas restantes presentan discrepancias más notables. El *grado de entrada* considera a *C2* como la más importante, seguida de *C5*, y ubica a *C6* en la tercera posición. En el caso de *PageRank*, *C2* vuelve a ocupar el primer lugar, pero *C6* desciende nuevamente a la tercera posición, a pesar de haber sido la más destacada en otras métricas. En general, puede observarse que los conceptos *C6* y *C2* son considerados entre los tres más importantes en la mayoría de las medidas, con la única excepción de la *intermediación*, donde *C2* desciende hasta la quinta posición en el ranking.

Concepto	Variable
C1	Número de Embarazos
C2	Glucosa
C3	Presión Sanguínea
C4	Grosor Cutáneo
C5	Insulina
C6	BMI
C7	DPF
C8	Edad

Tabla 5.11: Correspondencia entre códigos de concepto y nombres clínicos en el conjunto de datos de diabetes.

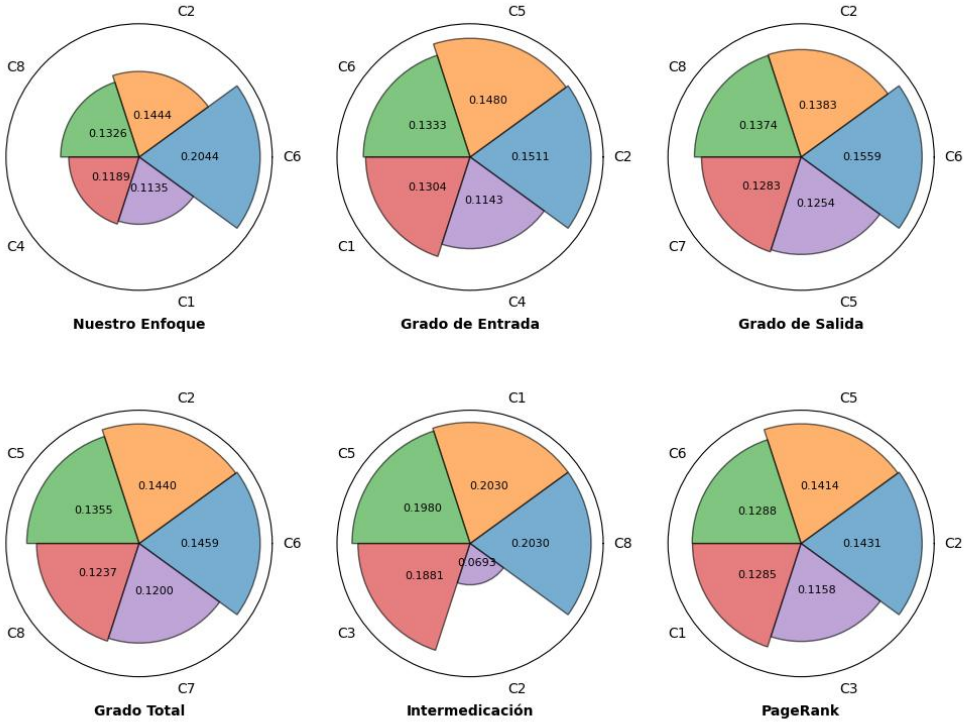

Figura 5.19: Comparación de la importancia de los conceptos según el método propuesto y distintas medidas de centralidad en grafos en el conjunto de datos de diabetes.

La Figura 5.20 presenta el ranking de relevancia propuesto por cada método de explicabilidad. Se observa que tanto *SHAP* como *FP* coinciden en identificar al concepto *C2* como el más relevante. Este es seguido en el ranking por los conceptos *C8*, *C6* o *C1*, según el modelo considerado. Cabe destacar que *C2* es la segunda variable con mayor multicolinealidad, seguida de *C6*, la cual es señalada por nuestro método como la más importante.

Figura 5.20: Commportancia del método propuesto y métodos SHAP y FP en el conjunto de datos diabetes.

5.5.2.4 Conjunto de Datos de Diagnóstico de Fallos en Vehículos Submarinos Autónomos

La Tabla 5.12 presenta la correspondencia entre el nombre de cada variable y su concepto *C* en los modelos construidos con el conjunto de datos de diabetes. La Figura 5.21 muestra los resultados de la comparación entre el método propuesto y las distintas medidas de centralidad basadas en grafos. Se observa que el método propuesto identifica al concepto *C7* como el más relevante, seguido por *C3* y *C6*. Las medidas de *grado de salida* e *intermediación* seleccionan como más importante al concepto *C4*, seguido de *C6* y *C2* en diferente orden. Por otro lado, las medidas de *grado de entrada*, *grado total* y *PageRank* consideran al concepto *C1* como el más relevante, seguido de *C4*. En este caso, el método propuesto no coincide con ninguna de las medidas de centralidad evaluadas.

Concepto	Nombre
C1	PWM
C2	Voltaje (V)
C3	Presión (Pa)
C4	Ángulo de Inclinación (°)
C5	Profundidad (m)
C6	Ángulo de Rodar (°)
C7	Velocidad Angular de Guinada (°/s)

Tabla 5.12: Correspondencia entre códigos de concepto y nombres en el conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos

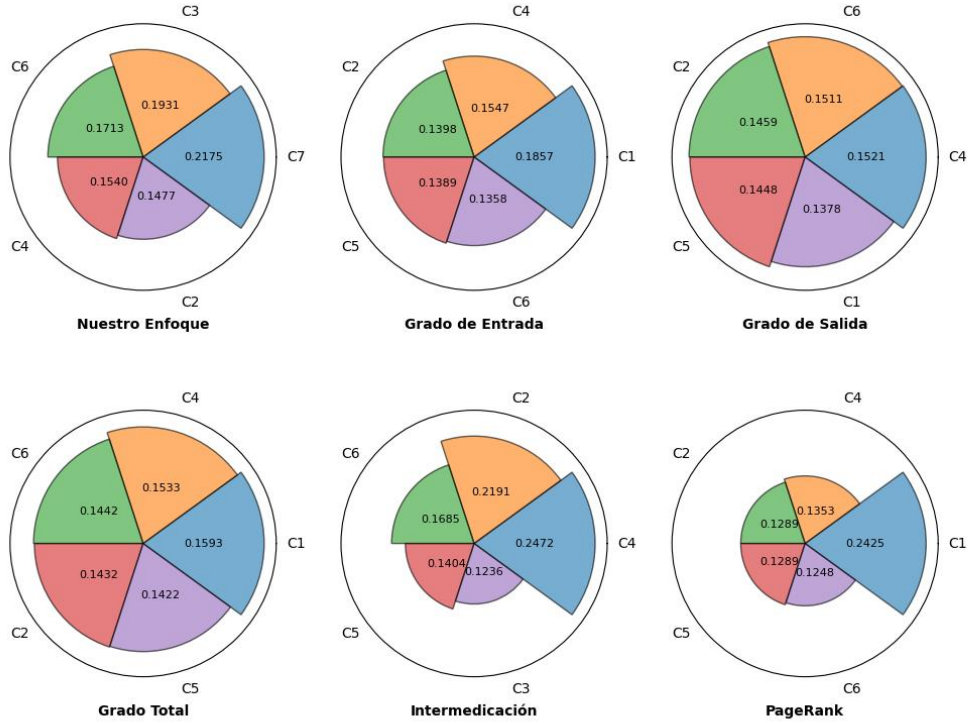

Figura 5.21: Comparación de la importancia de los conceptos según el método propuesto y distintas medidas de centralidad en grafos en el conjunto de diagnóstico de fallos en vehículos submarinos autónomos.

La Figura 5.22 presenta el ranking de relevancia propuesto por cada método de explicabilidad. En relación con los métodos de explicabilidad SHAP y FP, en este caso no se identifica una tendencia clara, ya que cada modelo considera distintas variables como las más importantes. En particular, FP señala a los conceptos *C5*, *C3* y *C1* como los más relevantes, sin que exista un claro consenso entre ellos. Por su parte, SHAP destaca a los conceptos *C3*, *C2* y *C4* como los de mayor importancia, lo que refuerza la ausencia de una tendencia dominante entre los modelos evaluados.

Una vez examinado el comportamiento en los cuatro conjuntos de datos, se observa que los resultados derivados de las medidas de centralidad en grafos difieren notablemente de los obtenidos con el método propuesto. Además, estas medidas de centralidad presentan discrepancias entre sí, reflejando diferentes criterios y jerarquías en la identificación de los conceptos más relevantes, lo que evidencia la complejidad de capturar la verdadera importancia dentro de redes complejas.

