Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Grado en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Grado

Automatización de Métodos de Validación de Explicaciones para Modelos de Aprendizaje Automático: un Estudio Observacional de la Relación entre Robustez y Precisión

Autor: Stefania Georgia Rac Raican

Tutor(a): Dr. Esteban García-Cuesta

Madrid, julio 2025

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ciencia de Datos e Inteligencia Artificial
Título: Automatización de Métodos de Validación de Explicaciones para Modelos de Aprendizaje Automático: un Estudio Observacional de la Relación entre Robustez y Precisión
Enero 2025

Autor: Stefania Georgia Rac Raican
Tutor:
Dr. Esteban García-Cuesta
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

La Inteligencia Artificial eXplicable (IAX) es un área emergente que busca mejorar la transparencia de los modelos de aprendizaje automático, abordando desafíos como la explicabilidad, la robustez y la gestión de sesgos. Este campo es esencial para proporcionar explicaciones comprensibles sobre cómo los modelos toman decisiones, lo cual es crucial tanto en aplicaciones críticas como en el cumplimiento de normativas de transparencia.

Este trabajo se basa en una metodología propuesta para validación de explicaciones, pero enfocado en evaluar específicamente la relación entre la precisión de un modelo y la robustez de las explicaciones generadas mediante técnicas de IAX post-hoc, como SHAP y LIME, que son agnósticas al modelo. A través de un enfoque experimental y observacional, se estudia cómo los cambios en la precisión del modelo afectan a la estabilidad y fidelidad de las explicaciones proporcionadas. Se emplean tres modelos de aprendizaje automático aplicados a problemas de clasificación binaria, y se introduce ruido controlado en los datos para simular diferentes niveles de precisión, evaluando cómo estas alteraciones impactan las explicaciones generadas. La metodología permite la validación sistemática de explicaciones de modelos complejos.

Abstract

eXplainable Artificial Intelligence (XAI) is an emerging field that aims to improve the transparency of machine learning models by addressing challenges such as explainability, robustness and bias management. This field is essential for providing understandable explanations of how models make decisions, which is crucial in both critical applications and compliance with transparency regulations.

This work is based on a proposed methodology for explanation validation but focuses on specifically evaluating the relationship between the accuracy of a model and the robustness of explanations generated using post-hoc IAX techniques, such as SHAP and LIME, which are model-agnostic. Through an experimental and observational approach, we study how changes in model accuracy affect the stability and fidelity of the explanations provided. Three machine learning models applied to binary classification problems are used, and controlled noise is introduced into the data to simulate different levels of accuracy, evaluating how these alterations impact the explanations generated. The methodology allows the systematic validation of complex model explanations.

Tabla de contenidos

Índice de Figuras

1 Introducción

Desde la antigüedad, los humanos han soñado con crear máquinas capaces de pensar [1]. La Inteligencia Artificial (en adelante, IA) se define como la capacidad de las máquinas de imitar funciones cognitivas humanas como la percepción, la resolución de problemas, la interacción en lenguaje natural e incluso la creatividad [2].

En el siglo XX, con la creación de las redes neuronales y el desarrollo de la computación, esta idea fue cobrando más forma. A mediados de los años 50, comenzó a utilizarse el término "inteligencia artificial", y a lo largo de ese siglo surgieron avances importantes como el perceptrón, los sistemas expertos y los primeros programas de aprendizaje automático y procesamiento del lenguaje natural. Después de atravesar los llamados "inviernos de IA" en las décadas de los 70 y 80, épocas en las que el interés y la inversión en investigación decayó, la llegada del aprendizaje profundo -*deep learning*-, así como mejoras computacionales y grandes cantidades de datos a partir de los años 2000, dieron lugar a una evidente revolución tecnológica. Hoy en día, la inteligencia artificial se integra cotidianamente en nuestra vida mediante asistentes virtuales, chatbots, vehículos autónomos y diagnósticos médicos avanzados [1].

En los últimos años, la IA ha ido tomando un papel aún más relevante, esto es, en gran parte, debido a los modelos de lenguaje de gran tamaño (LLM, por sus siglas en inglés). Estos modelos de aprendizaje profundo, como la serie GPT de OpenAI, BERT de Google o Llama de Meta, pueden interpretar y generar lenguaje como lo haría una persona, dando respuestas precisas y asistiendo en labores de traducción de diversos idiomas, síntesis de textos, redacción o desarrollo de código [3]. Son modelos fundacionales, es decir, redes neuronales que han sido entrenadas sobre enormes conjuntos de datos no etiquetados, habitualmente mediante aprendizaje no supervisado [4].

A pesar de las innegables ventajas que trae consigo la IA, también pueden producirse errores o sesgos. Por ejemplo, en varios sistemas de salud de Estados Unidos, un sistema de IA dio prioridad a pacientes blancos más sanos sobre pacientes negros más enfermos a la hora de recibir atención adicional. Esto ocurrió porque el modelo, al basarse en historiales clínicos previos, asumió un menor riesgo para los pacientes negros, no por estar más sanos, sino porque históricamente habían tenido menos acceso al sistema sanitario, de manera que se generó una base de datos incompleta y sesgada [5].

Además de los sesgos, los modelos de lenguaje actuales pueden generar respuestas incorrectas o inventadas, denominadas alucinaciones, que parecen lógicas y contextualmente adecuadas, aunque no lo sean. Estos errores impredecibles y difíciles de explicar ponen de manifiesto una limitación fundamental en el uso de la IA: su opacidad, comúnmente referida como el

problema de la “caja negra”. Como advierte el Supervisor Europeo de Protección de Datos (EDPS), hasta los propios desarrolladores pueden desconocer el razonamiento que hay detrás de ciertas decisiones, debido a la complejidad de los parámetros y a la falta de transparencia inherente de muchos modelos de aprendizaje profundo [6].

Ante esta problemática, surge el campo de la Inteligencia Artificial Explicable, cuyo objetivo es hacer comprensible el comportamiento de los sistemas de IA, dando explicaciones claras sobre el funcionamiento de los procesos internos que conducen a sus acciones y decisiones [6].

Ahora bien, esta necesidad de explicabilidad no es solo técnica o ética, sino también legal. El Reglamento de Inteligencia Artificial de la Unión Europea (AI Act), impone requisitos específicos de transparencia, trazabilidad y auditabilidad para los sistemas de IA, especialmente en ámbitos considerados de alto riesgo, como la sanidad, la justicia o el empleo. Según este marco regulatorio, los sistemas deben ser diseñados de forma que sus decisiones puedan ser entendidas y explicadas a las personas afectadas, lo que convierte a la IAX en un elemento clave para el cumplimiento normativo y la protección de los derechos fundamentales [7].

1.1 Motivación y definición del proyecto

Este trabajo se enmarca en la línea de investigación iniciada en el artículo “*On the transferability of local model-agnostic explanations of machine learning models to unseen data*”, desarrollado por Alba María López González y Esteban García-Cuesta [8], siendo este último el tutor del presente proyecto. Dicho artículo investiga hasta qué punto las explicaciones generadas por técnicas locales y agnósticas al modelo, como SHAP y LIME, son transferibles a datos no vistos.

La motivación principal de este trabajo surge del interés por profundizar y ampliar los hallazgos del estudio, aplicando su enfoque metodológico a nuevos conjuntos de datos. En concreto, se pretende analizar si existe una relación entre la precisión de un modelo de aprendizaje automático y la robustez de las explicaciones que genera, analizándolo en distintos contextos afectados por ruido y perturbaciones en los datos. Además, se plantea automatizar la metodología empleada, con intención de facilitar su aplicación a diferentes datasets.

1.2 Objetivos

La finalidad de este trabajo es por tanto ayudar a identificar la hipótesis de la existencia de una relación entre la precisión de un modelo de aprendizaje automático y la robustez de las explicaciones que genera. En este contexto, el concepto de *model multiplicity* nos dice que, para un mismo nivel de precisión, pueden existir varios modelos diferentes que logren resultados similares, implicando también diferentes explicaciones de dicho modelo. Esto sucede por una definición insuficiente del problema o el uso de un modelo demasiado complejo para el problema en el que se aplica, lo que permite la existencia de múltiples modelos equivalentes en rendimiento.

La hipótesis que guía este trabajo sostiene que las explicaciones serán más robustas cuanto más cerca esté el modelo de capturar de manera fiel la estructura subyacente del problema, es decir, la realidad que busca representar. A medida que un modelo alcanza una mayor precisión, se espera que el problema de la multiplicidad de modelos disminuya, ya que debería acercarse a una única representación más fidedigna del fenómeno latente. Por el contrario, a medida que la precisión se aleja del 100%, es más probable que existan modelos alternativos con la misma o similar precisión, y esto implicaría reducir la robustez de las explicaciones.

Por tanto, este trabajo se propone medir la degradación en la robustez de las explicaciones conforme disminuye la precisión del modelo. Se evaluará si esta degradación ocurre en mayor medida en ciertos rangos de precisión, por ejemplo, si el impacto es mayor entre el 90% y el 80%, en comparación con una disminución del 80% al 70%, del 70% al 60%, etc. El objetivo final es identificar si existe un punto de inflexión a partir del cual la degradación en la robustez de las explicaciones es tan significativa que deja de tener sentido emplearlas para interpretar el modelo.

Para alcanzar este objetivo, se plantean los siguientes subobjetivos:

1. Analizar la relación entre precisión y robustez de las explicaciones generadas para diferentes modelos de clasificación (XGBoost, SVM y MLP):
    - Comparar la robustez de explicaciones generadas por los modelos.
    - Evaluar si, para un mismo nivel de precisión, algunos modelos ofrecen explicaciones más robustas que otros.
2. Evaluar el impacto del *model multiplicity problem* en la fidelidad de las explicaciones:
    - Determinar si existe una tendencia en la que los modelos con alta precisión generen explicaciones más consistentes y robustas que aquellos con menor precisión.

3. Identificar puntos de inflexión en la precisión a partir de los cuales las explicaciones pierden robustez significativamente:
    - Analizar cómo varía la robustes en distintos intervalos de precisión.
    - Establecer si existe un umbral de precisión por debajo del cual la fidelidad de las explicaciones ya no es confiable.

2 Conceptos previos y Estado del Arte

2.1 Conceptos previos en IAX

2.1.1 Inteligencia Artificial Explicable (IAX)

A medida que los sistemas de IA han ido ganando protagonismo en sectores críticos como la sanidad, las finanzas, la defensa o en la aplicación de la ley, también ha aumentado la preocupación por su opacidad y difícil interpretación. Muchos de los modelos más potentes, especialmente los basados en aprendizaje profundo, funcionan como "cajas negras", proporcionando predicciones muy precisas sin revelar claramente el razonamiento subyacente [9].

En este contexto surge la Inteligencia Artificial eXplicable (en adelante, IAX), una disciplina emergente centrada en hacer comprensible el funcionamiento interno de estos sistemas mediante explicaciones a posteriori sobre cómo llegan a sus decisiones o predicciones [10]. El interés por la IAX ha crecido significativamente en los últimos años, como evidencian las búsquedas del término "Explainable AI" en Google Trends, reflejando una mayor conciencia y relevancia entre investigadores, profesionales y reguladores.

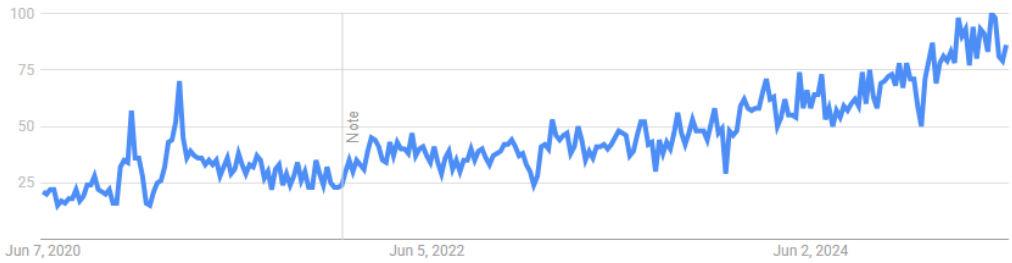


Fig. 1 Crecimiento del interés por la Inteligencia Artificial eXplicable (IAX)

Desde una perspectiva formal, la IAX busca esclarecer cómo se relacionan matemáticamente las entradas y salidas de los modelos, ofreciendo razones claras y humanas sobre sus decisiones, fortaleciendo así la confianza en sus resultados [10]. Asimismo, considera diversas características complementarias tales como la transparencia, robustez, equidad, ética y confianza [11], asegurando que los sistemas sean auditables, justos y que cumplan con estándares éticos y legales.

