Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Grado

Metodología de validación de explicaciones generadas mediante técnicas de IAX post-hoc para procesos de aprendizaje automático

Autor: Alba María López González
Tutor: Dr. Esteban García-Cuesta

Madrid, Junio - 2024

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: Metodología de validación de explicaciones generadas mediante técnicas de IAX post-hoc para procesos de aprendizaje automático

Junio - 2024

Autor: Alba María López González
Tutor: Dr. Esteban García-Cuesta
Inteligencia Artificial
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid

Resumen

La Inteligencia Artificial eXplicable (IAX) es un campo de investigación emergente que aborda los desafíos de explicabilidad, transparencia y sesgos en los Sistemas de Inteligencia Artificial (SIA). Una de sus iniciativas es generar explicaciones post-hoc de modelos de aprendizaje automático que faciliten la comprensión de sus resultados, justifiquen sus decisiones, corrijan los SIA durante su desarrollo y descubran nuevo conocimiento. Sin embargo, al igual que los modelos de aprendizaje automático, la calidad de estas explicaciones puede variar y causar que no siempre sean adecuadas para una tarea de predicción. Además, no existe un consenso en la literatura sobre cómo evaluarlas y validarlas.

Este proyecto propone una metodología para validar las explicaciones generadas mediante técnicas IAX post-hoc, agnósticas del modelo y en forma de importancia de características, para procesos de aprendizaje automático. Esta metodología simplifica la tarea de evaluación y validación de la explicabilidad de los SIA y facilita la selección de las explicaciones más adecuadas para un problema de clasificación. Para ejemplificar su funcionamiento, la metodología se aplica a diferentes problemas de clasificación, demostrando que todas sus etapas son esenciales para obtener conclusiones fiables.

Palabras Clave: IAX, explicabilidad, interpretabilidad, metodología, validación, métodos de importancia de características

Abstract

Explainable Artificial Intelligence (XAI) is an emerging research field focused on addressing the challenges of explainability, transparency, and biases in Artificial Intelligence Systems (AIS). One of its key initiatives is to generate post-hoc explanations of machine learning models, , which can help users understand the results, justify decisions, correct the AIS during development, and uncover new knowledge. However, like the machine learning models themselves, the quality of these explanations can vary and sometimes make them unsuitable for a specific prediction task. Additionally, there is no consensus in the literature on how to evaluate and validate them.

This project proposes a methodology to validate the explanations generated using post-hoc XAI methods, that are model-agnostic and based on feature importance. The proposed methodology simplifies the evaluation and validation of AIS explainability, facilitating the selection of the most suitable explanations for a classification problem. To demonstrate its utility, the methodology is applied to various classification problems, illustrating that it consists of a series of essential steps to obtain reliable conclusions.

Keywords: XAI, explainability, interpretability, methodology, validation, feature importance methods

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

Es evidente que la Inteligencia Artificial (IA) trae consigo una serie de ventajas económicas y sociales en diversos sectores. La personalización de servicios de reparto, la optimización de operaciones en cadenas de suministro y la gestión eficiente de recursos en tiempo real son ejemplos claros del impacto positivo que los Sistemas de Inteligencia Artificial (SIA) pueden tener. Sus beneficios van más allá del ámbito empresarial, influyendo a nivel nacional e incluso global.

Sin embargo, los SIA, a pesar de su potencial para impulsar el desarrollo socioeconómico, también presentan riesgos que pueden tener graves consecuencias para individuos y la sociedad y se pueden manifestar a lo largo de todo su ciclo de vida: desde la recolección de datos, su procesamiento y modelado, hasta su uso para tomar decisiones y su mantenimiento continuo. Desde sus comienzos, los SIA han estado envueltos en controversias debido a decisiones sesgadas o discriminatorias en áreas críticas como la sanidad, justicia penal y movilidad. Por ejemplo, a mediados de la década de 1990, una red neuronal diseñada para predecir cuáles pacientes con neumonía debían ser ingresados en el hospital obtuvo resultados superiores a los métodos tradicionales. Sin embargo, este sistema aprendió patrones contraintuitivos, sugiriendo que los pacientes con neumonía y asma tenían un menor riesgo de muerte, lo que era incorrecto [1]. En 2013, el modelo COMPAS de predicción de reincidencia criminal fue utilizado para sentenciar al Sr. Loomis, resultando en una pena de seis años de prisión influenciada de manera discriminatoria por su género y raza [2]. En 2018, un Uber autónomo causó la muerte de una mujer en Arizona al no clasificar correctamente el objeto en su camino y no considerar necesario frenar [3]. Además, existen muchos otros casos similares de sistemas que desarrollan y propagan sesgos, se enfrentan a problemas de transparencia y explicabilidad, y sufren brechas de privacidad y seguridad, además de otros problemas éticos relacionados con los datos y la IA [4].