De manera similar, los resultados ofrecidos por SHAP y FP también presentan diferencias significativas respecto al método propuesto. Esto se explica principalmente porque nuestro enfoque está diseñado para capturar la dinámica de las relaciones causales entre variables, mientras que SHAP y FP evalúan la importancia basándose fundamentalmente en asociaciones estadísticas directas. Además, el análisis exhaustivo realizado en todos los conjuntos de datos y modelos reveló que ambos métodos tienden a favorecer variables con alta correlación o multicolinealidad, lo que puede llevar a interpretaciones sesgadas o erróneas. En contraste, el método propuesto permite identificar de forma más robusta y precisa las variables verdaderamente relevantes, al considerar las interacciones causales subyacentes y la estructura dinámica del sistema, mejorando así la calidad, coherencia y utilidad de las explicaciones generadas.

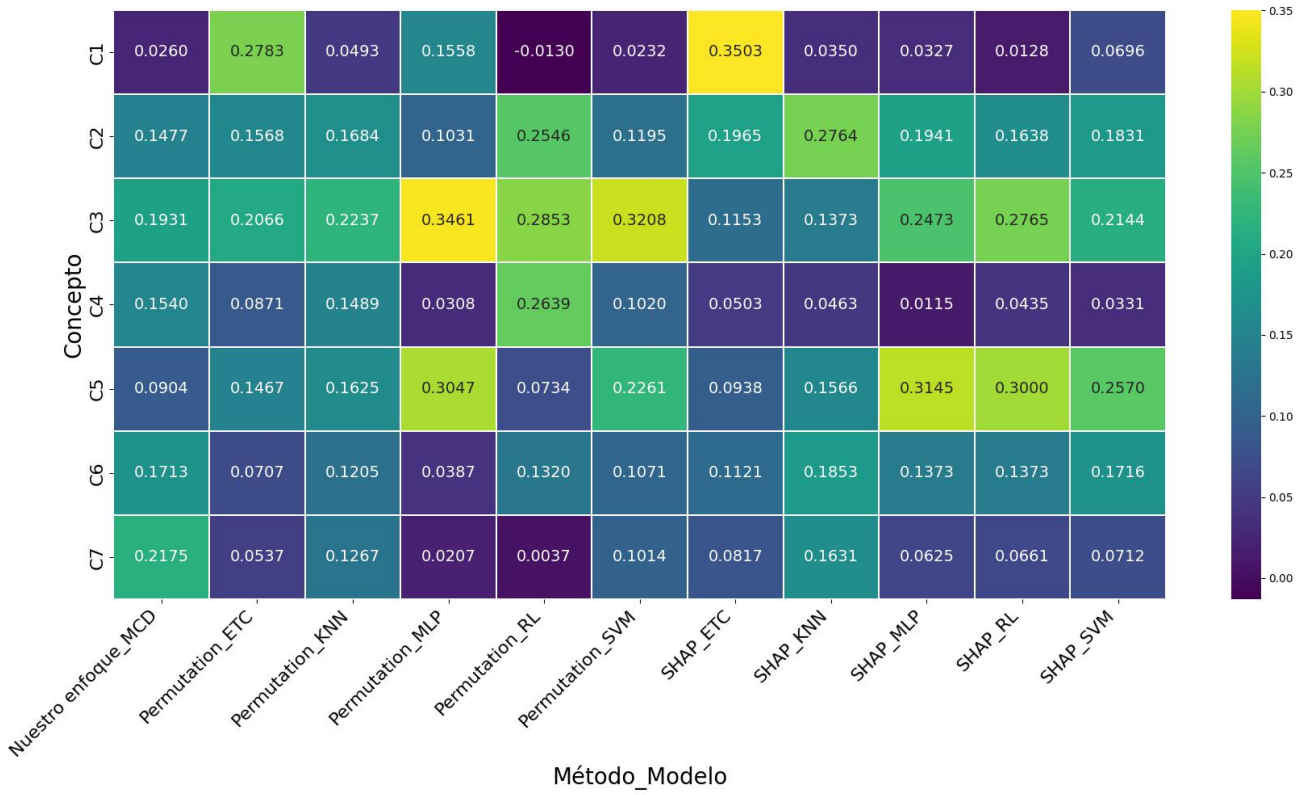

Figura 5.22: Commportancia del método propuesto y métodos SHAP y FP en el conjunto de diagnóstico de fallos en vehículos submarinos autónomos.

5.5.3 Comparación de la Calidad de los Métodos de Explicabilidad

Después de analizar los resultados obtenidos mediante el método propuesto, así como los proporcionados por SHAP y FP, se propone la aplicación de ROAR con el objetivo de evaluar la calidad de las explicaciones generadas. Se pretende comprobar que las explicaciones generadas por el método propuesto presentan una mayor calidad que las ofrecidas por SHAP y FP, las cuales pueden verse afectadas por problemas de *multicolinealidad*.

La tabla 5.13 presenta los índices de degradación calculados tras eliminar la primera variable relevante, las dos primeras variables revelantes, y las tres primeras variables relevantes, en el conjunto de datos de dengue según cada método de explicabilidad. Se observa que la mayor degradación se produce en los MCD, alcanzando un valor de 0.2850 al eliminar las tres variables mas relevantes. SHAP y FP presentan degradaciones similares para algunos modelos, e incluso mejores, cuando se eliminan una o dos variables (como el caso de SHAP y el modelo KNN al eliminar dos variables). A pesar de ello, el método propuesto muestra consistentemente la mayor degradación, lo que sugiere una mayor sensibilidad a las variables eliminadas y, por tanto, una identificación más precisa de las características relevantes

Vars elim.	**Método propuesto**	**SHAP**					**FP**				
	MCD	**ETC**	**KNN**	**LR**	**MLP**	**SVM**	**ETC**	**KNN**	**LR**	**MLP**	**SVM**
1	0.0625	0.0626	0.0621	0.0584	0.0627	0.0624	0.0626	0.0621	0.0584	0.0627	0.0624
2	0.1425	0.1293	0.1834	0.1241	0.1292	0.1292	0.0966	0.1316	0.1399	0.1012	0.1007
3	0.2850	0.2653	0.2711	0.2589	0.2699	0.2625	0.1210	0.2711	0.2052	0.1305	0.1285

Tabla 5.13: Índice de degradación basado en el *accuracy* al eliminar las variables más importantes en el conjunto de datos de dengue

La tabla 5.14 presenta los índices de degradación calculados tras eliminar la primera variable relevante, y las dos y tres primeras variables revelantes en el conjunto de datos de COVID-19 según cada método de explicabilidad. Se observa que la mayor degradación se produce en los MCD, con un valor de 0.14 al utilizar el método propuesto y eliminar las tres variables mas relevantes, seguida de LR con FP (0.1287) y de KNN con FP (0.1147). De nuevo, el método propuesto muestra una mayor capacidad para identificar las variables realmente relevantes.

Vars elim.	**Método propuesto**	**SHAP**					**FP**				
	MCD	**ETC**	**KNN**	**LR**	**MLP**	**SVM**	**ETC**	**KNN**	**LR**	**MLP**	**SVM**
1	0.02	0.0007	0.0007	-0.0010	0.0007	0.0007	0.0587	0.0710	0.0623	0.0570	0.0563
2	0.10	0.0587	0.0340	0.0363	0.0227	0.0217	0.0587	0.1137	0.0780	0.0577	0.0570
3	0.14	0.0767	0.0337	0.0373	0.0810	0.0370	0.0767	0.1147	0.1287	0.0777	0.0770

Tabla 5.14: Índice de degradación basado en el *accuracy* al eliminar las variables más importantes en el conjunto de datos de COVID-19

La tabla 5.15 contiene los índices de degradación del conjunto de datos de diabetes, y muestra que el mayor valor se alcanza con el método propuesto al eliminar las tres variables más relevantes, con una degradación de 0.2034. Incluso, al eliminar las dos variables mas relevantes sigue siendo mejor nuestro enfoque. En el caso de la eliminación de una variable tiene un comportamiento similar al resto.

Vars elim.	**Método propuesto**	**SHAP**					**FP**				
	MCD	**ETC**	**KNN**	**LR**	**MLP**	**SVM**	**ETC**	**KNN**	**LR**	**MLP**	**SVM**
1	0.0523	0.0400	0.0167	0.0567	0.0667	0.0567	0.0400	0.0167	0.0567	0.0667	0.0567
2	0.1491	0.0900	0.0400	0.0633	0.0767	0.0467	0.0900	0.0467	0.0633	0.1200	0.0967
3	0.2034	0.0900	0.0767	0.0700	0.1167	0.0967	0.0833	0.0767	0.0933	0.1167	0.0967

Tabla 5.15: Índice de degradación basado en el *accuracy* al eliminar las variables más importantes en el conjunto de datos de diabetes.