En definitiva, la IAX persigue no solo mejorar la comprensión técnica de los modelos, sino también garantizar su implementación responsable y confiable, particularmente en contextos sensibles con alto impacto en las personas.

2.1.2 Conceptos generales de IAX

Antes de profundizar en los aspectos específicos de la IAX, resulta esencial presentar algunas definiciones clave para facilitar una mejor comprensión de la terminología utilizada en este campo:

- **Interpretabilidad**: mide el grado en que un modelo puede ser comprendido por una persona, es decir, *cómo* toma sus decisiones a partir de las entradas. Un modelo interpretable permite estimar sus predicciones y reconocer posibles errores de manera intuitiva.
- **Explicabilidad**: es la capacidad de proporcionar descripciones claras y coherentes sobre las decisiones que toma un modelo, explicando *por qué* se ha tomado una determinada decisión a partir de las entradas proporcionadas.
- **Transparencia**: es la capacidad que tiene un modelo para ser comprendido por una persona en su totalidad. Un modelo es considerado transparente cuando sus componentes, incluyendo entradas, parámetros y procesos, pueden ser inspeccionados y explicados intuitivamente por los usuarios. Esta transparencia permite auditar el sistema completo, identificar sesgos y garantizar el cumplimiento de normativas éticas y legales.
- **Confianza**: es la capacidad de un modelo para ofrecer resultados consistentes y estables, conservando un comportamiento predecible y seguro incluso en condiciones variables o nuevas.
- **Robustez**: es la capacidad de un modelo para mantener su desempeño ante pequeñas variaciones, manipulaciones intencionadas o perturbaciones en los datos de entrada que podrían provocar resultados incorrectos o imprevistos.
- **Sesgo**: se refiere a errores sistemáticos en las predicciones del modelo, generalmente causados por datos de entrenamiento incompletos o no representativos. Estos sesgos pueden resultar en decisiones injustas o discriminatorias hacia ciertos grupos poblacionales.
- **Ética**: comprende el conjunto de normas y valores que guían el desarrollo y uso de los modelos, asegurando que estos sean transparentes, responsables y respetuosos con los derechos y valores de la sociedad.
- **Equidad**: implica el tratamiento justo y equitativo por parte del modelo hacia todos los grupos representados en los datos, evitando discriminaciones o sesgos basados en características como raza, género o edad.

Estos conceptos proporcionan una base terminológica solida que permite abordar de manera clara los retos técnicos, éticos y legales vinculados a la IAX, facilitando su aplicación efectiva y responsable [6, 10, 11].

2.1.3 Tipos de modelos según su interpretabilidad

Es fundamental distinguir entre los modelos que son interpretables por diseño y aquellos que requieren técnicas post-hoc para generar explicaciones. Esta distinción no solo tiene implicaciones técnicas, sino también prácticas normativas, puesto que condiciona el tipo de explicaciones que pueden ofrecerse y el grado de confianza que generan en ámbitos donde la transparencia es esencial. La transparencia puede estar incorporada de forma inherente al algoritmo por diseño, o bien obtenerse aplicando técnicas de explicabilidad a posteriori [13].

La relación entre interpretabilidad y precisión se ilustra en la siguiente figura, que muestra cómo los modelos más transparentes tienden a ser menos precisos, mientras que los más precisos suelen ser menos interpretables [9]. Este compromiso es conocido como *trade-off entre interpretabilidad y precisión*.

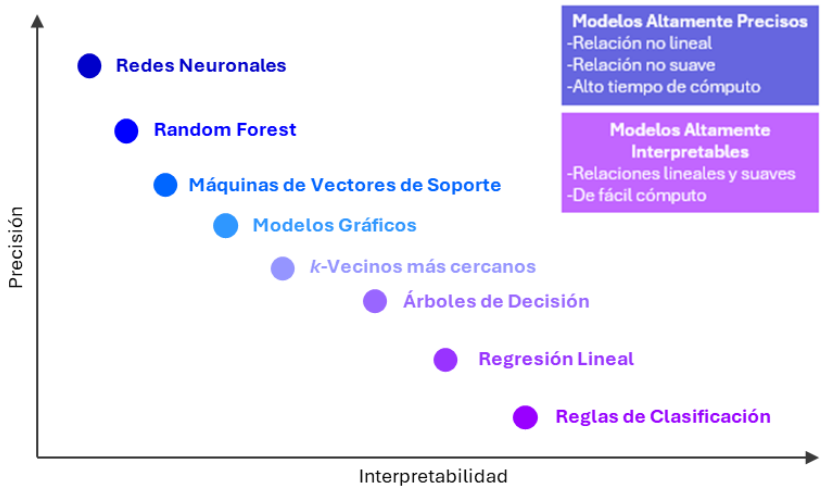


Fig. 2 Relación entre interpretabilidad y precisión de algunos modelos relevantes de aprendizaje automático [12].

Los modelos interpretables, también llamados ***white-box*** o de caja blanca, son aquellos cuyo funcionamiento interno puede ser entendido de forma directa por humanos. Se basan en estructuras simples, que permiten rastrear cómo las variables de entrada contribuyen al resultado final. Ejemplos típicos son la regresión lineal, la regresión logística o los árboles de decisión. Este tipo de modelos ofrece lo que se denomina interpretabilidad intrínseca o ante-hoc, es decir, son explicables desde su construcción sin necesidad de herramientas adicionales [10, 13].

En contraste, los modelos ***black-box*** o de caja negra presentan alta dimensionalidad, transformaciones no lineales y un gran número de parámetros, lo que los convierten en modelos complejos y opacos [11]. Suelen alcanzar un rendimiento predictivo superior, pero sacrifican interpretabilidad. Aunque se

conocen las entradas y salidas del modelo, su funcionamiento interno es difícil o imposible de comprender directamente. Este es el caso de las redes neuronales profundas (*Deep Neural Networks*, DNNs), los métodos ensambladores (*ensemble methods*) como Random Forest, los modelos de aprendizaje profundo (*Deep Learning*) o los *Transformers* [10]. En estos casos, la explicabilidad debe lograrse mediante técnicas post-hoc, es decir, aplicadas una vez entrenado el modelo.

Entre ambos extremos se encuentran los modelos ***gray-box*** o de caja gris, que permiten cierto grado de comprensión de su funcionamiento si están cuidadosamente diseñados. Ejemplos de este tipo son los sistemas difusos (*Fuzzy Systems*), las redes bayesianas (*Bayesian Networks*) o los modelos probabilísticos de difusión (*Diffusion Probabilistic Models*) [10].

La siguiente figura representa gráficamente la comparación entre modelos *white-box*, *gray-box* y *black-box*, según su nivel de interpretabilidad, precisión y adecuación a distintos contextos.

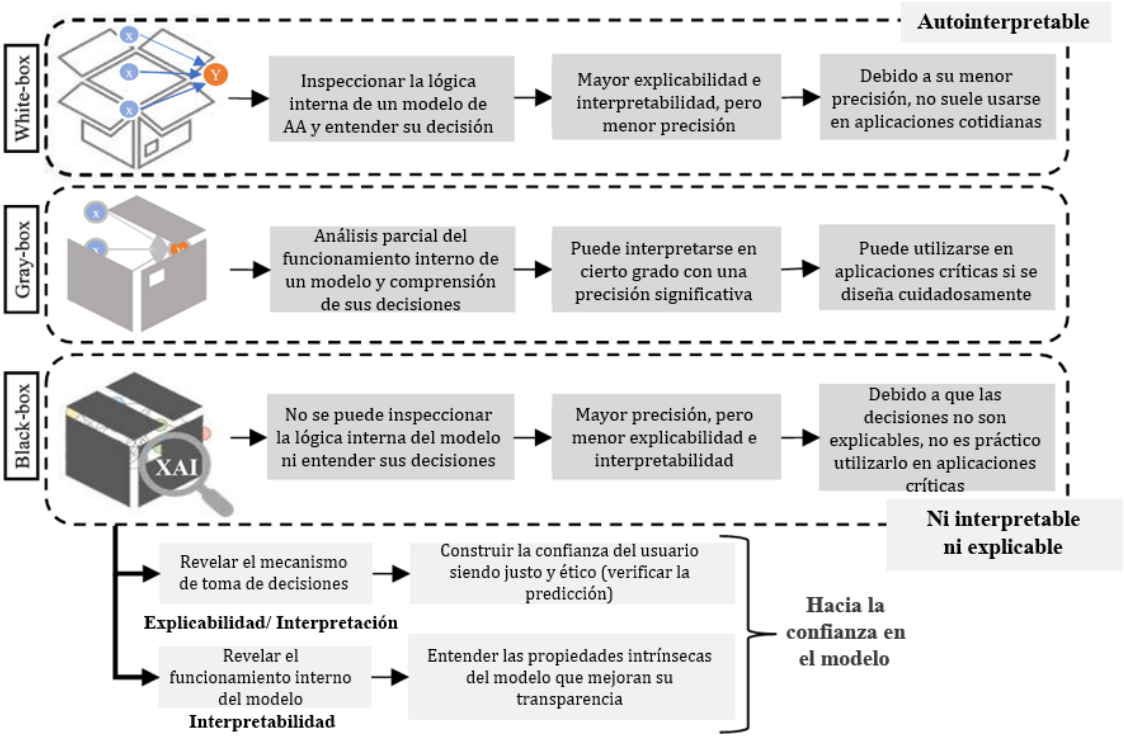


Fig. 3 Comparación entre modelos *white-box, gray-box* y *black-box* según su interpretabilidad, precisión y aplicabilidad [10].

En definitiva, la elección entre un tipo u otro de modelo dependerá de los objetivos específicos del usuario, considerando tanto sus ventajas y limitaciones como el *trade-off* existente entre interpretabilidad y precisión.

2.2 Estado del Arte en IAX

2.2.1 Clasificación de técnicas de explicabilidad

2.2.1.1 Alcance: explicaciones locales vs. globales

Una de las principales formas de clasificar las técnicas en IAX es según el alcance de sus explicaciones. En este sentido, se diferencian dos enfoques: las explicaciones locales y las explicaciones globales. Según [14], esta distinción es habitual literatura, ya que permite seleccionar el tipo de explicación más adecuado en función de los objetivos y del contexto de uso.

Las explicaciones locales se centran en justificar una predicción concreta realizada por el modelo para una instancia específica. Este enfoque es especialmente relevante en dominios donde es crucial entender las decisiones a nivel individual, como el ámbito médico, financiero o legal. Por ejemplo, ante la predicción de que un paciente padece una enfermedad una explicación local permite entender qué características clínicas concretas han llevado al modelo a emitir ese diagnóstico.

Mientras que las explicaciones globales ofrecen una visión general del comportamiento del modelo, proporcionando información sobre los patrones generales de su funcionamiento y destacando qué variables son más influyentes a nivel agregado. Este tipo de explicación resulta particularmente útil en tareas como auditorías, validación de modelos o cumplimiento normativo, donde el objetivo es validar y justificar el comportamiento general del modelo frente a *stakeholders* o autoridades [15, 16].

2.2.1.2 Aplicabilidad: model-agnostic vs. model-specific

Otra dimensión clave en la clasificación de técnicas de explicabilidad es su aplicabilidad, que permite diferenciarlas en función de su dependencia o no del tipo de modelo utilizado.

Las técnicas *model-agnostic* son independientes del modelo, que significa que pueden aplicarse a cualquier algoritmo de aprendizaje automático. Estas técnicas suelen operar a partir de las entradas y salidas del modelo, sin necesidad de conocer su estructura interna. Ejemplos representativos de este enfoque son LIME [17], SHAP [18], MCR [19] o LOCO [20].

Por otro lado, las técnicas *model-specific* están diseñadas para funcionar con modelos concretos, aprovechando su estructura interna para generar explicaciones más eficientes o precisas. Algunos ejemplos son Grad-CAM [21], DeepLift [22] o SmoothGrad [23].

2.2.1.3 Técnicas de explicabilidad post-hoc

Una vez presentada la clasificación general de las técnicas de explicabilidad en función de su alcance y su aplicabilidad, en los apartados siguientes se describen con mayor detalle algunas de las técnicas post-hoc más representativas en el ámbito de la IAX. En concreto, se analizan LIME, Shapley Values y SHAP, ampliamente utilizadas en la literatura.