En este contexto, surge la Inteligencia Artificial eXplicable (IAX) como un campo de investigación emergente dedicado a abordar estos desafíos, sin comprometer el desempeño de los SIA. La IAX se basa en la premisa de que una mayor explicabilidad de los SIA puede conducir a una mejor comprensión y control, y reducir así los demás riesgos asociados. La mayor explicabilidad puede ayudar a justificar las decisiones de los algoritmos, asegurando que no sean erróneas especialmente cuando son inesperadas; ayudar a desvelar las vulnerabilidades y errores de un sistema, ofreciendo la posibilidad de corregirlos durante su desarrollo y prevenir que se produzcan en

situaciones críticas; y ser una herramienta útil para descubrir nuevo conocimiento, al proporcionar estrategias para abordar determinadas tareas que excedan las capacidades humanas [5, 6]. Con esta intención, la IAX aborda desde la generación de explicaciones de modelos de IA complejos hasta el desarrollo de nuevos modelos interpretables, para que en esencia los humanos puedan comprender sus decisiones y resultados.

Aun así, esta disciplina presenta sus propios desafíos. De la misma manera que las predicciones de los modelos de aprendizaje automático pueden variar en calidad para una tarea de predicción, las explicaciones o modelos explicables generados también pueden hacerlo. Sin embargo, mientras que existen indicadores de rendimiento bien definidos para evaluar la precisión de las predicciones de los modelos, otros criterios, como la explicabilidad, no son tan fácilmente cuantificables [7].

En este capítulo se contextualizará, motivará y establecerán los objetivos del proyecto. Para ello, en la Sección 1.1 se presentará la reciente propuesta de regulación de la IA a nivel europeo y expondrá su relevancia para el trabajo, en la Sección 1.2 se motivará la necesidad de validar y guiar la selección de explicaciones y justificará la selección del trabajo, y en la Sección 1.3 se presentarán los objetivos y subobjetivos del proyecto.

1.1. Contexto del proyecto

Recientemente, la Comisión Europea ha propuesto el primer marco legal sobre IA, con el objetivo de crear un entorno en el que la innovación tecnológica pueda prosperar sin comprometer los derechos fundamentales de las personas [4].

La propuesta establece reglas armonizadas para el desarrollo, la comercialización y el uso de sistemas de IA en la Unión, siguiendo un enfoque proporcional basado en el riesgo. Esta define los «sistemas de IA de alto riesgo» como aquellos que pueden poner en peligro la salud, seguridad o derechos fundamentales, y determina que estos sistemas deberán cumplir con un conjunto de requisitos y procedimientos de evaluación de conformidad antes de que puedan ser comercializados en el mercado. Algunos ejemplos de estos sistemas son los sistemas de identificación biométrica, gestión de infraestructuras críticas o reclutamiento de personas. Sin embargo, para otros sistemas de IA definidos como «sistemas de IA de riesgo militado», cuyo riesgo se reduce al asociado con la falta de transparencia, solo se proponen obligaciones mínimas de transparencia para asegurar que los seres humanos estén informados cuando sea necesario. En esta categoría entrarían, por ejemplo, aquellos sistemas que utilizan chatbots o contenido de audio y video que constituye falsificaciones profundas («deep fakes»).

Así, desde la Comisión Europea se ofrece visibilidad sobre los desafíos de la IA y regula las obligaciones de transparencia, explicabilidad, privacidad y seguridad, entre otros, de los SIA en base al riesgo que supongan para los individuos y la sociedad.

1.2. Motivación del proyecto

En este contexto, la IAX es una herramienta con potencial para abordar los desafíos de los SIA presentes desde sus comienzos y consolidados por la Comisión Europea re-

cientemente. Entre los métodos más utilizados para mejorar la explicabilidad están la atribución de importancia de características de modelos existentes, perturbación de entradas o visualización, que se aplican sobre los modelos "caja negra" sin necesidad de conocer cómo han aprendido o funcionan.