Por último, en el conjunto de datos de diagnóstico de fallos en vehículos submarinos autónomos, la Tabla 5.16 presenta los resultados de los índices de degradación. Se observa nuevamente que el método propuesto produce la mayor degradación en el rendimiento al eliminar tres variables, en comparación con el resto de métodos. No obstante, tanto SHAP como FP en ETC ofrecen un valor de degradación cercano. En el resto de los casos, los índices de degradación son muy parecidos y significativamente menores.

Vars elim.	**Método propuesto**	**SHAP**					**FP**				
	MCD	**ETC**	**KNN**	**LR**	**MLP**	**SVM**	**ETC**	**KNN**	**LR**	**MLP**	**SVM**
1	0.0929	0.0929	0.0281	0.0200	0.0738	0.0176	0.0929	0.0548	0.0443	0.1057	0.0410
2	0.1471	0.1314	0.0329	0.0419	0.1471	0.0538	0.1510	0.1005	0.0814	0.1471	0.0538
3	0.2365	0.2176	0.0211	0.0676	0.1700	0.0610	0.2176	0.1186	0.0986	0.1619	0.0610

Tabla 5.16: Índice de degradación basado en el *accuracy* al eliminar las variables más importantes en el conjunto de datos diagnóstico de fallos en vehículos submarinos autónomos.

En todos los conjuntos de datos explorados, la métrica de calidad ROAR aplicado al método propuesto muestra consistentemente los mayores índices de degradación cuando se eliminan las tres primeras variables identificadas como más importantes. Este comportamiento indica que las variables seleccionadas por el método propuesto son efectivamente relevantes para el modelo, ya que su eliminación provoca un deterioro significativo en el rendimiento. En comparación, los métodos basados en SHAP y FP muestran degradaciones más moderadas o localizadas en modelos concretos, lo que sugiere que nuestro método propuesto tiene una mayor capacidad para detectar información verdaderamente esencial. Por tanto, los resultados respaldan la utilidad del enfoque propuesto para tareas de selección de características, especialmente en entornos críticos donde una identificación precisa de las variables clave es fundamental.

5.5.4 Análisis de las Propiedades de Explicabilidad en Nuestro Método de Explicabilidad

Se evaluó la calidad del método de explicabilidad propuesto utilizando las propiedades definidas en la Sección 2.3.3, las cuales permiten medir su calidad, utilidad y confiabilidad. Estas propiedades incluyen: *fidelidad*, *estabilidad*, *uniformidad*, *robustez* y *eficiencia*.

Para el cálculo de dichas propiedades, se emplearon $n = 10$ instancias distintas, sobre las cuales se aplicó el procedimiento de evaluación. El valor de n se limitó a 10 debido a que cada instancia requiere una verificación manual exhaustiva. Los resultados presentados en la Tabla 5.17 reflejan el desempeño del método de explicabilidad propuesto en cuatro dominios distintos: Dengue, COVID-19, Diabetes y diagnóstico de fallos en VAS. En general, el método muestra un comportamiento sólido en todas las propiedades evaluadas, lo que respalda su aplicabilidad en diversos contextos.

En cuanto a la **fidelidad**, se observa un rendimiento consistentemente alto en todos los conjuntos, con valores que oscilan entre 0.9162 y 0.9533. Esto indica que las explicaciones generadas son coherentes con el comportamiento del modelo de clasificació subyacente, representando adecuadamente los factores que influyen en sus decisiones.

La **consistencia** se evaluó mediante dos métricas complementarias: *estabilidad* y *uniformidad*. Respecto a la *estabilidad*, se obtuvo un valor nulo (0.0) en todos los dominios. Este resultado se debe a la naturaleza determinista del método, basado en las relaciones causales directas e indirectas codificadas en el MCD, las cuales no varían al ejecutar múltiples veces el proceso de explicabilidad sobre una misma instancia. Como consecuencia, tanto las activaciones como las explicaciones generadas permanecen invariantes, garantizando una estabilidad total entre ejecuciones. Por su parte, la *uniformidad* alcanzó valores elevados en todos los conjuntos (entre 0.9397 y 0.9911), lo que indica que las relevancias asignadas a las distintas características están distribuidas de manera balanceada. Es decir, el método no concentra toda la importancia explicativa en unas pocas variables, sino que reconoce el aporte de múltiples características en la decisión del modelo.

En relación con la **robustez**, los valores obtenidos se encuentran en un rango de 0.7231 a 0.7561. Estos resultados sugieren que las explicaciones son razonablemente estables frente a perturbaciones pequeñas en las entradas. Aunque el método responde a las modificaciones en los datos, mantiene una consistencia suficiente como para considerarse robusto. Cabe señalar que esta propiedad puede estar parcialmente influenciada por la sensibilidad inherente del modelo de clasificación ante dichas perturbaciones.

Finalmente, respecto a la **eficiencia**, se observan diferencias notables entre los dominios. El conjunto de Diabetes presenta el menor valor de C_s (0.0013), lo cual refleja una generación de explicaciones altamente eficiente en ese contexto. En contraste, los conjuntos de Dengue y COVID-19 muestran valores más elevados (0.1167 y 0.1072, respectivamente), posiblemente debido a una mayor complejidad estructural del modelo o del grafo causal utilizado. No obstante, todos los valores se mantienen dentro de márgenes aceptables, lo que confirma que el método es computacionalmente viable, incluso en escenarios con recursos limitados.

En conjunto, estos resultados evidencian que el método de explicabilidad propuesto cumple satisfactoriamente con los criterios de calidad establecidos: alta fidelidad, consistencia perfecta, uniformidad adecuada, robustez razonable y eficiencia operativa. Estas características lo posicionan como una herramienta confiable y útil para la interpretación de modelos en dominios sensibles o críticos.

Propiedad	Dengue	COVID-19	Diabetes	Fallo en VASs
Fidelidad	0.9392	0.9264	0.9533	0.9162
Estabilidad	0	0	0	0
Uniformidad	0.9911	0.9540	0.9685	0.9397
Robustez	0.7248	0.7231	0.7561	0.7300
Eficiencia	0.1167	0.1072	0.0013	0.0172

Tabla 5.17: Resultados promedio de las propiedades evaluadas del método de explicabilidad

Capítulo 6

Conclusiones y líneas futuras

6.1 Resumen

Este trabajo presenta un nuevo método para mejorar la explicabilidad en los MCDs. A diferencia de los enfoques clásicos usados en MCDs que se basan en medidas de la teoría de grafos para la obtención de la explicabilidad a partir de la imagen final del modelo o de técnicas de explicabilidad adaptadas a los MCDs, este enfoque propone una explicabilidad dinámica y causal. Se centra en el comportamiento dinámico del modelo durante cada iteración del proceso de inferencia, considerando cómo las relaciones directas e indirectas entre conceptos influyen en la evolución de las activaciones.

El método fue evaluado en profundidad mediante una serie de experimentos comparativos. En primer lugar, se contrastaron las explicaciones generadas con aquellas obtenidas a través de medidas clásicas de centralidad de grafos, comúnmente utilizadas en MCDs. Se comprobó que el enfoque propuesto ofrece explicaciones más representativas y útiles, ya que no se limita a analizar la estructura estática del modelo, sino que incorpora la dinámica de la inferencia, capturando el papel que cada concepto desempeña a lo largo del tiempo. Adicionalmente, se comparó el método con técnicas de explicabilidad ampliamente utilizadas en la literatura, como SHAP y FP. Dado que estos métodos no pueden aplicarse directamente sobre MCDs, se entrenaron distintos modelos de IA, como redes neuronales y árboles de decisión, utilizando los mismos conjuntos de datos. Se diseñaron cuidadosamente los experimentos: se adaptaron los datos, se construyeron los modelos, se midió su rendimiento y se generaron las explicaciones con SHAP y FP.

La calidad de las explicaciones se evaluó desde dos perspectivas. En primer lugar, se utilizó la técnica ROAR. El método propuesto mostró una mayor degradación de rendimiento que SHAP y FP, lo cual indica que identifica de forma más precisa las variables clave. En segundo lugar, se comprobó la robustez de las explicaciones generadas, evaluando el conjunto de propiedades fundamentales que debe cumplir un método de explicabilidad para garantizar su calidad, utilidad y confiabilidad, y se verificó que las explicaciones propuestas cumplen con dichas propiedades.