2.2.1.3.1 LIME

LIME (*Local Interpretable Model-agnostic Explanations*) es una técnica de explicabilidad post-hoc, local y agnóstica al modelo (*model-agnostic*) que permite generar explicaciones comprensibles para modelos complejos considerados como cajas negras. Fue propuesta por Ribeiro, Singh y Guestrin en 2016 con el objetivo de aumentar la confianza de los usuarios en los modelos de aprendizaje automático, facilitando la interpretación de sus predicciones individuales [17]. Su funcionamiento se basa en construir un modelo interpretable, llamado modelo subrogado, que aproxima el comportamiento del modelo complejo, pero únicamente en la vecindad de una instancia específica. Para ello, genera perturbaciones de la instancia a explicar, obtiene las predicciones del modelo sobre estas muestras sintéticas y ajusta un modelo simple (como una regresión lineal o un árbol poco profundo) ponderando las muestras según su proximidad a la instancia original mediante un *kernel* de distancia. Este procedimiento permite aproximar la frontera de decisión del modelo complejo de forma local [24]. Este enfoque se ilustra en la Fig. 4, donde la frontera de decisión compleja del modelo (zona azul y rosa) se aproxima localmente mediante un modelo lineal (línea discontinua) alrededor de la instancia de interés (cruz roja).

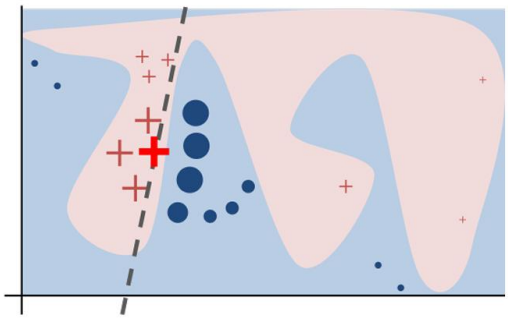


Fig. 4 Ejemplo ilustrativo de la intuición de LIME [17].

Formalmente, LIME resuelve la siguiente optimización:

$$\xi(x) = \arg\min_{g \in G} L\,(f, g, \pi_x) + \Omega(g)$$

donde f es el modelo complejo, g es el modelo interpretable seleccionado del conjunto G, $L(f, g, \pi_x)$ mide la pérdida de fidelidad local (cuánto se asemeja g a

f en la vecindad de la instancia x), π_x es la función de proximidad y $\Omega(g)$ penaliza la complejidad del modelo interpretable.

Este enfoque permite identificar qué características han contribuido más, positiva o negativamente, a la predicción de la instancia seleccionada.

2.2.1.3.2 Shapley Values

Los Shapley Values son un método proveniente de la teoría de juegos cooperativos que permite atribuir de forma justa la contribución de cada característica a la predicción de un modelo de aprendizaje automático [25]. La idea es considerar que cada característica actúa como un «jugador» en un juego, donde la predicción del modelo representa el «beneficio» que debe ser distribuido entre las características según su contribución individual.

Formalmente, el valor de Shapley de una característica i se define como la media de sus contribuciones marginales a todas las posibles coaliciones de características:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

donde N es el conjunto total de características, S es un subconjunto de características que no contiene la característica i y $v(S)$ es la función de valor, que representa la predicción del modelo considerando solo las características presentes en el subconjunto S, mientras que las restantes se reemplazan por valores de referencia (por ejemplo, la media del dataset).

De forma intuitiva, el valor de Shapley para una característica mide cuánto cambia la predicción al añadir dicha característica a diferentes coaliciones posibles de otras características. La suma de todos los valores Shapley es igual a la diferencia entre la predicción del modelo y la media del conjunto de datos, cumpliendo con la propiedad de eficiencia.

2.2.1.3.3 SHAP

SHAP (*SHapley Additive exPlanations*) es un método propuesto por Lundberg y Lee (2017) que permite explicar predicciones individuales de modelos de aprendizaje automático mediante una atribución justa de la importancia de las características, basándose en la teoría de juegos y en los valores de Shapley [18]. Aunque conceptualmente es una extensión directa de los valores de Shapley, SHAP introduce una formulación específica que unifica diversas metodologías previas de explicabilidad post-hoc, como LIME, y proporciona un marco teórico sólido con propiedades deseables que garantizan explicaciones consistentes [26].

La principal innovación de SHAP es reformular las explicaciones como un modelo aditivo lineal de atribuciones sobre un espacio binario de coaliciones, donde cada característica puede estar presente (1) o ausente (0) en la explicación:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

donde z' es el vector binario que indica la presencia o ausencia de cada característica en la coalición, ϕ_i es el valor de SHAP de la característica i, su contribución a la predicción, ϕ_0 es el valor base o predicción esperada cuando no se conoce ninguna característica (*baseline)* y M es el número total de características.

Este modelo aditivo permite que la suma de todas las contribuciones de las características más el valor base reproduzca exactamente la predicción del modelo para la instancia de interés, lo que corresponde a la propiedad de exactitud local.

El cálculo exacto de los valores SHAP es computacionalmente costoso, ya que requiere evaluar todas las posibles coaliciones de características. Para resolver este problema, se han propuesto distintos métodos de estimación: KernelSHAP (método agnóstico al modelo), TreeSHAP (variante optimizada para modelos basadas en árboles) y otros como DeepSHAP (para redes neuronales).

Más allá de las explicaciones locales, SHAP permite realizar interpretaciones globales del modelo mediante la agregación de los valores SHAP de múltiples instancias, lo que permite analizar la importancia global de las características, sus dependencias e interacciones.

2.2.2 Validación de IAX

Uno de los principales desafíos en la IAX es la validación de las explicaciones generadas. A diferencia de las métricas tradicionales de evaluación de modelos, como la precisión o el AUC, no existe un consenso claro sobre cómo medir la calidad de las explicaciones. De hecho, diferentes técnicas de IAX, en concreto, de importancia de características, pueden ofrecer explicaciones divergentes sobre el mismo modelo y conjunto de datos, lo que puede conducir a interpretaciones distintas [18].

En este contexto, resulta esencial establecer criterios que permitan evaluar si las explicaciones generadas son fiables, coherentes y útiles. Entre los aspectos más relevantes se encuentran la fidelidad, la robustez, la transferibilidad y la usabilidad. En este trabajo, se abordan dos de ellos de manera específica: la fidelidad y la robustez [8].

2.2.2.1 Fidelidad

La fidelidad mide hasta qué punto una explicación refleja con precisión el comportamiento real del modelo. En otras palabras, se considera que una explicación es fiel si las características que identifica como importantes son efectivamente las que más influyen en la predicción del modelo.

Por ejemplo, si una explicación asigna una alta importancia a la variable *edad* en un modelo de predicción de riesgo cardiovascular, se espera que eliminar dicha variable o alterarla provoque un cambio significativo en las predicciones.

Existen diferentes metodologías para evaluar la fidelidad, entre las que destacan:

- Sanity Checks [27]: consisten en aleatorizar los pesos del modelo o las etiquetas de los datos. Si tras esta aleatorización las explicaciones permanecen inalteradas respecto a las originales, se considera que la explicación no depende del modelo real y, por tanto, carece de fidelidad.
- ROAR (*RemOve And Retrain*) [28]: evalúa cómo afecta al rendimiento del modelo la eliminación de las características que la explicación considera más importantes. Si al eliminar dichas características el rendimiento del modelo cae significativamente, en mayor medida que al eliminar características aleatorias, se considera que la explicación es fiel.

Este tipo de evaluación se considera fundamental, ya que una explicación que no refleje adecuadamente el comportamiento del modelo puede inducir a interpretaciones erróneas y decisiones equivocadas.

2.2.2.2 Robustez

La robustez evalúa la estabilidad de las explicaciones frente a pequeñas variaciones en los datos o en los modelos. Una explicación se considera robusta si presenta resultados consistentes bajo condiciones similares. Tal y como se plantea en [8], la robustez puede abordarse desde dos perspectivas:

- Robustez a cambios en la distribución: evalúa si las instancias similares generan explicaciones similares. Es decir, pequeñas perturbaciones en las entradas no deberían provocar grandes cambios en las explicaciones. Una baja robustez en este contexto puede indicar que la explicación está sobreajustada a datos concretos, perdiendo su capacidad de generalización.
- Robustez a multiplicidad predictiva: analiza si distintos modelos, entrenados sobre el mismo conjunto de datos y con un rendimiento comparable, generan explicaciones coherentes entre sí. Si las explicaciones son muy diferentes, se puede concluir que dependen en

exceso de las particularidades del modelo y no reflejan adecuadamente los patrones en los datos.

Para evaluar la robustez se pueden emplear diferentes métricas, como:

- Coeficientes de correlación (Pearson o Spearman), que miden el grado de acuerdo en el orden de importancia de las variables entre diferentes explicaciones.
- NDCG (*Normalized Discounted Cumulative Gain*), que compara rankings ponderando las discrepancias en las posiciones más altas del ranking, dando más peso a las características importantes.
- Aproximación de Lipschitz, que cuantifica la sensibilidad de las explicaciones antes pequeñas perturbaciones en las instancias de entrada.

Además, la robustez se extiende al concepto de transferibilidad, entendido como la capacidad de las explicaciones generadas sobre el conjunto de entrenamiento para mantenerse válidas a datos no vistos. Esta propiedad puede ser relevante en entornos donde la distribución de los datos puede experimentar ligeras variaciones a lo largo del tiempo.

2.2.2.3 Metodologías de validación

La validación de las explicaciones es uno de los principales retos en el ámbito de la IAX. Durante años, muchos trabajos se han basado en estrategias cualitativas, como la inspección visual de ejemplos individuales o la validación por parte de expertos, lo que puede resultar insuficiente y conducir a conclusiones poco fiables [29, 30]. Para superar estas limitaciones, algunos autores han propuesto metodologías basadas en datos sintéticos o la aplicación de metodologías específicas de ciertos dominios, como las directrices clínicas para imágenes médicas [31, 32]. Sin embargo, estas metodologías no siempre son extrapolables a problemas del mundo real.

Frente a estas limitaciones, el artículo de referencia [8] propone una metodología de validación sistemática, centrada en dos dimensiones: fidelidad y robustez, cada una evaluada a través de dos bloques. En total, la metodología se estructura en cuatro pruebas principales:

- Evaluación de **fidelidad**:
    - Sanity Checks: verifican que las explicaciones dependan realmente del modelo y de los datos. Consisten en pruebas de aleatorización de los datos (*data randomization*) y de los parámetros del modelo (*model randomization*). Si las explicaciones permanecen inalteradas tras estas perturbaciones, se considera que no son fieles al modelo.

        - ROAR: evalúa si las características identificadas como importantes por la explicación son efectivamente relevantes para el modelo. Para ello, se analiza cómo se degrada el rendimiento al eliminar progresivamente las variables con mayor importancia, comparándolo con la eliminación de variables seleccionadas aleatoriamente.
- Evaluación de **robustez**:
    - Robustez a multiplicidad predictiva: analiza la estabilidad de las explicaciones cuando existen múltiples modelos que alcanzan un rendimiento predictivo similar. Se espera que estos modelos generen explicaciones consistentes. La similitud entre los vectores de importancia se cuantifica mediante métricas como el NDCG y RMSE.
    - Robustez a cambios en la distribución: evalúa si las explicaciones se mantienen consistentes al generarse sobre diferentes particiones del mismo conjunto de datos. Esta prueba permite estimar la capacidad de las explicaciones para generalizar frente a ligeras variaciones en la distribución, simulando escenarios de aplicación en datos futuros o en producción.

Este marco metodológico permite una evaluación rigurosa de las explicaciones, abordando tanto su fidelidad al modelo como su robustez ante la variabilidad del problema, y constituye la base sobre la que se desarrolla el presente trabajo.

3 Desarrollo

Este capítulo recoge el desarrollo completo del presente trabajo, abordando tanto la metodología seguida como el diseño experimental implementado para alcanzar los objetivos planteados. En primer lugar, se presentan los conjuntos de datos seleccionados y los modelos utilizados en el estudio. Posteriormente, se detalla el *pipeline* diseñado para el entrenamiento de modelos con distintos niveles de rendimiento, introduciendo ruido de manera controlada. Finalmente, se presentan las técnicas de explicabilidad utilizadas, así como la métrica aplicada para evaluar la fidelidad y la robustez de las explicaciones generadas.

3.1 Metodología

Es importante recordar que en este trabajo no se tiene como objetivo encontrar modelos de máximo rendimiento como tal, sino analizar cómo la pérdida de precisión afecta a la robustez y la fidelidad de las explicaciones generadas. Este enfoque adopta una perspectiva observacional, en la que la degradación del modelo mediante la introducción de ruido permite estudiar cómo varían las explicaciones generadas por técnicas como SHAP y LIME. Este análisis es especialmente relevante en el contexto de la multiplicidad de modelos, donde diferentes modelos pueden alcanzar una precisión similar, pero ofrecer explicaciones diferentes.