Como se ha comentado, este tipo de técnicas tienen sus propias limitaciones. Desde la literatura, se cuantifican dichas limitaciones mediante una multitud de métricas que definen cualidades deseables que las técnicas deben cumplir, como adecuación, fidelidad, robustez o usabilidad. Dependiendo del dominio de aplicación, tipo de explicación, tipo de datos o conocimientos previos del usuario para el que se generan las explicaciones, es posible o relevante evaluar unas u otras cualidades de las explicaciones. Además, no existe un consenso acerca de las cualidades deseables para cada tipo de explicación, y se utiliza terminología diferente para referirse a conceptos similares. Todo ello puede generar confusión a los desarrolladores que pretendan cumplir con los requisitos de explicabilidad sobre los SIA o usuarios que pretendan iniciarse en el campo.

1.3. Objetivos

En este trabajo tiene como objetivo principal el desarrollo una metodología de validación de explicaciones generadas mediante técnicas de IAX para procesos de aprendizaje automático, simplificando la tarea evaluación y validación de la explicabilidad en este tipo de sistemas. Además, esta metodología pretende asistir en la selección de las mejores explicaciones posibles para una tarea de clasificación y ante multiplicidad predictiva. Dado que sería imposible abordar el total de tipos de explicaciones generadas, este trabajo se centra en la evaluación y validación de las explicaciones post-hoc, agnósticas del modelo, en forma de importancia de características. Con todo ello, esencialmente, se busca fomentar la transparencia y la confianza en los SIA, de manera que puedan ser implementados de manera segura en ámbitos críticos y se puedan aprovechar así sus potenciales beneficios.

Para ello, se plantean los siguientes subobjetivos:

- Subobjetivo 1: Definir las métricas y diseñar pruebas de validación.
    - Establecer las métricas pertinentes para evaluar la calidad de las explicaciones para métodos post-hoc, agnósticos del modelo, qie generen explicaciones en forma de importancia de características.
    - Seleccionar o diseñar pruebas adecuadas para evaluar las métricas definidas, considerando la diversidad de bases de datos, modelos y técnicas de explicabilidad.
- Subobjetivo 2: Desarrollar una metodología de explicabilidad.
    - Integrar las pruebas definidas en un marco metodológico coherente y replicable, que sirva tanto para garantizar la generación de explicaciones adecuadas como para guiar la selección de las mejores explicaciones para una tarea de clasificación.
    - Proporcionar una descripción clara del uso de esta metodología para alcanzar ambos propósitos.

- Subobjetivo 3: Ejemplificar la aplicación de la metodología.
  - Seleccionar una variedad de bases de datos, modelos y técnicas de explicabilidad para generar explicaciones y poner en práctica la metodología.
  - Demostrar el uso de la metodología en diferentes escenarios y extraer conclusiones de sus ventajas y limitaciones.

1.4. Estructura del Documento

La estructura del documento se divide en tres secciones fundamentales. En primer lugar, la Sección 2 presenta el marco teórico que proporciona los fundamentos y contexto necesarios el desarrollo del proyecto. A continuación, en la Sección 3 se expone la metodología de evaluación propuesta, y detallan las bases de datos, modelos y técnicas de explicabilidad sobre las que será ejemplificada. Además, en la Sección 4 se discuten los resultados obtenidos en cada fase de la metodología sobre los diferentes conjuntos de datos, y finalmente, en la Sección 5 se extraen conclusiones generales, evalúa el impacto del proyecto y sugieren futuras líneas de investigación.

Capítulo 2

Conceptos previos y Estado del Arte

2.1. Conceptos previos en IAX

2.1.1. Confianza en IAX

La confianza es considerada un constructo social de gran trascendencia por psicólogos de diversos campos. Una definición comúnmente aceptada señala que es “el estado psicológico en el que se acepta una postura de vulnerabilidad respecto a otro, en vista a las expectativas positivas de sus intenciones o comportamiento”[8]. Sin la capacidad de confiar, no podríamos aceptar las incertidumbres cotidianas y formar relaciones mediante las que funcionar en el mundo social [9].