Todas las evaluaciones se realizaron sobre cuatro conjuntos de datos distintos, y en todos los casos el método propuesto mostró resultados consistentes y superiores. Particularmente, los resultados evidenciaron diferencias significativas con respecto a SHAP y FP. Tanto SHAP como FP tendieron a identificar como más importantes aquellas variables con alta multicolinealidad, lo cual compromete la calidad de las explicaciones **blue** ya que en presencia de multicolinealidad el modelo no puede distinguir claramente entre variables altamente correlacionadas. Esto provoca que la importancia se distribuya

arbitrariamente entre ellas, lo que dificulta identificar cuáles factores tienen un verdadero impacto en la predicción.

En contraste, el método propuesto no se ve afectado por este problema, ya que evalúa la relevancia de las variables considerando su impacto efectivo dentro del proceso inferencial del modelo. Por tanto, estas diferencias resaltan la importancia de adoptar enfoques que integren el análisis causal dinámicamente para una interpretación más fiel y completa de los modelos, especialmente en dominios complejos donde la simple correlación puede resultar insuficiente o engañosa.

6.2 Hallazgos

A continuación se detallan los hallazgos mas relevantes derivados del análisis y evaluación del enfoque propuesto. Se resaltan aspectos fundamentales que evidencian las ventajas y aportes significativos del método propuesto frente a enfoques tradicionales:

- **Desarrollo de un método de explicabilidad *post-hoc* para MCDs.**
  Se ha diseñado y validado un método de explicabilidad específico para modelos de clasificación basados en MCDs que proporciona explicaciones precisas, confiables y coherentes con el comportamiento del modelo.

- **Incorporación de la causalidad en la explicabilidad.**
  A diferencia de la mayoría de los métodos existentes, el enfoque integra fundamentos causales para explicar no solo qué influye, sino por qué influye, aspecto que ha sido escasamente explorado en la literatura sobre MCDs.

- **Consideración del comportamiento dinámico del modelo.**
  El método aprovecha la dinámica interna del proceso de inferencia, analizando cómo las influencias directas e indirectas evolucionan a lo largo de las iteraciones, un aspecto poco abordado en trabajos previos sobre explicabilidad en MCDs.

- **Robustez y confiabilidad demostradas.**
  El método de explicabilidad cumple con las propiedades fundamentales establecidas en la literatura que aseguran su calidad, utilidad y confiabilidad en diferentes dominios y configuraciones.

- **Generación de explicaciones visuales**
  El método permite representar gráficamente las rutas causales y el flujo de influencia entre conceptos, lo que facilita la interpretación por parte de expertos humanos. Estas representaciones visuales son especialmente útiles en contextos críticos, como el médico, donde la comprensión clara de las decisiones del modelo es fundamental.

6.3 Limitaciones

A pesar de los resultados positivos obtenidos con el enfoque propuesto, es importante reconocer una serie de limitaciones que condicionan su aplicabilidad y generalización. Estas limitaciones están relacionadas con la naturaleza del método, el contexto de aprendizaje en el que se probo, como con las herramientas disponibles actualmente para el trabajo con MCDs. A continuación, se enumeran los principales aspectos identificados durante el desarrollo y la evaluación del método:

- **Coste computacional:** El método presenta un mayor coste computacional en comparación con los enfoques estáticos, debido al análisis de caminos dinámicos y la integración de influencias temporales durante el proceso de inferencia.
- **Requiere convergencia del modelo:** No todos los modelos basados en MCDs garantizan llegar a un estado estable (converge), lo que afecta la consistencia de las explicaciones generadas.
- **Limitaciones en el manejo de problemas multiclase:** La herramienta empleada para la construcción de los MCDs presenta dificultades para trabajar con problemas de clasificación multiclase con más de tres clases, limitando su aplicación en conjuntos de datos con mayor número de clases. A su vez, no existe otra herramienta abierta que permita el desarrollo de modelos de clasificación multiclases con MCDs .
- **Experimentacion solo con modelos supervisados de clasificación:** El método propuesto fue concebido para proporcionar explicabilidad en modelos de MCDs aplicados tanto a tareas de clasificación como de predicción. No obstante, en este trabajo su desarrollo se ha centrado únicamente en el contexto de clasificación. La razón principal ha sido la inexistencia de herramientas abiertas que permitan construir MCDs orientados a tareas de predicción. Las pocas soluciones disponibles son de uso privado, desarrolladas por laboratorios de investigación específicos.

6.4 Trabajos Futuros

El presente trabajo ha abierto nuevas líneas de investigación en el ámbito de la explicabilidad dinámica aplicada a los (MCD). A continuación, se proponen algunas direcciones prometedoras para su desarrollo futuro:

- **Extenderlo a modelos descriptivos basados en MCD (sin una variable objetivo):** Extender el enfoque propuesto a MCDs que describan la dinamica de un sistema, que no requieren explícitamente un concepto objetivo de salida. En este caso, el objetivo sería evaluar la importancia relativa de cada concepto en el sistema, considerando su influencia global sobre el comportamiento de la red para llegar a un estado estable. Esta adaptación permitiría aplicar el método en contextos donde no se dispone de una variable objetivo bien definida, como sistemas descriptivos o de simulación, un muy común uso de los MCD.
- **Aplicarlo en problemas de regresión (predicción):** Probar el enfoque propuesto a contextos de regresión, donde existe una variable objetivo continua a predecir. En este escenario, el objetivo sería calcular la influencia dinámica de los conceptos sobre dicha variable, lo cual permitiría identificar cuáles son los factores más determinantes en el resultado de la regresión. Eso permitiría evaluar nuestro método en todas las tareas de aprendizaje supervisado.
- **Mejorar su eficiencia computacional:** Investigar estrategias de optimización que reduzcan el coste computacional del cálculo de caminos e influencias, especialmente en modelos de MCDs de gran escala. Esto podría incluir técnicas de poda, heurísticas para seleccionar los caminos más relevantes, o paralelización del proceso de cálculo.
- **Explorar nuevas funciones de penalización:** Evaluar el impacto de distintas funciones de penalización dentro del proceso de cálculo de importancia de los caminos indirectos. Probar alternativas a la función empleada actualmente podría mejorar la sensibilidad del método a distintos patrones de interacción.

Estas líneas futuras permitirán consolidar el enfoque propuesto y ampliar su aplicabilidad a una mayor variedad de problemas, reforzando su utilidad en entornos reales que demandan transparencia y comprensión en los procesos de toma de decisiones automatizados.

Bibliografía

[1] Z. C. Lipton, *The Mythos of Model Interpretability*, 2017. arXiv: 1606.03490 [cs.LG]. dirección: https://arxiv.org/abs/1606.03490.

[2] F. Doshi-Velez y B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning*, 2017. arXiv: 1702.08608 [stat.ML]. dirección: https://arxiv.org/abs/1702.08608.

[3] M. T. Ribeiro, S. Singh y C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, 2016. arXiv: 1602.04938 [cs.LG]. dirección: https://arxiv.org/abs/1602.04938.

[4] S. M. Lundberg y S.-I. Lee, "A unified approach to interpreting model predictions", en *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ép. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, págs. 4768-4777, ISBN: 9781510860964.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh y D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", en *Proceedings of the IEEE international conference on computer vision*, 2017, págs. 618-626.

[6] B. Kosko, "Fuzzy cognitive maps", *International Journal of Man-Machine Studies*, vol. 24, n.º 1, págs. 65-75, 1986, ISSN: 0020-7373. DOI: https://doi.org/10.1016/S0020-7373(86)80040-2. dirección: https://www.sciencedirect.com/science/article/pii/S0020737386800402.

[7] E. I. Papageorgiou y J. L. Salmeron, "A review of fuzzy cognitive maps research during the last decade", *IEEE transactions on fuzzy systems*, vol. 21, n.º 1, págs. 66-79, 2012.

[8] J. Aguilar, "A survey about Fuzzy Cognitive maps papers", *International Journal of Computational Cognition*, vol. 3, ene. de 2005.

[9] W. Stach, L. A. Kurgan y W. Pedrycz, "Numerical and linguistic prediction of time series with the use of fuzzy cognitive maps", *IEEE transactions on fuzzy systems*, vol. 16, n.º 1, págs. 61-72, 2008.