La metodología seguida se basa en un enfoque experimental que permite analizar la relación entre la precisión de los modelos y la estabilidad de sus explicaciones bajo diferentes condiciones. Este planteamiento metodológico se inspira en el trabajo de referencia [8], que analiza la capacidad de las explicaciones para generalizar a datos no vistos. A partir de dicha base, este trabajo amplía el enfoque introduciendo un procedimiento de degradación controlada del rendimiento de los modelos, con el objetivo de estudiar cómo dicha degradación afecta a la fidelidad y la robustez de las explicaciones generadas.

La degradación se lleva a cabo aplicando ruido únicamente a los datos de entrenamiento, mientras que el conjunto de test se mantiene intacto. Este procedimiento permite ajustar el desempeño de los modelos a cinco niveles predefinidos de AUROC: 90, 80, 70, 60 y 50. De este modo, se simulan escenarios en los que el modelo pierde capacidad de predecir, pero de forma controlada y manteniendo constante su arquitectura y configuración.

Para el análisis se emplean tres modelos de aprendizaje automático: XGBoost, SVM y MLP, combinados con dos técnicas post-hoc de explicabilidad: SHAP y LIME. Ambas proporcionan explicaciones de tipo local, es decir, interpretaciones sobre cómo el modelo realiza una predicción para una instancia específica. Con el fin de obtener una visión global del comportamiento del modelo, las explicaciones generadas por SHAP y LIME se han promediado,

lo que permite construir un ranking global de importancia de características. Este enfoque facilita el análisis de tendencias generales, aunque supone perder parte de la información específica de cada instancia.

La evaluación de las explicaciones se aborda desde dos perspectivas:

- Por un lado, se analiza su fidelidad, entendida como la capacidad de las explicaciones para reflejar correctamente el comportamiento del modelo. En este contexto, la fidelidad se evalúa observando cómo varían las explicaciones a medida que el modelo pierde precisión. Se comparará las explicaciones obtenidas en la carpeta con mayor rendimiento (normalmente AUROC = 90) con las de las carpetas degradadas, analizando si son significativamente diferentes.
- Por otro lado, se evalúa su robustez, es decir, la estabilidad de las explicaciones frente a distintas particiones del mismo conjunto de datos, pero manteniendo el mismo nivel de rendimiento (es decir, dentro de la misma carpeta de AUROC). Una explicación se considera robusta si las características identificadas como relevantes se mantienen consistentes entre particiones.

Para cuantificar la similitud entre las explicaciones en ambos casos, se emplea la métrica NDCG (*Normalized Discounted Cumulative Gain*), que permite comparar rankings de importancia de características, dando mayor peso a las primeras posiciones del ranking, es decir, a las variables más influyentes según el modelo.

Este análisis permite estudiar cómo la pérdida progresiva de rendimiento afecta a la calidad de las explicaciones generadas. Se pretende identificar si existe un punto en el que las explicaciones dejan de ser lo suficientemente fiables como herramienta de interpretación, ya sea porque pierden fidelidad, o porque disminuye su robustez. En definitiva, este enfoque permite evaluar si hay un nivel de rendimiento a partir del cual deja de ser razonable confiar en en las explicaciones generadas por los modelos, es decir, si existe un umbral de rendimiento a partir del cual dejan de ser útiles o interpretables.

3.2 Diseño experimental

3.2.1 Conjuntos de datos

Para llevar a cabo el estudio se han seleccionado 5 conjuntos de datos, escogidos por su uso en investigaciones previas relacionadas con la explicabilidad en inteligencia artificial [8, 33, 34]. Cabe destacar que todos ellos corresponden a problemas de clasificación binaria y presentan un formato de datos tabular.

El conjunto de datos **Dementia** [35] contiene información longitudinal de 150 sujetos, con edades comprendidas entre 60 y 96 años, sometidos a entre dos y cuatro resonancias magnéticas a lo largo del tiempo. El dataset consta de 373 instancias y 15 variables, y permite clasificar a los pacientes con o sin demencia.

El conjunto **Breast Cancer Wisconsin (Diagnostic)** [36] incluye 569 instancias y 30 variables, que describen características extraídas de imágenes digitales obtenidas mediante aspiración con aguja fina (AAF) de masas mamarias. Su objetivo es distinguir entre tumores benignos y malignos.

El dataset **Heart Disease** [37], obtenido del repositorio UCI Machine Learning Repository, contiene información clínica de 303 pacientes y está compuesto por 13 variables. Corresponde específicamente al subconjunto de Cleveland, ampliamente utilizado en estudios de aprendizaje automático para la predicción de la enfermedad cardíaca.

Por último, se incluyen los dos conjuntos de datos utilizados en el trabajo de referencia [8]:

COVID19, que contiene 275 instancias y 191 variables, derivadas de datos de espectrometría de masas (MALDI-TOF) de muestras nasofaríngeas de pacientes, en el rango de 5 a 20 kDa.

Census Income [38], que cuenta con 48.842 instancias y 14 variables. Este conjunto permite predecir si los ingresos anuales de una persona superan los 50.000 dólares a partir de datos censales. Siguiendo la metodología del trabajo original [8], se ha utilizado únicamente un subconjunto aleatorio de 5000 (*) muestras debido a las limitaciones computacionales asociadas al entrenamiento de algunos modelos.

En adelante, y por simplicidad, los conjuntos de datos se referirán como *dementia*, *breast-cancer*, *heart-disease*, *covid* y *census-income*.

TABLA I

CONJUNTOS DE DATOS: NÚMERO DE INSTANCIAS Y VARIABLES

Dataset	Nº instancias	Nº variables
dementia	373	15
breast-cancer	569	30
heart-disease	303	13
covid	275	191
census-income	5.000 (*)	14

3.2.2 Modelos seleccionados

Asimismo, se han generado explicaciones utilizando tres tipos de modelos de aprendizaje supervisado entrenados sobre los diferentes conjuntos de datos: XGBoost (*eXtreme Gradient Boosting*), SVM (*Support Vector Machine*) y MLP (*Multi-Layer Perceptron*). Esta selección se debe tanto a su amplio uso en diversos contextos, como por haber sido empleados previamente en el artículo de referencia [8].

Cada modelo pertenece a una familia distinta de algoritmos, lo que permite analizar el comportamiento de las explicaciones bajo distintas estructuras internas y grados de complejidad. A continuación, se describen brevemente sus principales características:

- **XGBoost**

XGBoost es una implementación optimizada del algoritmo de árboles de decisión potenciados por gradiente (*gradient boosted decision trees*) [39]. Este método consiste construye modelos secuenciales donde cada nuevo árbol busca corregir los errores del anterior, optimizando una función de pérdida a través de técnicas basadas en gradiente.

XGBoost destaca por su eficiencia, escalabilidad y capacidad para manejar valores ausentes. Además, incorpora técnicas de regularización (L1 y L2) que no están presentes en las implementaciones tradicionales de *boosting*, lo que contribuye a reducir el sobreajuste.

El funcionamiento del algoritmo se basa en ajustar iterativamente árboles de decisión donde cada árbol se entrena sobre los residuos (diferencia entre las predicciones del modelo y los valores reales) del modelo anterior. Este proceso se ilustra en la Fig. 5, donde se observa cómo cada iteración ajusta los pesos de los datos en función de los errores anteriores, mejorando progresivamente el rendimiento del modelo.

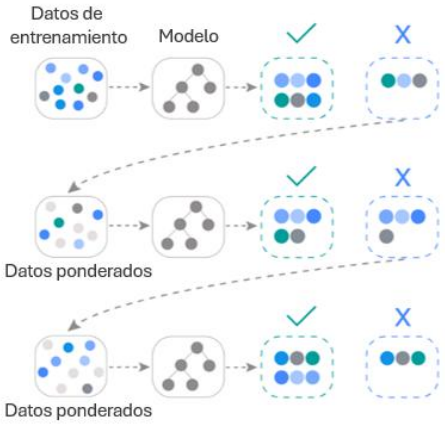


Fig. 5 Esquema del funcionamiento del algoritmo de *boosting* [39].

- **SVM**

Las Máquinas de Vectores de Soporte (SVM) son algoritmos supervisados que buscan encontrar el hiperplano óptimo que maximiza el margen entre clases en un espacio de N dimensiones [40]. Este hiperplano se define de forma que maximiza la distancia entre los puntos más cercanos de cada clase, los vectores de soporte, lo que mejora la capacidad de generalización del modelo. Este concepto se muestra en la Fig. 6.

Cuando los datos no son linealmente separables, como suele ocurrir en la práctica, se aplica el «*kernel trick*», que transforma los datos a un espacio de mayor dimensionalidad donde sí es posible encontrar un hiperplano separador. Entre los *kernels* más comunes se encuentran: lineal, polinómico, radial (RBF) y sigmoide.

Las SVM son especialmente eficaces en problemas de alta dimensionalidad y cuando el número de instancias es reducido con relación al número de características.

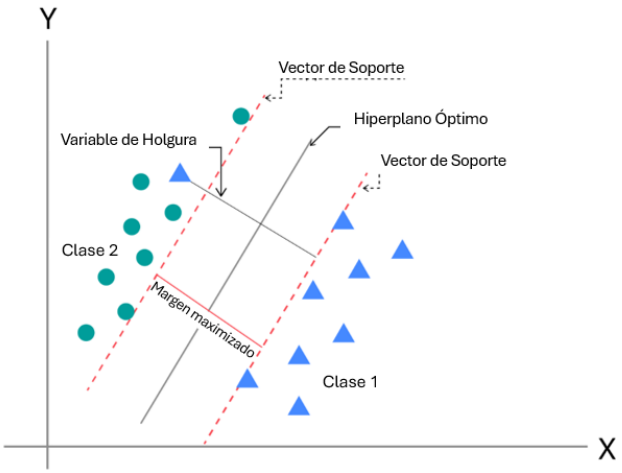


Fig. 6 Representación del hiperplano óptimo en SVM [40].

- **MLP**

El Perceptrón Multicapa (MLP) es un tipo de red neuronal de tipo *feedforward*, en la que los datos fluyen en una única dirección, desde la capa de entrada hacia la de salida [41]. Su estructura, representada en la Fig. 7, consta de una capa de entrada, una o varias capas ocultas y una capa de salida, con todas las neuronas interconectadas entre capas consecutivas.

Cada neurona calcula la suma ponderada de sus entradas, añade un sesgo y aplica una función de activación no lineal, lo que permite al modelo capturar relaciones complejas entre las variables. Sin estas funciones de activación, la red se comportaría como un modelo lineal.

El entrenamiento se realiza mediante el algoritmo de *backpropagation*, que ajusta los pesos de las conexiones minimizando la función de pérdida mediante el descenso del gradiente. Las capas ocultas permiten descomponer problemas complejos en representaciones progresivamente más abstractas, lo que hace que este tipo de redes sea capaz de abordar tareas como clasificación de imágenes, procesamiento del lenguaje natural, sistemas de recomendación o diagnósticos médicos.

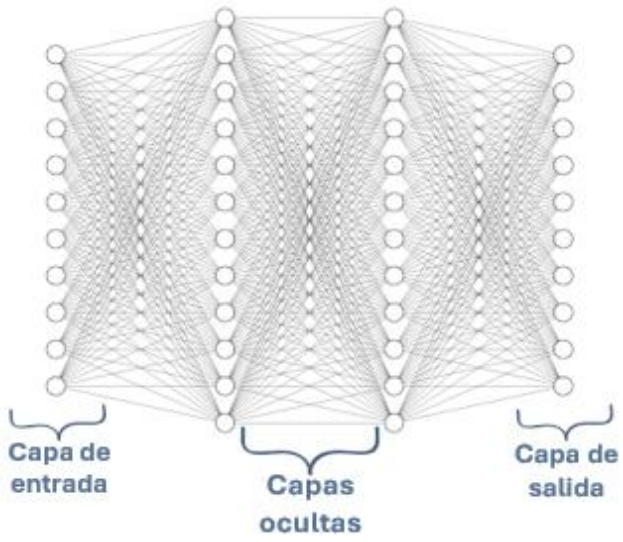


Fig. 7 Estructura de un Perceptrón Multicapa (MLP) [41].

3.2.3 Pipeline de entrenamiento y generación de modelos

Con el objetivo de garantizar la reproducibilidad, escalabilidad y estandarización del proceso experimental, se ha diseñado un pipeline que permite aplicar de forma sistemática los mismos pasos a todos los conjuntos de datos seleccionados.