En el contexto de la IA, la confianza se refiere a la expectativa de si “un modelo actuará según lo previsto al enfrentar un problema dado”[10], y repercute en su aceptación y utilización para la toma de decisiones críticas. A pesar del potencial de los SIA, como se ha comentado, los usuarios a menudo son escépticos ante sus predicciones debido a la falta de confianza en que el sistema realmente entienda sus circunstancias personales y actúe en su interés. Este fenómeno se conoce como ”aversión al algoritmo” (que en inglés se conoce como «algorithm aversion») y puede constituir un obstáculo en el progreso de muchas empresas que desarrollan este tipo de sistemas [11]. Así, surgen preguntas como: ¿Cómo podemos confiar en un modelo que sabemos que comete errores esporádicos, si no proporciona ninguna justificación aceptable para sus decisiones? ¿Cómo podemos delegar un trabajo en algo que razona de una manera que no podemos seguir ni entender? [12] Como dijo la investigadora y emprendedora Nurit Nobel: “Estos algoritmos están diseñados para facilitar la elección por parte de los humanos. Teniendo esto en cuenta, es sorprendente descubrir cuán poca investigación sobre el comportamiento humano se suele incorporar en su diseño”[13].

La ciencia del comportamiento busca conectar la IA con los humanos, estudiando los diversos factores que contribuyen a la confianza de un usuario en un SIA, como su robustez, transparencia y explicabilidad. Estos factores están intrínsecamente relacionados; al fomentar cualquiera de ellos, no solo se mejora la confianza en el sistema, sino que también se fortalecen los demás [14]. Basándose en ellos, la confianza en un sistema de IA se puede definir tanto de manera subjetiva como objetiva [15].

Aprovechando los conocimientos de la ciencia del comportamiento, se podría reforzar la confianza de los usuarios en los SIA y conseguir que sean aún más eficientes en la toma de decisiones responsables. Una manera de abordar este desafío es explicando el comportamiento de los modelos, es decir, fomentando la explicabilidad. Para ello, y sin comprometer el desempeño de los SIA, surge la Inteligencia Artificial eXplicable (IAX). Así, la confianza se puede lograr cuando el modelo puede proporcionar explicaciones de sus decisiones, ya que resulta lógico pensar que una persona puede tener más confianza en usar un modelo si lo entiende [15].

2.1.2. Explicabilidad e Interpretabilidad

A menudo existe confusión entre los términos de explicabilidad e interpretabilidad. Ambos se usan de manera intercambiable en la literatura, junto con otros como comprensibilidad (que en inglés se diferencia entre «understandability» y «comprehensibility»), pero en realidad, hacen referencia a cualidades distintas de los modelos [5].

La explicabilidad se refiere a la capacidad de describir y expresar el comportamiento del modelo y sus decisiones finales en términos entendibles por humanos no expertos [16]. De esta manera, sus usuarios finales y demás usuarios interesados serían capaces de entender la relación matemática entre las entradas del modelo y sus salidas y justificar *por qué* el modelo ha tomado una decisión u otra. Por otra parte, la interpretabilidad se refiere a la capacidad comprender *cómo* un sistema de IA toma las decisiones que toma.

Aún así, ambos términos están relacionados, ya que un modelo interpretable es explicable si sus operaciones pueden ser entendidas por humanos [14]. Estos también guardan relación con los conceptos de transparencia, robustez y sesgos, ya que la explicabilidad fomenta la transparencia del proceso de toma de decisiones del modelo y facilita la evaluación de robustez y mitigación de sesgos.

2.1.3. El problema de "caja negra"

Normalmente, se utilizan modelos de aprendizaje automático (ML) como las redes de neuronas profundas (DNN) para buscar relaciones complejas en datos con alta precisión [17]. Aun así, es tal la cantidad de factores que influencian la relación que una DNN extrae de los datos que resulta complicado entender e interpretar su funcionamiento. Incluso la DNN más sencilla puede estar compuesta por varias capas, filtros y neuronas, que aumentan exponencialmente con el número de parámetros a entrenar y dificultan la capacidad de examinar sus conexiones [18]. Por esta razón, estos modelos se conocen como modelos de "caja negra" (del inglés, «black-box»): un término que alude a la incapacidad de entender las relaciones que extraen. Esto genera un problema de interpretabilidad.

Típicamente, son los modelos más complejos los que obtienen mejores resultados en términos de accuracy (u otras métricas de evaluación) que los modelos más sencillos. Por ello, desarrollar el mejor modelo normalmente implica aumentar su complejidad, lo que tiende a reducir su interpretabilidad. En algunos casos, resulta asumible perder cierta interpretabilidad para conseguir el mejor modelo posible. Sin embargo, cuando esto no es viable, la IAX propone generar explicaciones de los modelos complejos a posteriori sin necesidad de alterarlos, pero logrando la explicabilidad deseada.