[10] S. Bueno y J. L. Salmeron, "Benchmarking main activation functions in fuzzy cognitive maps", *Expert systems with Applications*, vol. 36, n.º 3, págs. 5221-5229, 2009.

[11] G. Nápoles, N. Ranković e Y. Salgueiro, "On the interpretability of Fuzzy Cognitive Maps", *Knowledge-Based Systems*, vol. 281, pág. 111 078, 2023, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.111078. dirección: https://www.sciencedirect.com/science/article/pii/S0950705123008286.

[12] K. Kokkinos, E. Lakioti, E. Papageorgiou, K. Moustakas y V. Karayannis, "Fuzzy cognitive map-based modeling of social acceptance to overcome uncertainties in establishing waste biorefinery facilities", *Frontiers in Energy Research*, vol. 6, pág. 112, 2018.

[13] B. G. Giles, C. S. Findlay, G. Haas, B. LaFrance, W. Laughing y S. Pembleton, "Integrating conventional science and aboriginal perspectives on diabetes using fuzzy cognitive maps", *Social science & medicine*, vol. 64, n.º 3, págs. 562-576, 2007.

[14] C. Shearer, "The CRISP-DM model: the new blueprint for data mining", *Journal of data warehousing*, vol. 5, n.º 4, págs. 13-22, 2000.

[15] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal y A. Taly, "Explainable AI in Industry", en *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ép. KDD '19, Anchorage, AK, USA: Association for Computing Machinery, 2019, págs. 3203-3204, ISBN: 9781450362016. DOI: 10.1145/3292500.3332281. dirección: https://doi.org/10.1145/3292500.3332281.

[16] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay y A. Maruthamuthu, "A Literature Review on Applications of Explainable Artificial Intelligence (XAI)", *IEEE Access*, vol. 13, págs. 41 111-41 140, 2025. DOI: 10.1109/ACCESS.2025.3546681.

[17] Z. Wang, Y. Liu, A. Arumugam Thiruselvi y A. Hamou-Lhadj, "XAIport: A Service Framework for the Early Adoption of XAI in AI Model Development", en *Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, ép. ICSE-NIER'24, Lisbon, Portugal: Association for Computing Machinery, 2024, págs. 67-71, ISBN: 9798400705007. DOI: 10.1145/3639476.3639759. dirección: https://doi.org/10.1145/3639476.3639759.

[18] R. M. Munshi, L. Cascone, N. Alturki, O. Saidani, A. Alshardan y M. Umer, "A novel approach for breast cancer detection using optimized ensemble learning framework and XAI", *Image and Vision Computing*, vol. 142, pág. 104 910, 2024, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2024.104910. dirección: https://www.sciencedirect.com/science/article/pii/S0262885624000131.

[19] A. Namrita Gummadi, J. C. Napier y M. Abdallah, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems", *IEEE Access*, vol. 12, págs. 71 024-71 054, 2024. DOI: 10.1109/ACCESS.2024.3402446.

[20] D. Tchuente, J. Lonlac y B. Kamsu-Foguem, "A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications", *Computers in Industry*, vol. 155, pág. 104 044, 2024, ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2023.104044. dirección: https://www.sciencedirect.com/science/article/pii/S016636152300194X.

[21] G. Bonifazi, F. Cauteruccio, E. Corradini et al., "A model-agnostic, network theory-based framework for supporting XAI on classifiers", *Expert Systems with Applications*, vol. 241, pág. 122 588, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.122588. dirección: https://www.sciencedirect.com/science/article/pii/S0957417423030907.

[22] S. Nazat, L. Li y M. Abdallah, "XAI-ADS: An Explainable Artificial Intelligence Framework for Enhancing Anomaly Detection in Autonomous Driving Systems", *IEEE Access*, vol. 12, págs. 48 583-48 607, 2024. DOI: 10.1109/ACCESS.2024.3383431.

[23] O. Arreche, T. Guntur y M. Abdallah, "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems", *Applied Sciences*, vol. 14, n.º 10, 2024, ISSN: 2076-3417. DOI: 10.3390/app14104170. dirección: https://www.mdpi.com/2076-3417/14/10/4170.

[24] M. Naiseh, A. Simkute, B. Zieni, N. Jiang y R. Ali, "C-XAI: A conceptual framework for designing XAI tools that support trust calibration", *Journal of Responsible Technology*, vol. 17, pág. 100 076, 2024, ISSN: 2666-6596. DOI: https://doi.org/10.1016/j.jrt.2024.100076. dirección: https://www.sciencedirect.com/science/article/pii/S2666659624000027.

[25] L. Longo, M. Brcic, F. Cabitza et al., "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions", *Information Fusion*, vol. 106, pág. 102 301, 2024, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2024.102301. dirección: https://www.sciencedirect.com/science/article/pii/S1566253524000794.

[26] C. Zhao, J. Liu y E. Parilina, "ShapG: New feature importance method based on the Shapley value", *Engineering Applications of Artificial Intelligence*, vol. 148, pág. 110 409, 2025, ISSN: 0952-1976. DOI: 10.1016/j.engappai.2025.110409.

[27] F. van Mourik, M. A. Haeri, F. A. Bukhsh y F. Ahmed, "IterSHAP: An XAI-Based Feature Selection Method for Small High-Dimensional Datasets", en *Proceedings of the Future Technologies Conference (FTC) 2024, Volume 2*, ép. Lecture Notes in Networks and Systems, Germany: Springer, 2024, págs. 526-545, ISBN: 9783031731211. DOI: 10.1007/978-3-031-73122-8_35.

[28] Z. Guo, Z. Wu, T. Xiao, C. Aggarwal, H. Liu y S. Wang, "Counterfactual Learning on Graphs: A Survey", English (US), *Machine Intelligence Research*, vol. 22, n.º 1, págs. 17-59, feb. de 2025, Publisher Copyright: © The Author(s) 2025., ISSN: 2731-538X. DOI: 10.1007/s11633-024-1519-z.

[29] J. Sun, M. Gao, H. Wang y Q. Dong, "Recursive Counterfactual Deconfounding for image recognition", *Knowledge-Based Systems*, vol. 315, pág. 113 245, 2025, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2025.113245. dirección: https://www.sciencedirect.com/science/article/pii/S0950705125002928.

[30] E. Kaufman y A. Levy, "Explainable AI Approach using Near Misses Analysis", *arXiv preprint arXiv:2411.16895*, 2024, Submitted on 25 Nov 2024. dirección: https://doi.org/10.48550/arXiv.2411.16895.

[31] H. Löfström, T. Löfström, U. Johansson y C. Sönströd, "Calibrated explanations: With uncertainty information and counterfactuals", *Expert Systems with Applications*, vol. 246, pág. 123 154, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2024.123154. dirección: https://www.sciencedirect.com/science/article/pii/S0957417424000198.

[32] A. Wang, Y. Pruksachatkun, N. Nangia et al., "SuperGLUE: a stickier benchmark for general-purpose language understanding systems", en *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[33] R. Achtibat, S. M. V. Hatefi, M. Dreyer et al., "AttnLRP: attention-aware layer-wise relevance propagation for transformers", en *Proceedings of the 41st International Conference on Machine Learning*, ép. ICML'24, Vienna, Austria: JMLR.org, 2024.

[34] Y. Bakish, I. Zimerman, H. Chefer y L. Wolf, *Revisiting LRP: Positional Attribution as the Missing Ingredient for Transformer Explainability*, jun. de 2025. DOI: 10.48550/arXiv.2506.02138.

[35] J. Wei, X. Wang, D. Schuurmans et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", en *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho y A. Oh, eds., vol. 35, Curran Associates, Inc., 2022, págs. 24 824-24 837. dirección: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[36] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang y J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications", *Information Fusion*, vol. 81, págs. 59-83, 2022, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.11.003. dirección: https://www.sciencedirect.com/science/article/pii/S1566253521002281.

[37] S. Beckers, *Causal Explanations and XAI*, 2022. arXiv: 2201.13169 [cs.AI]. dirección: https://arxiv.org/abs/2201.13169.

[38] S. S. Kumar y C. Soares, "Causal Inference Explanations for Graph Neural Networks", en *9th Causal Inference Workshop at UAI 2024*, 2024. dirección: https://openreview.net/forum?id=xB99i5yHtm.

[39] J. Peters, D. Janzing y B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017, ISBN: 9780262037319.

[40] I. Bica, A. M. Alaa y M. van der Schaar, "Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders", *CoRR*, vol. abs/1902.00450, 2019. arXiv: 1902.00450. dirección: http://arxiv.org/abs/1902.00450.