Para ello, se ha implementado una clase denominada **DatasetHelper**, que centraliza las funciones específicas necesarias para cada dataset:

- **Lectura de datos:** cada conjunto de datos presenta un formato diferente (.csv, .xlsx, .data, etc.), por lo que dispone una función dedicada a su correcta carga.
- **Preprocesamiento:** incluye la limpieza de datos, transformación de variables y preparación de los datos de entrada. Esta función devuelve el conjunto de características (dfX) y la variable objetivo (dfy) listos para el entrenamiento.
- **Definición de modelos:** cada dataset cuenta con funciones específicas para instanciar los modelos XGBoost, SVM y MLP, integrando una búsqueda de hiperparámetros mediante validación cruzada (cv=10) con el objetivo de optimizar el rendimiento en función del AUROC.

El espacio de búsqueda de hiperparámetros se ha definido siguiendo las recomendaciones del tutor, buscando un equilibrio entre eficiencia computacional y rendimiento del modelo. Además, todos los modelos son inicializados con una semilla aleatoria fija para garantizar la reproducibilidad del experimento.

Una vez preparados los datos, se generan 20 particiones train/test (identificadas como ps = 0 a 19), utilizando el 80% de los datos para entrenamiento y el 20% restante para prueba. Los tres modelos seleccionados (XGBoost, SVM y MLP) se entrenan sobre cada partición del conjunto de entrenamiento, aplicando el nivel de ruido correspondiente. La introducción controlada de ruido permite ajustar el rendimiento del modelo hasta alcanzar un valor de AUROC lo más próximo posible al objetivo definido, como se detalla en el apartado 3.2.4.

Finalizado el entrenamiento, se generan las explicaciones correspondientes para cada modelo utilizando SHAP y LIME, posteriormente se calcula el NDCG, que permite evaluar la fidelidad y robustez de las explicaciones obtenidas en cada carpeta de AUROC y para cada clasificador.

Todo el pipeline, Fig. 8, está diseñado para ejecutarse de forma iterativa sobre cada uno de los conjuntos de datos. Esto permite aplicar automáticamente el mismo procedimiento de carga, preprocesamiento, particionado, entrenamiento, degradación y evaluación de explicaciones, garantizando la consistencia y la comparabilidad entre los distintos datasets y modelos.

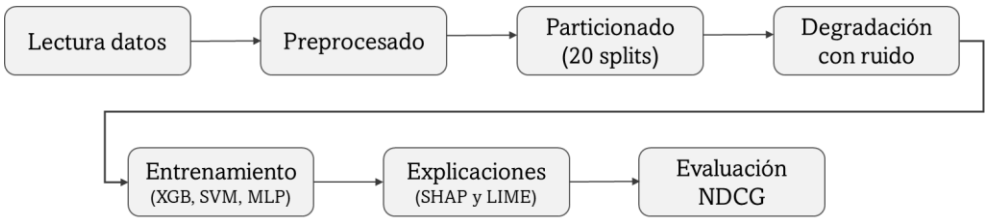


Fig. 8 Pipeline experimental desarrollado. El flujo incluye las etapas de lectura y preprocesamiento de los datos, particionado y degradación con ruido, entrenamiento de modelos (XGBoost, SVM y MLP), generación de explicaciones (SHAP y LIME) y evaluación de la robustez mediante NDCG.

Cabe señalar que no todos los conjuntos de datos ni todos los modelos logran alcanzar los cinco niveles de AUROC planteados. Este comportamiento, ya observado en [8], se confirma igualmente en este trabajo. En concreto, en el dataset *census-income*, el modelo XGBoost alcanza como máximo un AUROC de 80, mientras que SVM y MLP solo llegan a 70. De manera similar, en el dataset *heart-disease*, el modelo MLP no consigue alcanzar el AUROC 90, por lo que el análisis entre niveles se realiza tomando como referencia la carpeta de 80 para las comparaciones (80-70, 80-60 y 80-50), criterio que también se aplica en el caso de *census-income*.

Asimismo, con el fin de garantizar la fiabilidad de los análisis de fidelidad entre niveles, se establece como criterio que las comparaciones se lleven a cabo únicamente cuando existan al menos 5 *splits* válidos coincidentes entre los niveles considerados. Por ejemplo, en la comparación entre AUROC 90 y 80, deben existir al menos cinco particiones donde ambos niveles estén disponibles para la misma partición (90/p=0 vs 80/p=0, 90/p=1 vs 80/p=1, etc.). En caso de no alcanzar este mínimo, se verifica si es posible realizar la comparación tomando como referencia la carpeta de AUROC inmediatamente inferior.

3.2.4 Método de degradación del modelo mediante ruido

Con el fin de analizar el comportamiento de las explicaciones generadas por los modelos bajo distintos niveles de rendimiento, se ha diseñado un procedimiento de degradación controlada. Este proceso permite ajustar la capacidad predictiva de cada modelo a valores concretos de **AUROC** (*Area Under the Receiver Operating Characteristic*).

La métrica de evaluación empleada es el área bajo la curva ROC (AUROC), que compara la tasa de verdaderos positivos y falsos positivos. Esta métrica es especialmente adecuada puesto que es invariante frente a desequilibrios en la distribución de clases, asegurando una evaluación justa y representativa del rendimiento del modelo.

La degradación se realiza introduciendo ruido de tipo impulso (*impulse noise*), que consiste en reemplazar un porcentaje de los valores originales de cada variable por valores aleatorios generados dentro del rango definido por sus valores mínimo y máximo.

El procedimiento se aplica de forma iterativa siguiendo los siguientes pasos:

1. Para cada conjunto de datos, se generan 20 particiones train/test.
2. Sobre cada partición, se entrena cada uno de los tres modelos (XGBoost, SVM y MLP) con diferentes niveles de ruido, explorando porcentajes comprendidos entre 0% y 100%, en incrementos de 2%.
3. Para cada nivel de ruido, se calcula el AUROC en el conjunto de test.
4. Se selecciona el nivel de ruido que maximiza el número de particiones cuyo AUROC se encuentra dentro de una ventana de tolerancia de ± 5 puntos respecto al valor objetivo.
5. Una vez identificado este nivel de ruido, se generan y almacenan los modelos correspondientes a cada uno de los cinco niveles objetivo de AUROC: 90, 80, 70, 60 y 50.

No obstante, es importante señalar que no siempre es posible alcanzar todos los niveles de AUROC definidos. Dependiendo de las características del dataset y del modelo, puede ocurrir que ciertos niveles, especialmente los más altos, no sean alcanzables, como sucede en el dataset *census-income*, donde XGBoost

solo llega a un AUROC de 80 y SVM y MLP a 70, o en *heart-disease* con MLP, que no alcanza el nivel de 90. En estos casos, no se generan modelos en las carpetas correspondientes a esos niveles.

Para cada base de datos y cada clasificador, se entrenaron 100 modelos, organizados en 5 carpetas según el nivel objetivo de AUROC (90, 80, 70, 60 y 50), con 20 modelos en cada carpeta, correspondientes a las particiones train/test.

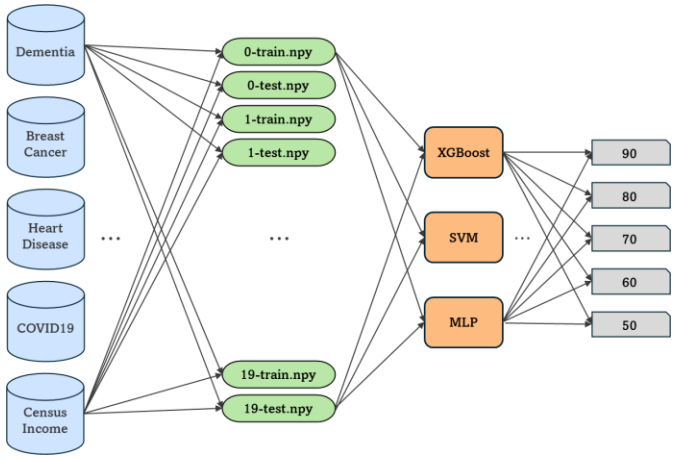


Fig. 9 Esquema del proceso de entrenamiento y degradación de los modelos. Cada conjunto de datos se divide en 20 particiones train/test para entrenar tres tipos de modelos (XGBoost, SVM y MLP), agrupados según cinco niveles objetivo de AUROC.

En consecuencia, el estudio comprende un máximo de 1500 modelos entrenados, resultado de la combinación de 5 conjuntos de datos × 20 particiones × 3 modelos × 5 niveles de AUROC. Esta estructura permite analizar de forma sistemática cómo varían las explicaciones generadas por los modelos a medida que su rendimiento se degrada.

3.2.5 Métricas de evaluación de la explicabilidad

La evaluación de la calidad de las explicaciones se ha realizado usando la métrica **NDCG**, ampliamente utilizada en problemas de ranking y adaptada en este contexto para medir la similitud entre vectores de importancia de características. Permite cuantificar hasta qué punto se mantienen las posiciones relativas de las características más relevantes cuando el rendimiento del modelo empeora o cuando se entrenan modelos distintos, pero con el mismo rendimiento. Se han considerado dos escenarios de evaluación:

- **Fidelidad (NDCG inter-nivel)**: evalúa la similitud de las explicaciones de un modelo con alto rendimiento (normalmente con AUROC 90) respecto a sus versiones degradadas (80, 70, 60, 50). La comparación se realiza sobre los mismos *splits*, es decir 90/p=0 vs 80/p=0, 90/p=1 vs 80/p=1,

y así sucesivamente. Este análisis permite comprobar si las explicaciones son significativamente diferentes.

- **Robustez (NDCG intra-nivel):** mide la consistencia de las explicaciones entre diferentes splits dentro de un mismo nivel de AUROC. Un valor alto indicaría que, aunque las particiones sean distintas, el modelo identifica de forma consistente las mismas características como relevantes, a pesar de reducirse el rendimiento.

En ambos casos, el NDCG se calcula sobre los vectores promedio de importancia de características en cada combinación de dataset, modelo, nivel de AUROC y *split*. Los resultados se almacenan de forma estructurada para facilitar su posterior análisis y visualización.

Cálculo del NDCG

El NDCG se basa en el cálculo previo del *Discounted Cumulative Gain* (DCG), que pondera la relevancia de cada elemento según su posición en el ranking:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

donde rel_i es la relevancia del elemento en la posición i.

Este valor se normaliza dividiendo por el *Ideal Discounted Cumulative Gain* (IDCG), que corresponde al DCG obtenido cuando los elementos están perfectamente ordenados según su relevancia:

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

El valor de NDCG está comprendido entre 0 y 1, donde 1 indica un ranking perfecto. Esta normalización permite realizar comparaciones coherentes entre diferentes rankings o modelos.

No obstante, esta métrica presenta ciertas limitaciones, ya que no penaliza directamente la aparición de elementos irrelevantes en el ranking, y su valor depende de la calidad y completitud de las puntuaciones de relevancia proporcionadas [42, 43].

4 Resultados

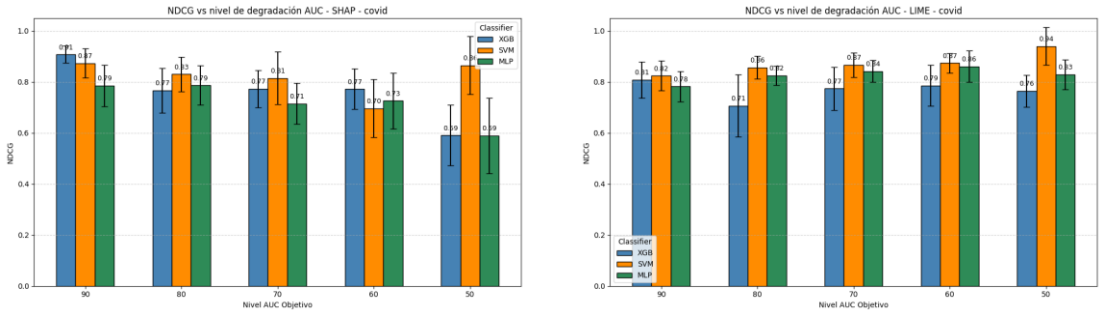
En esta sección se presentan los resultados obtenidos tras aplicar la metodología propuesta. Se analizan tanto la robustez como la fidelidad de las explicaciones generadas con SHAP y LIME en función del rendimiento de los modelos.

4.1 Robustez de las explicaciones

En este apartado se analiza cómo afecta la pérdida de rendimiento del modelo a la robustez de las explicaciones generadas. El objetivo es comprobar si, a medida que disminuye el AUROC, las explicaciones se vuelven menos estables entre diferentes particiones del mismo conjunto de datos. Estas particiones simulan ligeros cambios en la distribución de los datos, comparables con datos futuros, cuya distribución probablemente sea ligeramente diferente. Una menor robustez indicaría que el modelo deja de identificar de forma consistente las mismas características como relevantes.