2.1.4. Explicabilidad ante-hoc vs. post-hoc

Existen también modelos de aprendizaje automático más simples, como la regresión lineal, regresión logística o los árboles de decisión que encuentran relaciones más sencillas en los datos. Esto se debe a que parten de una hipótesis de linealidad o sub-linealidad, que limita su complejidad [19]. De esta manera, modelos son interpretables (e incluso explicables) de forma ante-hoc. Sin embargo, dado que la realidad suele ser altamente compleja y no lineal, estos modelos suelen tener un rendimiento inferior en términos de precisión y no resultan una solución prometedora para problemas complejos a pesar de las necesidades de explicabilidad presentes.

Por ello, los esfuerzos de la IAX se centran en dos desafíos: generar explicaciones de modelos de IA complejos (explicabilidad post-hoc) o desarrollar nuevos modelos explicables (explicabilidad ante-hoc). Algunos autores argumentan que es preferible desarrollar y entrenar modelos inherentemente interpretables en vez de explicar modelos "caja negra" a posteriori, ya que en algunos casos las explicaciones pueden ofrecer una falsa sensación de confianza [hater]. Aunque esto pueda ser cierto, debido a la existencia de una multitud de modelos entrenados para diversos problemas complejos, que han requerido una gran capacidad de cómputo, en muchos contextos no resulta posible o interesante someterlos a reentrenamiento. En este contexto, los métodos de explicabilidad post-hoc proporcionan interpretabilidad sobre un modelo ya aprendido sin necesidad de conocer cómo ha aprendido o funciona, lo que permite trabajar sobre modelos existentes.

Además, existe un amplio abanico de técnicas de explicabilidad post-hoc, que generan explicaciones para diferentes contextos, mediante diferentes mecanismos y de diferentes formas. Así, es posible encontrar la técnica de más adecuada para dotar de explicabilidad a un modelo y problema de predicción específico.

2.2. Estado del Arte en IAX

2.2.1. Taxonomía de técnicas de explicabilidad post-hoc

En la literatura no existe consenso acerca de cómo categorizar las diferentes técnicas de explicabilidad post-hoc. Algunas revisiones proponen categorías excluyentes entre las que dividir el conjunto de técnicas, resultando en una clasificación simplista [20, 21]; mientras que otras aportan demasiados detalles sobre las diferencias entre unas técnicas y otras, lo que dificulta entender su relación [16].

Por lo tanto, en este caso se propondrá una taxonomía múltiple, dividida en los 4 principales aspectos en función de los que se puede clasificar. Estos son su **alcance** (qué datos explica), **aplicabilidad** (qué modelos explica), **funcionamiento** (cómo genera la explicación), **resultado** (qué tipo de explicación genera). Para cada uno de ellos, se plantean un conjunto finito de categorías en las que clasificar todas las técnicas. Así, cada técnica se puede clasificar en una categoría diferente para cada uno de los elementos, facilitando encontrar similitudes y diferencias entre las mismas, y seleccionar la más apropiada para cada contexto. La taxonomía propuesta está inspirada en la taxonomía propuesta por Speith [22] y se observa en la Figura 2.1.