[41] N. O. Breuer, A. Sauter, M. Mohammadi y E. Acar, *CAGE: Causality-Aware Shapley Value for Global Explanations*, 2024. arXiv: 2404.11208 [cs.AI]. dirección: https://arxiv.org/abs/2404.11208.

[42] M. Cinquini y R. Guidotti, "Causality-Aware Local Interpretable Model-Agnostic Explanations", en *Explainable Artificial Intelligence*. Springer Nature Switzerland, 2024, págs. 108-124, ISBN: 9783031638008. DOI: 10.1007/978-3-031-63800-8_6. dirección: http://dx.doi.org/10.1007/978-3-031-63800-8_6.

[43] Y. Cheng, X. Song, Z. Wang, Q. Zhong, K. He y J. Suo, *Causally-informed Deep Learning towards Explainable and Generalizable Outcomes Prediction in Critical Care*, 2025. arXiv: 2502.02109 [cs.LG]. dirección: https://arxiv.org/abs/2502.02109.

[44] O. Lang, I. Traynis e Y. Liu, "Explaining counterfactual images", *Nature Biomedical Engineering*, vol. 9, págs. 287-289, 2025, Published 20 December 2023, Issue Date March 2025. DOI: 10.1038/s41551-023-01164-5. dirección: https://doi.org/10.1038/s41551-023-01164-5.

[45] E. I. Papageorgiou, *Fuzzy Cognitive Maps for Applied Sciences and Engineering: From Fundamentals to Extensions and Learning Algorithms*. Springer, 2019.

[46] L. Zadeh, "Fuzzy sets", *Information and Control*, vol. 8, n.º 3, págs. 338-353, 1965, ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(65)90241-X. dirección: https://www.sciencedirect.com/science/article/pii/S001999586590241X.

[47] R. Axelrod, *Structure of decision: The cognitive maps of political elites*. Princeton University Press, 1976.

[48] B. Kosko, *Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence*. USA: Prentice-Hall, Inc., 1992, ISBN: 0136123341.

[49] W. Hoyos, J. Aguilar y M. Toro, "A clinical decision-support system for dengue based on fuzzy cognitive maps", *Health Care Management Science*, vol. 25, n.º 4, págs. 666-681, 2022, © 2022. The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature, ISSN: 1572-9389. DOI: 10.1007/s10729-022-09611-6. dirección: https://doi.org/10.1007/s10729-022-09611-6.

[50] G. Nápoles, M. Leon Espinosa, I. Grau, K. Vanhoof y R. Bello, "Fuzzy Cognitive Maps Based Models for Pattern Classification: Advances and Challenges", en *Soft Computing Based Optimization and Decision Models: To Commemorate the 65th Birthday of Professor José Luis "Curro" Verdegay*, D. A. Pelta y C. Cruz Corona, eds. Cham: Springer International Publishing, 2018, págs. 83-98, ISBN: 978-3-319-64286-4. DOI: 10.1007/978-3-319-64286-4_5. dirección: https://doi.org/10.1007/978-3-319-64286-4_5.

[51] G. Papakostas, D. Koulouriotis, A. Polydoros y V. Tourassis, "Towards Hebbian learning of Fuzzy Cognitive Maps in pattern classification problems", *Expert Systems with Applications*, vol. 39, n.º 12, págs. 10 620-10 629, 2012, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2012.02.148. dirección: https://www.sciencedirect.com/science/article/pii/S0957417412004113.

[52] J. Aguilar, "Different dynamic causal relationship approaches for cognitive maps", *Applied Soft Computing*, vol. 13, n.º 1, págs. 271-282, 2013, ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2012.08.037. dirección: https://www.sciencedirect.com/science/article/pii/S1568494612003948.

[53] G. Nápoles, M. L. Espinosa, I. Grau y K. Vanhoof, "FCM Expert: Software Tool for Scenario Analysis and Pattern Classification Based on Fuzzy Cognitive Maps", *International Journal on Artificial Intelligence Tools*, vol. 27, n.º 07, pág. 1 860 010, 2018. DOI: 10.1142/S0218213018600102. eprint: https://doi.org/10.1142/S0218213018600102. dirección: https://doi.org/10.1142/S0218213018600102.

[54] E. I. Papageorgiou, "Learning Algorithms for Fuzzy Cognitive Maps—A Review Study", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, n.º 2, págs. 150-163, 2012. DOI: 10.1109/TSMCC.2011.2138694.

[55] Z. Ren, "Learning Fuzzy Cognitive Maps by a Hybrid Method Using Nonlinear Hebbian Learning and Extended Great Deluge Algorithm", *Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference, MAICS 2012*, vol. 841, ene. de 2012.

[56] G. Felix, G. Nápoles, R. Falcon et al., "A review on methods and software for fuzzy cognitive maps", *Artificial Intelligence Review*, vol. 52, n.º 3, págs. 1707-1737, 2019. DOI: 10.1007/s10462-017-9575-1. dirección: https://doi.org/10.1007/s10462-017-9575-1.

[57] W. Hoyos, J. Aguilar y M. Toro, "Federated learning approaches for fuzzy cognitive maps to support clinical decision-making in dengue", *Engineering Applications of Artificial Intelligence*, vol. 123, pág. 106 371, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.106371. dirección: https://www.sciencedirect.com/science/article/pii/S0952197623005559.

[58] W. Hoyos, J. Aguilar y M. Toro, "PRV-FCM: An extension of fuzzy cognitive maps for prescriptive modeling", *Expert Systems with Applications*, vol. 231, pág. 120 729, 2023, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.120729. dirección: https://www.sciencedirect.com/science/article/pii/S0957417423012319.

[59] W. Pedrycz, "Federated FCM: Clustering Under Privacy Requirements", *IEEE Transactions on Fuzzy Systems*, vol. 30, n.º 8, págs. 3384-3388, 2022. DOI: 10.1109/TFUZZ.2021.3105193.

[60] G. D. Karatzinis, N. A. Apostolikas, Y. S. Boutalis y et al., "Fuzzy Cognitive Networks in Diverse Applications Using Hybrid Representative Structures", *International Journal of Fuzzy Systems*, vol. 25, págs. 2534-2554, 2023. DOI: 10.1007/s40815-023-01564-4. dirección: https://doi.org/10.1007/s40815-023-01564-4.

[61] C. Hwang y F. C.-H. Rhee, "Uncertain Fuzzy Clustering: Interval Type-2 Fuzzy Approach to C-Means", *IEEE Transactions on Fuzzy Systems*, vol. 15, n.º 1, págs. 107-120, 2007. DOI: 10.1109/TFUZZ.2006.889763.

[62] M. Kolahdoozi, A. Amirkhani, M. H. Shojaeefard y et al., "A novel quantum inspired algorithm for sparse fuzzy cognitive maps learning", *Applied Intelligence*, vol. 49, págs. 3652-3667, 2019. DOI: 10.1007/s10489-019-01476-7. dirección: https://doi.org/10.1007/s10489-019-01476-7.

[63] L. Parreño y F. Pablo-Martí, "Fuzzy cognitive maps for municipal governance improvement", *PLOS ONE*, vol. 19, n.º 2, e0294962, 2024, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0294962. dirección: https://doi.org/10.1371/journal.pone.0294962.

[64] N. Adabavazeh, M. Nikbakht, A. Amindoust y S. Ali Hassanzadeh-Tabrizi, "The identification and analysis of pivotal factors influencing the corrosion of natural gas pipelines using fuzzy cognitive map", *Engineering Failure Analysis*, vol. 166, pág. 108 806, 2024, ISSN: 1350-6307. DOI: https://doi.org/10.1016/j.engfailanal.2024.108806. dirección: https://www.sciencedirect.com/science/article/pii/S1350630724008525.

[65] Z. Song, Z. Zhang, F. Lyu, M. Bishop, J. Liu y Z. Chi, "From Individual Motivation to Geospatial Epidemiology: A Novel Approach Using Fuzzy Cognitive Maps and Agent-Based Modeling for Large-Scale Disease Spread", *Sustainability*, vol. 16, n.º 12, 2024, ISSN: 2071-1050. DOI: 10.3390/su16125036. dirección: https://www.mdpi.com/2071-1050/16/12/5036.

[66] M. Bevilacqua, F. Ciarapica, G. Marcucci y G. Mazzuto, "Fuzzy Cognitive Maps analysis of Green Supply Chain Management: a case study approach", *IFAC-PapersOnLine*, vol. 53, n.º 2, págs. 17 481-17 486, 2020, 21st IFAC World Congress, ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2020.12.2124. dirección: https://www.sciencedirect.com/science/article/pii/S2405896320327762.