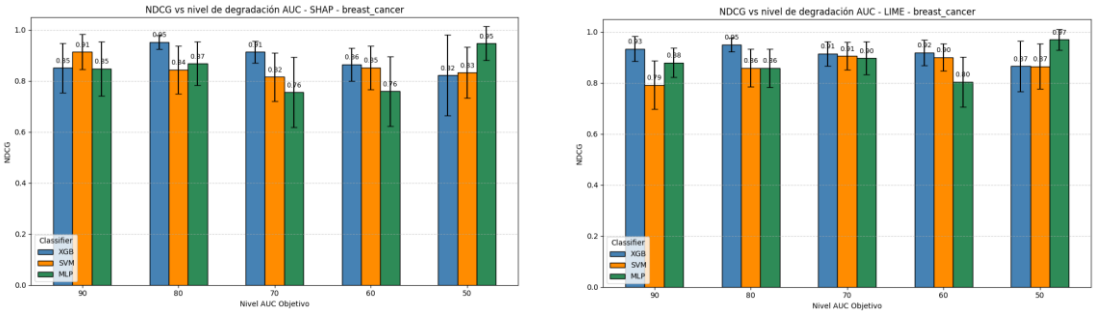
Cabe destacar que no todos los modelos han logrado alcanzar los cinco niveles de AUROC planteados (90, 80, 70, 60 y 50), tal y como se mencionó en la sección anterior. Por ejemplo, en el dataset *census-income*, los modelos XGBoost, SVM y MLP no alcanzan el nivel 90, quedándose en un máximo de 80 o 70 según el clasificador. De manera similar, en *heart-disease*, el modelo MLP no consigue alcanzar un AUROC de 90. Esta limitación se refleja en algunas gráficas, especialmente en los niveles superiores, donde ciertos puntos aparecen ausentes al no haberse alcanzado el rendimiento deseado.

Para cada dataset, se calculó el NDCG medio entre las explicaciones globales de todos los subconjuntos dentro de cada carpeta de AUROC, como se muestra en las figuras Fig. 10, Fig. 11 y las del Anexo A.1.



(a) SHAP (b) LIME

Fig. 10 NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC en el dataset *covid*, generadas con (a) SHAP y (b) LIME.



(a) SHAP (b) LIME

Fig. 11 NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC en el dataset *breast-cancer*, generadas con (a) SHAP y (b) LIME.

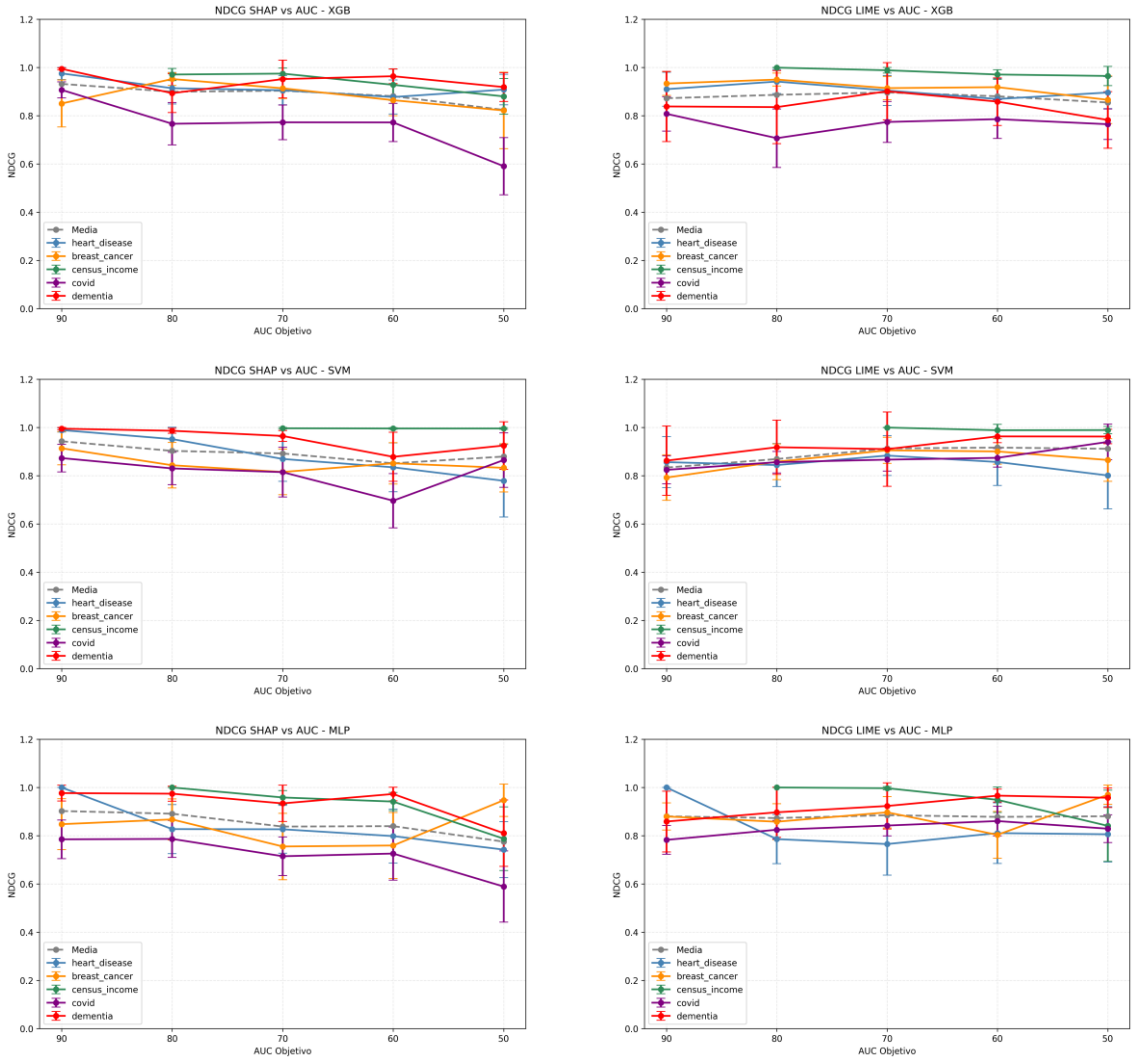
En el dataset *covid*, las explicaciones generadas con SHAP para XGB y MLP presentan cierta tendencia decreciente a medida que el rendimiento de los modelos disminuye, mientras que en SVM no se observa este comportamiento. Por el contrario, con LIME se aprecia un aumento del NDCG en SVM y MLP, mientras que en XGB se mantiene estable. Este comportamiento puede deberse a la naturaleza dispersa del dataset, como se señala en [8], donde se concluye que SHAP funciona mejor en dataset dispersos de baja densidad, mientras que LIME ofrece mejores resultados en conjuntos de datos muy densos.

En el caso del dataset *breast-cancer*, no se aprecia una tendencia clara en las explicaciones generadas por SHAP. XGBoost y MLP no muestran un patrón definido, esto mismo ocurre para las demás bases de datos, cuyos resultados se muestran en el Anexo A.1. Aun así, las explicaciones generadas por SHAP para SVM parecen mostrar un leve descenso. Para LIME, XGBoost es el único modelo que muestra una ligera tendencia decreciente, mientras que SVM y MLP no decrecen.

De forma general, y considerando también los resultados mostrados en las figuras del Anexo A.1, no se observa una tendencia consistente de que el NDCG disminuya sistemáticamente a medida que se degrada el modelo. Este efecto parece ser más evidente únicamente en los datasets *covid* y *heart-disease*, donde sí se aprecia un descenso del NDCG conforme se reduce el AUROC.

Finalmente, para ofrecer una visión global del comportamiento de la robustez en los cinco conjuntos de datos, se presenta la Fig. 12. En ella se muestran los resultados agregados por dataset: la columna izquierda corresponde a las explicaciones generadas con SHAP, y la de la derecha, a las generadas con LIME. Cada fila representa un clasificador en el siguiente orden: XGBoost, SVM y MLP. Además, en cada gráfico se incluye una línea discontinua que representa la media del NDCG en cada nivel de AUROC, lo que permite visualizar de forma más clara la tendencia general.

En general, se aprecian ciertas tendencias como las descritas anteriormente para *covid* y *breast-cancer*, pero no es posible generalizarlas al resto de conjuntos de datos. Esto queda reflejado en la línea discontinua que representa la media, la cual corrobora que no existe un patrón claro ni consistente en la evolución de la robustez en función del AUROC.



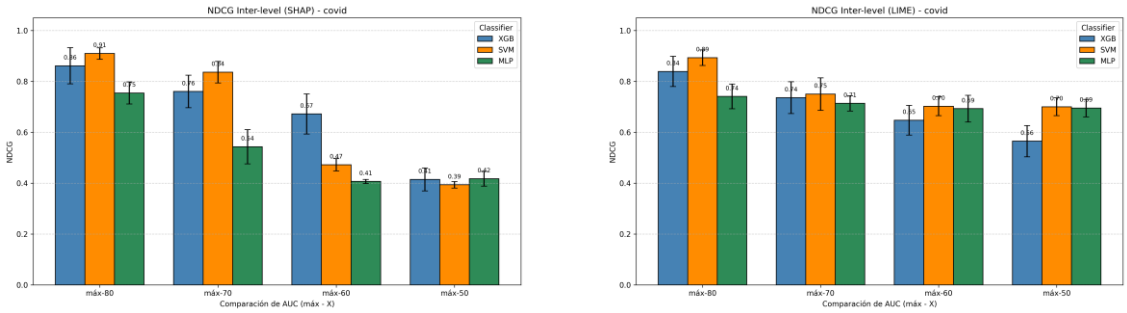
(a) SHAP (b) LIME

Fig. 12 NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC para todos los datasets, generadas con (a) SHAP y (b) LIME.

4.2 Fidelidad de las explicaciones

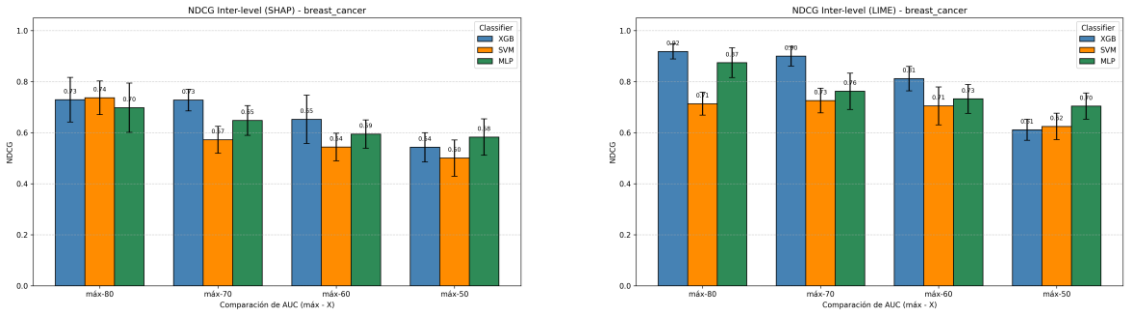
En este apartado se analiza la fidelidad de las explicaciones, es decir, la capacidad de mantenerse consistentes respecto a las generadas por el modelo con mayor rendimiento alcanzado. Este modelo se toma como referencia, o *baseline*, el nivel más alto de AUROC disponible para cada combinación de dataset y clasificador, que como se ha explicado anteriormente, no siempre es 90 debido a las limitaciones específicas de algunos modelos o conjuntos de datos. Cabe mencionar también que, tal como se indicó en el desarrollo, el mínimo de splits coincidentes entre niveles de AUROC es de 5. En caso de no alcanzarse este mínimo, se utilizaría la siguiente carpeta disponible con suficientes splits válidos.

El NDCG medio entre las explicaciones del modelo de referencia y las de los modelos degradados en cada uno de los niveles de AUROC inferiores, se muestran en las Fig.13, Fig.14 y las del Anexo A.2. A partir de estas gráficas se evalúa la fidelidad.



(a) SHAP (b) LIME

Fig. 13 NDCG medio entre las explicaciones del modelo de referencia y las de los modelos degradados en el dataset *covid*, generadas con (a) SHAP y (b) LIME.



(a) SHAP (b) LIME

Fig. 14 NDCG medio entre las explicaciones del modelo de referencia y las de los modelos degradados en el dataset *breast-cancer*, generadas con (a) SHAP y (b) LIME.