[67] C. Kadaifci, S. Karadayi-Usta y O. Yanmaz, "An analysis of consumer opinions on waste medicine management utilizing fermatean fuzzy cognitive mapping", *Environmental Development*, vol. 49, pág. 100 961, 2024, ISSN: 2211-4645. DOI: https://doi.org/10.1016/j.envdev.2023.100961. dirección: https://www.sciencedirect.com/science/article/pii/S2211464523001616.

[68] A. Ekici, Ş. Ö. Ekici, I. Ö. Yumurtacı Hüseyinoğlu y F. Watson, "A fuzzy cognitive map approach to understand agricultural system and food prices in Türkiye: Policy recommendations for national food security", *Systems Research and Behavioral Science*, vol. 41, n.º 3, págs. 471-495, 2024. DOI: https://doi.org/10.1002/sres.2989. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sres.2989. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1002/sres.2989.

[69] A. Adadi y M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", *IEEE Access*, vol. 6, págs. 52 138-52 160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

[70] B. Mittelstadt, C. Russell y S. Wachter, "Explaining Explanations in AI", en *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ép. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, págs. 279-288, ISBN: 9781450361255. DOI: 10.1145/3287560.3287574. dirección: https://doi.org/10.1145/3287560.3287574.

[71] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, 2019. arXiv: 1910.10045 [cs.AI]. dirección: https://arxiv.org/abs/1910.10045.

[72] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf y G. Z. Yang, "XAI–Explainable artificial intelligence", *Science Robotics*, vol. 4, n.º 37, eaay7120, 2019. DOI: 10.1126/scirobotics.aay7120. dirección: https://doi.org/10.1126/scirobotics.aay7120.

[73] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence*, vol. 267, págs. 1-38, 2019, ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007. dirección: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[74] B. Y. Lim, A. K. Dey y D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems", en *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ép. CHI '09, Boston, MA, USA: Association for Computing Machinery, 2009, págs. 2119-2128, ISBN: 9781605582467. DOI: 10.1145/1518701.1519023. dirección: https://doi.org/10.1145/1518701.1519023.

[75] J. Jung, S. Kang, J. Choi, R. El-Kareh, H. Lee y H. Kim, "Evaluating the impact of explainable AI on clinicians' decision-making: A study on ICU length of stay prediction", *International Journal of Medical Informatics*, vol. 201, pág. 105 943, 2025. DOI: 10.1016/j.ijmedinf.2025.105943.

[76] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman y A. Galstyan, *A Survey on Bias and Fairness in Machine Learning*, 2022. arXiv: 1908.09635 [cs.LG]. dirección: https://arxiv.org/abs/1908.09635.

[77] J. Buolamwini y T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", en *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler y C. Wilson, eds., ép. Proceedings of Machine Learning Research, vol. 81, PMLR, 23–24 Feb de 2018, págs. 77-91. dirección: https://proceedings.mlr.press/v81/buolamwini18a.html.

[78] A. Fuster, P. Goldsmith-pinkham, T. Ramadorai y A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets", *The Journal of Finance*, vol. 77, págs. 5-47, dic. de 2021. DOI: 10.1111/jofi.13090.

[79] Z. Obermeyer, B. Powers, C. Vogeli y S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, vol. 366, n.º 6464, págs. 447-453, 2019, Accessed Jun 24, 2025. DOI: 10.1126/science.aax2342.

[80] C. Blakeney, G. Atkinson, N. Huish, Y. Yan, V. Metris y Z. Zong, *Measure Twice, Cut Once: Quantifying Bias and Fairness in Deep Neural Networks*, 2021. arXiv: 2110.04397 [cs.LG]. dirección: https://arxiv.org/abs/2110.04397.

[81] R. Bommasani, D. A. Hudson, E. Adeli et al., *On the Opportunities and Risks of Foundation Models*, 2022. arXiv: 2108.07258 [cs.LG]. dirección: https://arxiv.org/abs/2108.07258.

[82] S. Wachter, B. Mittelstadt y L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *International Data Privacy Law*, vol. 7, n.º 2, págs. 76-99, jun. de 2017, ISSN: 2044-3994. DOI: 10.1093/idpl/ipx005. eprint: https://academic.oup.com/idpl/article-pdf/7/2/76/17932196/ipx005.pdf. dirección: https://doi.org/10.1093/idpl/ipx005.

[83] 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, sep. de 2017, ISSN: 2371-9621. DOI: 10.1609/aimag.v38i3.2741. dirección: http://dx.doi.org/10.1609/aimag.v38i3.2741.

[84] U.S. Congress, *Algorithmic Accountability Act*, https://www.congress.gov/bill/117th-congress/house-bill/6580, 2022.

[85] European Commission, *Regulation (EU) 2024/XXXX of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, https://eur-lex.europa.eu/, Accessed June 2025, 2024.

[86] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods", en *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ép. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, págs. 2239-2250, ISBN: 9781450393522. DOI: 10.1145/3531146.3534639. dirección: https://doi.org/10.1145/3531146.3534639.

[87] W. Samek y K.-R. Müller, "Towards Explainable Artificial Intelligence", en *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019, págs. 5-22, ISBN: 9783030289546. DOI: 10.1007/978-3-030-28954-6_1. dirección: http://dx.doi.org/10.1007/978-3-030-28954-6_1.

[88] R. N. McCauley y W. Bechtel, "Explanatory Pluralism and Heuristic Identity Theory", *Theory & Psychology*, vol. 11, n.º 6, págs. 736-760, 2001. DOI: 10.1177/0959354301116002. eprint: https://doi.org/10.1177/0959354301116002. dirección: https://doi.org/10.1177/0959354301116002.

[89] G. Vilone y L. Longo, "Classification of Explainable Artificial Intelligence Methods through Their Output Formats", *Machine Learning and Knowledge Extraction*, vol. 3, n.º 3, págs. 615-661, 2021, ISSN: 2504-4990. DOI: 10.3390/make3030032. dirección: https://www.mdpi.com/2504-4990/3/3/32.

[90] K. Sokol y P. Flach, "Explainability fact sheets: a framework for systematic assessment of explainable approaches", en *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ép. FAT* '20, ACM, ene. de 2020, págs. 56-67. DOI: 10.1145/3351095.3372870. dirección: http://dx.doi.org/10.1145/3351095.3372870.

[91] M. Mersha, K. Lam, J. Wood, A. K. Alshami y J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction", *Neurocomputing*, vol. 599, pág. 128 111, 2024.

[92] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, vol. 1, n.º 5, págs. 206-215, 2019. DOI: 10.1038/s42256-019-0048-x. dirección: https://doi.org/10.1038/s42256-019-0048-x.

[93] C. Molnar, *Interpretable Machine Learning*. Lulu.com, 2020. dirección: https://christophm.github.io/interpretable-ml-book/.

[94] A. Ghorbani, A. Abid y J. Zou, *Interpretation of Neural Networks is Fragile*, 2018. arXiv: 1710.10547 [stat.ML]. dirección: https://arxiv.org/abs/1710.10547.

[95] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo et al., "A perspective on explainable artificial intelligence methods: SHAP and LIME", *Advanced Intelligent Systems*, vol. 7, n.º 1, pág. 2 400 304, 2025.

[96] D. C. Castro, I. Walker y B. Glocker, "Causality matters in medical imaging", *Nature Communications*, vol. 11, n.º 1, pág. 3673, 2020. DOI: 10.1038/s414.

[97] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd. Cambridge: Cambridge University Press, 2009.

[98] J. Pearl, M. Glymour y N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ: John Wiley & Sons, 2020, vol. 88, págs. 256-258, Paperback, $46.75, ISBN: 978-1-1191-8684-7. DOI: 10.1111/insr.12369.

[99] R. Moraffah, M. Karami, R. Guo, A. Raglin y H. Liu, "Causal Interpretability for Machine Learning - Problems, Methods and Evaluation", *SIGKDD Explor. Newsl.*, vol. 22, n.º 1, págs. 18-33, mayo de 2020, ISSN: 1931-0145. DOI: 10.1145/3400051.3400058. dirección: https://doi.org/10.1145/3400051.3400058.

[100] B. Schölkopf, "Causality for Machine Learning", en *Probabilistic and Causal Inference*. ACM, feb. de 2022, págs. 765-804, ISBN: 9781450395861. DOI: 10.1145/3501714.3501755. dirección: http://dx.doi.org/10.1145/3501714.3501755.

[101] K. Imai, L. Keele y T. Yamamoto, "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects", *Statistical Science*, vol. 25, nov. de 2010. DOI: 10.1214/10-STS321.

[102] E. I. Papageorgiou, K. Poczęta y C. Laspidou, "Hybrid model for water demand prediction based on fuzzy cognitive maps and artificial neural networks", en *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2016, págs. 1523-1530. DOI: 10.1109/FUZZ-IEEE.2016.7737871.

[103] C. Stylios y P. Groumpos, "Modeling complex systems using fuzzy cognitive maps", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 34, n.º 1, págs. 155-162, 2004. DOI: 10.1109/TSMCA.2003.818878.

[104] M. Obiedat y S. Samarasinghe, "A novel semi-quantitative Fuzzy Cognitive Map model for complex systems for addressing challenging participatory real life problems", *Applied Soft Computing*, vol. 48, págs. 91-110, 2016.

[105] M. F. Hatwágner, E. Yesil, M. F. Dodurka, E. Papageorgiou, L. Urbas y L. T. Kóczy, "Two-Stage Learning Based Fuzzy Cognitive Maps Reduction Approach", *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2938-2952, 2018. DOI: 10.1109/TFUZZ.2018.2793904.

[106] M. F. Hatwágner y L. T. Kóczy, "Novel methods of FCM model reduction", en *Computational Intelligence and Mathematics for Tackling Complex Problems 2*, Springer, 2022, págs. 101-112.

[107] E. I. Papageorgiou, M. F. Hatwágner, A. Buruzs y L. T. Kóczy, "A concept reduction approach for fuzzy cognitive map models in decision making and management", *Neurocomputing*, vol. 232, págs. 16-33, 2017, Advances in Fuzzy Cognitive Maps Theory, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2016.11.060. dirección: https://www.sciencedirect.com/science/article/pii/S0925231216315727.

[108] M. Tyrovolas, N. D. Kallimanis y C. Stylios, *Advancing Explainable AI with Causal Analysis in Large-Scale Fuzzy Cognitive Maps*, 2024. arXiv: 2405.09190 [cs.AI]. dirección: https://arxiv.org/abs/2405.09190.

[109] G. Nápoles, Y. Salgueiro, I. Grau y M. L. Espinosa, "Recurrence-aware long-term cognitive network for explainable pattern classification", *IEEE transactions on cybernetics*, vol. 53, n.º 10, págs. 6083-6094, 2022.

[110] M. Tyrovolas, X. S. Liang y C. Stylios, "Information flow-based fuzzy cognitive maps with enhanced interpretability", *Granular Computing*, vol. 8, n.º 6, págs. 2021-2038, 2023.

[111] J. Y. Yen, "Finding the k shortest loopless paths in a network", *management Science*, vol. 17, n.º 11, págs. 712-716, 1971.

[112] Secretaría de Salud de Medellín, *Dengue and dengue grave dataset*, Accessed: 2025-06-29, 2020. dirección: http://medata.gov.co/dataset/dengue.

[113] I. C. C. Group, E. Garcia-Gallo, L. Merson et al., "ISARIC-COVID-19 dataset: A Prospective, Standardized, Global Dataset of Patients Hospitalized with COVID-19", *Scientific Data*, vol. 9, pág. 454, 2022. DOI: 10.1038/s41597-022-01534-9. dirección: https://doi.org/10.1038/s41597-022-01534-9.

[114] D. Dua y E. K. Taniskidou, *UCI Machine Learning Repository*, 2017. dirección: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.

[115] M. Salem Alzboon, M. Al-Batah, M. Alqaraleh, A. Abuashour y A. Fuad Bader, "A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes", *arXiv e-prints*, arXiv-2506, 2025.

[116] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum y S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data", *Heliyon*, vol. 10, n.º 2, 2024.

[117] V. Chang, J. Bailey, Q. A. Xu y Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms", *Neural Computing and Applications*, vol. 35, n.º 22, págs. 16 157-16 173, 2023.

[118] A. Ahmed, J. Khan, M. Arsalan et al., "Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset", en *Healthcare*, MDPI, vol. 13, 2024, pág. 37.

[119] D. Ji, X. Yao, S. Li, Y. Tang e Y. Tian, "Autonomous underwater vehicle fault diagnosis dataset", *Data in brief*, vol. 39, pág. 107 477, 2021.

[120] J. Neter, M. H. Kutner, C. J. Nachtsheim, W. Wasserman et al., "Applied linear statistical models", 1996.

[121] G. James, D. Witten, T. Hastie, R. Tibshirani et al., *An introduction to statistical learning*. Springer, 2013, vol. 112.

[122] F. Murtagh y A. Heck, *Multivariate data analysis*. Springer Science & Business Media, 2012, vol. 131.

[123] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, págs. 321-357, 2002.

[124] H. Klein, K. Asseo, N. Karni et al., "Onset, duration and unresolved symptoms, including smell and taste changes, in mild COVID-19 infection: a cohort study in Israeli patients", *Clinical Microbiology and Infection*, vol. 27, n.º 5, págs. 769-774, 2021, ISSN: 1198-743X. DOI: https://doi.org/10.1016/j.cmi.2021.02.008. dirección: https://www.sciencedirect.com/science/article/pii/S1198743X21000835.

[125] CIDRAP, University of Minnesota, *Israeli study finds 2.6 % COVID breakthrough infection rate*, https://www.cidrap.umn.edu/israeli-study-finds-26-covid-breakthrough-infection-rate, Accessed: 2025-07-03, 2023.

[126] S. Kulkarni, A. Chaudhari y R. Vaidya, *Detection of Covid19 Cases using ML*, https://www.kaggle.com/code/mykeysid10/detection-of-covid19-cases-using-ml, Kaggle code repository, accessed 2025-07-03, 2023.

[127] D. Freedman y P. Diaconis, "On the histogram as a density estimator: L 2 theory", *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, n.º 4, págs. 453-476, 1981.

[128] M. C. Data, M. Komorowski, D. C. Marshall, J. D. Salciccioli e Y. Crutain, "Exploratory data analysis", *Secondary analysis of electronic health records*, págs. 185-203, 2016.

[129] R. Gnanadesikan y M. B. Wilk, "Probability plotting methods for the analysis of data", *Biometrika*, vol. 55, n.º 1, págs. 1-17, 1968.

[130] S. Hooker, D. Erhan, P.-J. Kindermans y B. Kim, "A Benchmark for Interpretability Methods in Deep Neural Networks", en *Advances in Neural Information Processing Systems*, vol. 32, 2019. dirección: https://proceedings.neurips.cc/paper_files/paper/2019/file/8e2c8063a119d3dffeb2fdf39c9d77f7-Paper.pdf.

[131] R. Poli, J. Kennedy y T. Blackwell, "Particle Swarm Optimization: An Overview", *Swarm Intelligence*, vol. 1, n.º 1, págs. 33-57, oct. de 2007. DOI: 10.1007/s11721-007-0002-0.

[132] 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)*, IEEE, 1998, págs. 69-73. DOI: 10.1109/ICEC.1998.699146.

[133] Y. V. 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: SocProS 2018, Volume 2*, Springer, 2019, págs. 197-206.

[134] 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.

[135] T. Akiba, S. Sano, T. Yanase, T. Ohta y M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework", en *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2019, págs. 2623-2631. DOI: 10.1145/3292500.3330701.

[136] P. Geurts, D. Ernst y L. Wehenkel, "Extremely randomized trees", *Machine learning*, vol. 63, págs. 3-42, 2006.

[137] C. Cortes y V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, págs. 273-297, 1995.

[138] D. E. Rumelhart, G. E. Hinton y R. J. Williams, "Learning representations by back-propagating errors", *nature*, vol. 323, n.º 6088, págs. 533-536, 1986.

[139] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, n.º 2, págs. 215-232, 1958.

[140] T. Cover y P. Hart, "Nearest neighbor pattern classification", *IEEE transactions on information theory*, vol. 13, n.º 1, págs. 21-27, 1967.

[141] A. Altmann, L. Toloşi, O. Sander y T. Lengauer, "Permutation importance: a corrected feature importance measure", *Bioinformatics*, vol. 26, n.º 10, págs. 1340-1347, 2010.

Universidad de Alcalá
Escuela Politécnica Superior

ESCUELA POLITECNICA
SUPERIOR

Universidad
de Alcalá