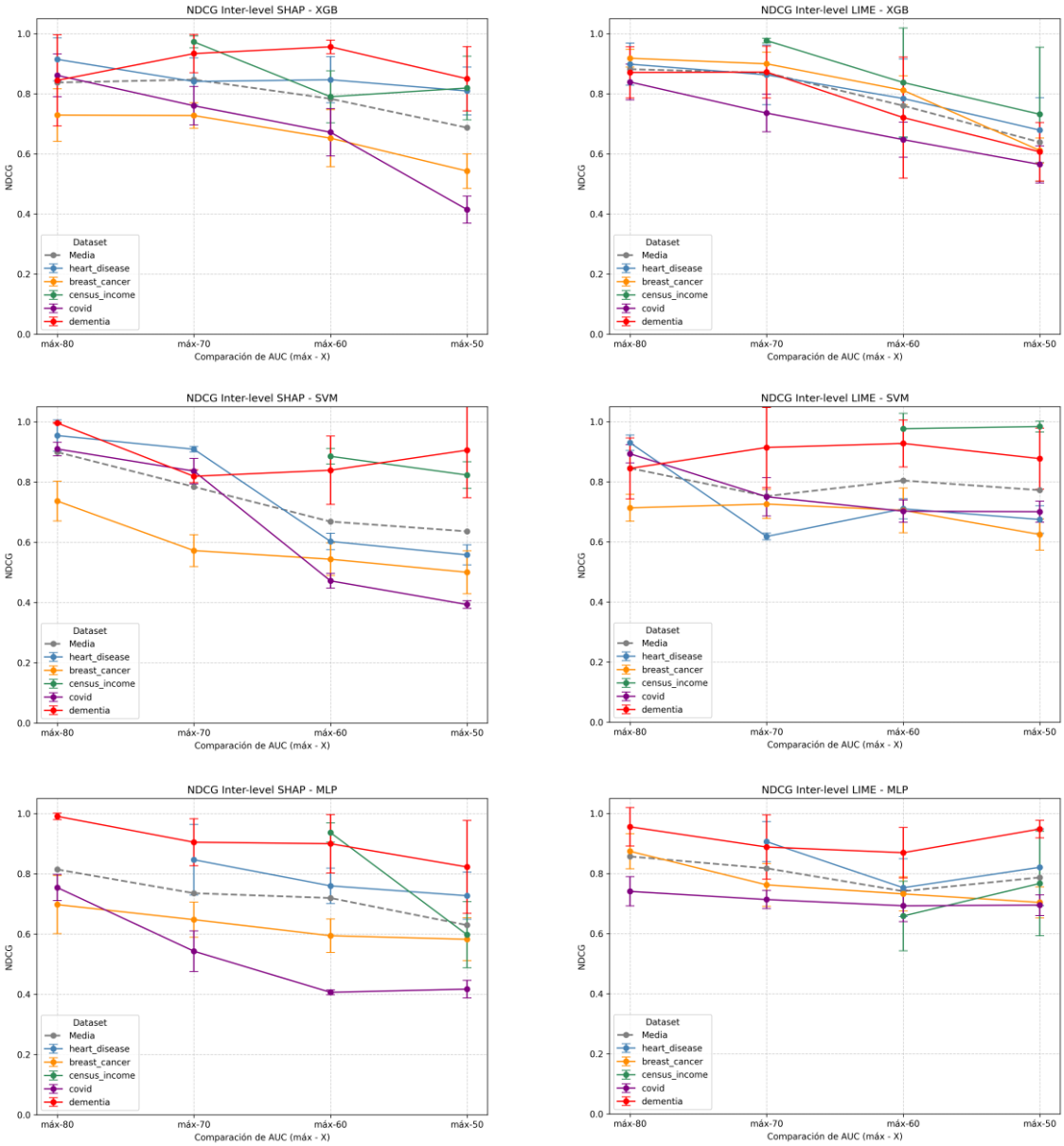
En el dataset *covid*, las explicaciones generadas por SHAP presentan una clara tendencia descendiente en el NDCG conforme disminuye el AUROC, comportamiento que se observa de manera consistente en los tres clasificadores. Un patrón similar, aunque menos pronunciado, se aprecia en las explicaciones generadas con LIME, donde el NDCG disminuye también en todos los modelos.

En caso del dataset *breast-cancer*, las explicaciones generadas por SHAP muestran una tendencia descendente más suave, aunque igualmente presente en los tres clasificadores. Para LIME, tanto XGBoost como MLP presentan una disminución gradual del NDCG a medida que el rendimiento del modelo se degrada, mientras que SVM se mantiene relativamente estable hasta la comparación con la carpeta de AUROC 50, donde se produce un descenso más acentuado.

Este comportamiento parece repetirse para los demás dataset del Anexo A.2, salvo excepciones como *dementia*, en la que, para las explicaciones generadas con LIME, SVM presenta una tendencia creciente en lugar de decreciente.

Finalmente, para ofrecer una visión global del comportamiento de la fidelidad de los cinco conjuntos de datos, se presenta la Fig.15. En ella se muestran los resultados agregados por dataset: la columna de la izquierda corresponde a las explicaciones generadas con SHAP y la de la derecha a las generadas con LIME. Cada fila representa un clasificador en el siguiente orden: XGBoost, SVM y MLP. Además, en cada gráfico se incluye una línea discontinua que representa la media del NDCG en cada nivel de AUROC.

En términos generales, sí se aprecia una disminución de la fidelidad generalizada para todos los conjuntos de datos y modelos, con la excepción del dataset *dementia*, donde las explicaciones generadas por SHAP para XGB y SVM, y por LIME para SVM y MLP, presentan una tendencia creciente en ciertos puntos. Sin embargo, para todas las bases de datos, se puede concluir que las explicaciones generadas para los modelos degradados difieren significativamente de las generadas por los modelos baseline. Es decir, a medida que el AUROC empeora, las explicaciones del modelo comienzan a identificar características cada vez más diferentes.



(a) SHAP (b) LIME

Fig. 15 NDCG medio inter-nivel entre las explicaciones de los modelos de referencia (mayor AUROC alcanzado) y las de los modelos degradados para todos los datasets, generadas con (a) SHAP y (b) LIME.

5 Conclusiones y Trabajo Futuro

5.1 Conclusiones

En términos generales, las explicaciones de los modelos analizados muestran ser fidedignas, ya que siguen una tendencia de degradación conforme disminuye el AUROC. Esto indica que, a medida que el rendimiento del modelo se reduce, las explicaciones pierden consistencia con respecto a las generadas por el modelo de mayor precisión (*baseline*), es decir, las características que explican son significativamente diferentes. No obstante, la robustez de las explicaciones no se degrada de manera generalizada, esta degradación solo ocurre en los datasets de *covid* y *breast-cancer*. En estos casos, la fidelidad también se había degradado, lo cual, según la metodología del artículo [8], es una condición necesaria para considerar los resultados de robustez como válidos.

En cuanto a los factores que podrían explicar estas degradaciones, especialmente en el caso de *covid*, la densidad de los datos parece ser un factor relevante. Tal como se menciona en el artículo, los datasets más dispersos tienden a mostrar una mayor variabilidad en las explicaciones.

Por último, aunque se observa una degradación en la fidelidad de las explicaciones a medida que disminuye el rendimiento de los modelos, no se encuentra una relación clara y consistente entre la disminución del AUROC y la degradación de la robustez. En base a este estudio, no se puede demostrar que la robustez de las explicaciones dependa exclusivamente de la precisión del modelo, sino que está influenciada por otros factores, como la estructura y naturaleza de los datos.

5.2 Trabajo futuro

Una mejora clave para futuros trabajos será optimizar el número de *splits* mínimos por carpeta para realizar las comparaciones más robustas. Actualmente, los resultados dependen de contar con suficientes particiones en cada nivel de AUROC, y aumentar el número de *splits* válidos permitirá obtener comparaciones más representativas.

Además de los datasets tabulares utilizados en este trabajo, sería interesante extenderlo a otros tipos de datos, como imágenes. Evaluar cómo las técnicas de explicabilidad, como SHAP y LIME (u otras que se adapten al caso de estudio), se comportan en dominios más complejos, como las imágenes, proporcionará una nueva dimensión al estudio de la robustez y fidelidad. El uso de modelos como CNNs (Redes Neuronales Convolucionales) para la clasificación de

imágenes permitirá explorar cómo afecta la relación entre precisión y explicabilidad en este tipo de datos.

A través de estas investigaciones futuras, se podrá no solo mejorar la metodología y aplicabilidad de las explicaciones generadas, sino también aumentar la confianza en el uso de modelos interpretables en diversas áreas del aprendizaje automático.

6 Análisis de Impacto

Este trabajo se alinea estrechamente con los Objetivos de Desarrollo Sostenible (ODS) establecidos por la Organización de las Naciones Unidas. En el ámbito social y cultural, se mejora la transparencia de los modelos de IA, lo que promueve la confianza pública en este tipo de tecnologías. Al hacer las decisiones de los modelos más explicables, se reduce el riesgo de discriminación, apoyando una sociedad más inclusiva y justa, alineada con el ODS 16.

En cuanto a la legislación europea, este trabajo está relacionado con el *AI Act*, que promueve la transparencia y la ética en el uso de la IA. Las técnicas de explicabilidad aplicadas contribuyen a la implementación de este marco normativo, garantizando que los modelos sean comprensibles y responsables, cumpliendo con las exigencias legales.

Desde la perspectiva de la transparencia en los modelos, este proyecto pretende ayudar a asegurar que los sistemas de IA no sean "cajas negras". Al proporcionar explicaciones claras y accesibles, se facilita la supervisión de resultados, lo que reduce los riesgos asociados con el uso de la IA en áreas críticas como la justicia (ODS 16), la sanidad (ODS 3) y la educación (ODS 4), contribuyendo a una sociedad más responsable y sostenible (ODS 9).

7 Bibliografía

[1] "Inteligencia Artificial," *Unesco.org*, 2022. https://www.unesco.org/es/artificial-intelligence

[2] T. Mucci, "History of artificial intelligence," *Ibm.com*, Oct. 21, 2024. https://www.ibm.com/think/topics/history-of-artificial-intelligence

[3] IBM, "Grandes modelos de lenguaje," *Ibm.com*, Nov. 02, 2023. https://www.ibm.com/es-es/think/topics/large-language-models?utm_source=chatgpt.com

[4] R. Merritt, "What Are Foundation Models?," *NVIDIA Blog*, Mar. 13, 2023. https://blogs.nvidia.com/blog/what-are-foundation-models/

[5] T. A. James, "Confronting the Mirror: Reflecting on Our Biases Through AI in Health Care," *Harvard.edu*, Sep. 24, 2024. https://postgraduateeducation.hms.harvard.edu/trends-medicine/confronting-mirror-reflecting-our-biases-through-ai-health-care

[6] "TechDispatch #2/2023 - Explainable Artificial Intelligence | European Data Protection Supervisor," *www.edps.europa.eu*, Mar. 15, 2024. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence_en

[7] European Union, "Regulation - EU - 2024/1689 - EN - EUR-Lex," *Europa.eu*, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[8] A. M. L. González and E. García-Cuesta, "On the transferability of local model-agnostic explanations of machine learning models to unseen data," 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), Madrid, Spain, 2024, pp. 1-10, doi: 10.1109/EAIS58494.2024.10570001.

[9] Prasad Pasam Thulasiram, "EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): ENHANCING TRANSPARENCY AND TRUST IN MACHINE LEARNING MODELS," *Philpapers.org*, 2025. https://philpapers.org/rec/THUEAI

[10] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, no. 101805, p. 101805, Apr. 2023, doi: https://doi.org/10.1016/j.inffus.2023.101805.

[11] Emrullah ŞAHiN, N. N. Arslan, and Durmuş Özdemir, "Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Computing and Applications*, Nov. 2024, doi: https://doi.org/10.1007/s00521-024-10437-2.

[12] M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," in IEEE Access, vol. 7, pp. 137184-137206, 2019, doi: 10.1109/ACCESS.2019.2942390.

[13] W. Ding, M. Abdel-Basset, H. Hawash, and A. M. Ali, “Explainability of Artificial Intelligence Methods, Applications and Challenges: A Comprehensive Survey,” *Information Sciences*, Oct. 2022, doi: https://doi.org/10.1016/j.ins.2022.10.013.

[14] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2239–2250. https://doi.org/10.1145/3531146.3534639

[15] L. Longo *et al.*, “Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions,” *Information Fusion*, vol. 106, p. 102301, Jun. 2024, doi: https://doi.org/10.1016/j.inffus.2024.102301.

[16] “Explainability II: global explanations, proxy models, and interpretable models - RBC Borealis,” *RBC Borealis*, Apr. 06, 2023. https://rbcborealis.com/research-blogs/explainability-ii-global-explanations-proxy-models-and-interpretable-models/

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” Feb. 2016, doi: https://doi.org/10.48550/arxiv.1602.04938.

[18] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *arXiv.org*, Nov. 24, 2017. https://arxiv.org/abs/1705.07874v2

[19] A. Fisher, C. Rudin, and F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously,” *arXiv.org*, Dec. 23, 2019. http://arxiv.org/abs/1801.01489 (accessed Feb. 20, 2024).

[20] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, “Distribution-Free Predictive Inference For Regression,” *arXiv (Cornell University)*, Jan. 2016, doi: https://doi.org/10.48550/arxiv.1604.04173.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: https://doi.org/10.1007/s11263-019-01228-7.

[22] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” *arXiv.org*, Oct. 12, 2019. https://arxiv.org/abs/1704.02685

[23] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: removing noise by adding noise,” *arXiv.org*, Jun. 12, 2017. https://arxiv.org/abs/1706.03825

[24] Christoph Molnar, “5.7 Local Surrogate (LIME) | Interpretable Machine Learning,” *Github.io*, Sep. 18, 2019. https://christophm.github.io/interpretable-ml-book/lime.html

[25] C. Molnar, *5.9 Shapley Values | Interpretable Machine Learning*. Available: https://christophm.github.io/interpretable-ml-book/shapley.html

[26] C. Molnar, *5.10 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning*. 2022. Available: https://christophm.github.io/interpretable-ml-book/shap.html

[27] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity Checks for Saliency Maps,” Oct. 2018, doi: https://doi.org/10.48550/arxiv.1810.03292.

[28] S. Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and B. Kim, “A Benchmark for Interpretability Methods in Deep Neural Networks,” *arXiv (Cornell University)*, Jan. 2018, doi: https://doi.org/10.48550/arxiv.1806.10758.

[29] F. Doshi-Velez and B. Kim, “Considerations for Evaluation and Generalization in Interpretable Machine Learning,” *The Springer Series on Challenges in Machine Learning*, pp. 3–17, 2018, doi: https://doi.org/10.1007/978-3-319-98131-4_1.

[30] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: https://doi.org/10.1073/pnas.1900654116.

[31] Y. Liu, S. Khandagale, C. White, and W. Neiswanger, “Synthetic Benchmarks for Scientific Research in Explainable Machine Learning,” *arXiv.org*, Nov. 04, 2021. https://arxiv.org/abs/2106.12543

[32] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, “Guidelines and evaluation of clinical explainable AI in medical image analysis,” *Medical Image Analysis*, p. 102684, Nov. 2022, doi: https://doi.org/10.1016/j.media.2022.102684.

[33] V. Sharma and D. Midhunchakkaravarthy, "XGBoost Classification of XAI based LIME and SHAP for Detecting Dementia in Young Adults," 2023 14th International Conference on Computing Communication and Networking

Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10307791.

[34] P. Paudel, R. Saud, S. K. Karna and M. Bhandari, "Determining the Major Contributing Features to Predict Breast Cancer Imposing ML Algorithms with LIME and SHAP," 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2023, pp. 1-7, doi: 10.1109/ICECET58911.2023.10389217.

[35] M. Mittal, F. Amenta, and Nalini Chintalapudi, "Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM)," vol. 1, Jul. 2019, doi: https://doi.org/10.17632/tsy6rbc5d4.1.

[36] W. Wolberg, O. Mangasarian, N. Street, and W. Street. "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1993. [Online]. Available: https://doi.org/10.24432/C5DW2B.

[37] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. "Heart Disease," UCI Machine Learning Repository, 1989. [Online]. Available: https://doi.org/10.24432/C52P4X.

[38] Kohavi, R. (1996). Census Income [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5GP7S.

[39] E. Kavlakoglu and E. Russi, "XGBoost," *Ibm.com*, May 09, 2024. https://www.ibm.com/think/topics/xgboost

[40] IBM, "Support Vector Machine," *IBM*, Dec. 12, 2023. https://www.ibm.com/think/topics/support-vector-machine

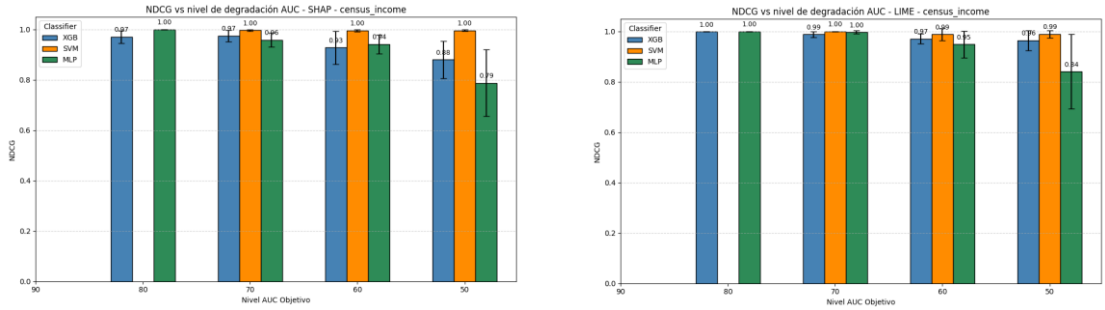
[41] Sidharth, "Multi-Layer Perceptron Explained: A Beginner's Guide," *QuarkML*, Jan. 25, 2023. https://www.quarkml.com/2023/01/multi-layer-perceptron-a-complete-overview.html

[42] "Normalized Discounted Cumulative Gain (NDCG) explained," *www.evidentlyai.com*. https://www.evidentlyai.com/ranking-metrics/ndcg-metric

[43] "Normalized Discounted Cumulative Gain (NDCG): Where To Use It," *Arize AI*, Mar. 07, 2024. https://arize.com/blog-course/ndcg/

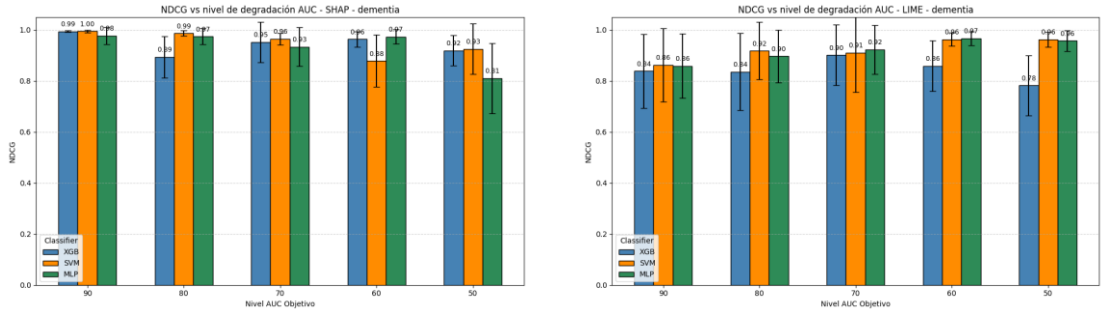
8 Anexos

A.1. Histogramas de robustez de las explicaciones



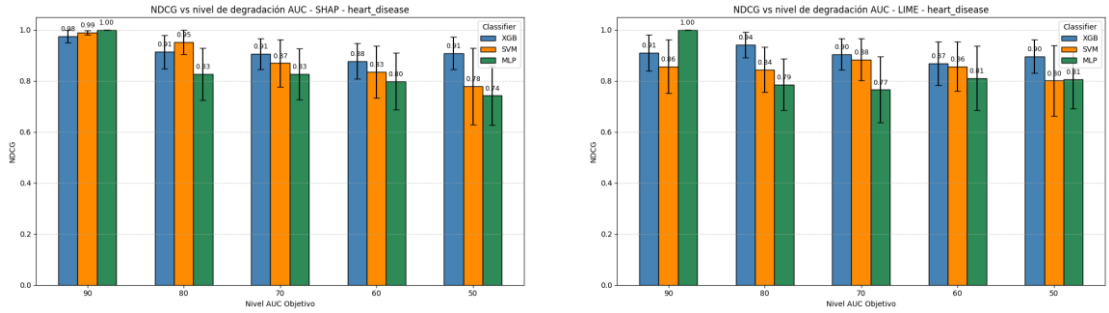
(a) SHAP (b) LIME

Fig. A.1.1 NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC en el dataset *census-income*, generadas con (a) SHAP y (b) LIME.



(a) SHAP (b) LIME

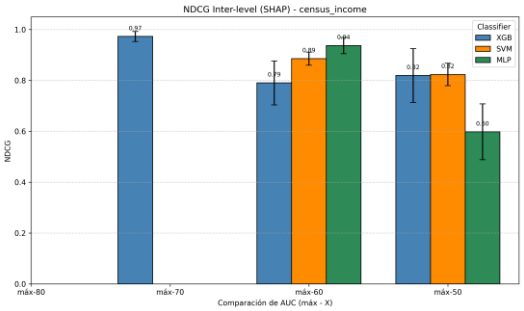
Fig. A.1.2. NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC en el dataset *dementia*, generadas con (a) SHAP y (b) LIME.



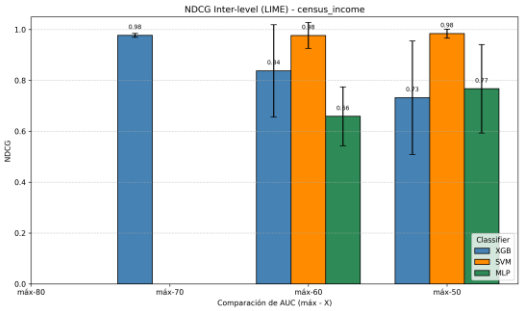
(a) SHAP (b) LIME

Fig. A.1.3. NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP por cada carpeta de AUROC en el dataset *heart-disease*, generadas con (a) SHAP y (b) LIME.

A.2. Histogramas de fidelidad de las explicaciones

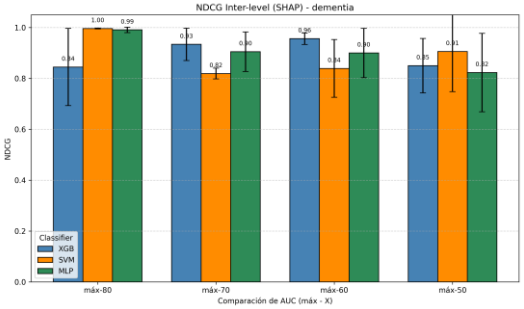


(a) SHAP

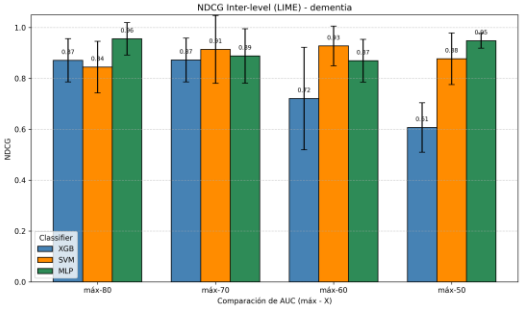


(b) LIME

Fig. A.2.1. NDCG medio entre las explicaciones del modelo de referencia (mayor AUROC alcanzado) y las de los modelos degradados en el dataset *census-income*, generadas con (a) SHAP y (b) LIME.

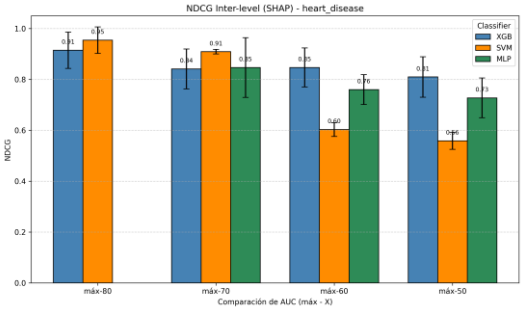


(a) SHAP

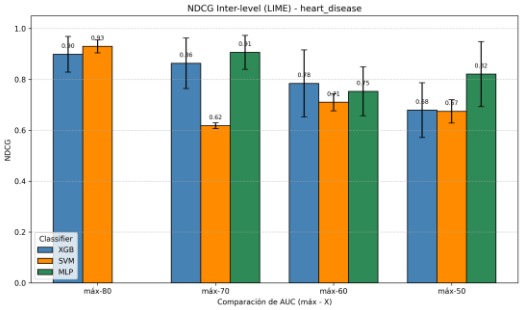


(b) LIME

Fig. A.2.2. NDCG medio entre las explicaciones del modelo de referencia (mayor AUROC alcanzado) y las de los modelos degradados en el dataset *dementia*, generadas con (a) SHAP y (b) LIME.



(a) SHAP



(b) LIME

Fig. A.2.3. NDCG medio entre las explicaciones del modelo de referencia (mayor AUROC alcanzado) y las de los modelos degradados en el dataset *heart-disease*, generadas con (a) SHAP y (b) LIME.

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Wed Jul 02 23:40:47 CEST 2025
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)