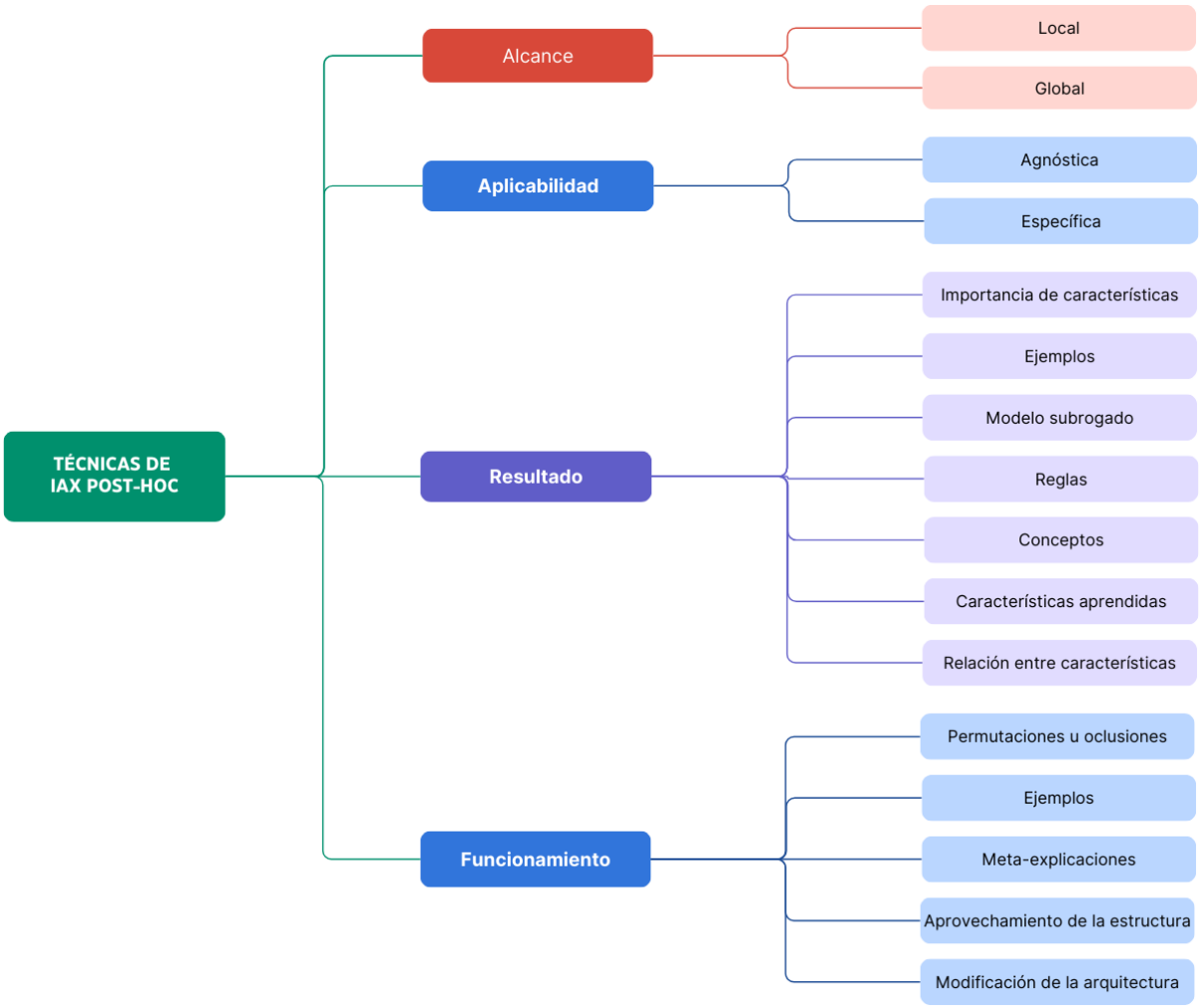


Figura 2.1: Taxonomía de técnicas post-hoc de IAX

2.2.1.1. Alcance

El alcance de una técnica se refiere a lo generales que son sus explicaciones. Las técnicas que explican el modelo y los datos en su conjunto se conocen como técnicas **globales**. Aun así, no todas las técnicas generan explicaciones que se puedan interpretar en el contexto global del modelo o conjunto de datos. Estas técnicas se conocen como técnicas **locales**, y explican instancias concretas o subconjuntos de instancias de manera más precisa, pero a veces inconsistente en el contexto global.

2.2.1.2. Aplicabilidad

Por otra parte, la aplicabilidad se refiere al conjunto de modelos sobre los que se puede implementar. Aquellas técnicas que buscan explicar el comportamiento de los modelos de aprendizaje automático de una manera general, sin depender de ningún tipo de modelo o dominio de aplicación, son técnicas **agnósticas**, mientras que aquellas diseñadas para un tipo de modelo o dominio concreto, son técnicas **específicas**. Estas últimas, a pesar de no proporcionar soluciones generalizables, funcionan de manera óptima para el caso para el que fueron diseñadas.

2.2.1.3. Resultado

Además, el resultado de la técnica de explicabilidad es la forma final que cobra la explicación extraída del modelo y datos. Las explicaciones más comunes obtienen el valor de **importancia de las características**, que reflejan el grado en el que una característica específica contribuye a los resultados obtenidos por el modelo, y se generan mediante técnicas conocidas como métodos de importancia de características [23]. En el caso en el que los datos sobre los que se apliquen sean imágenes, estas técnicas mostrarán la importancia de los diferentes píxeles en el resultado de la clasificación del modelo CNN, mediante "mapas de calor" (del inglés, «heatmaps») para facilitar la interpretabilidad por parte de los usuarios [16].

Otros modos de explicaciones incluyen ejemplos, modelos subrogados, reglas, y menos comúnmente se utilizan conceptos, características aprendidas o relación entre características. Los métodos basados en **ejemplos** seleccionan instancias concretas del conjunto de datos para explicar el comportamiento de los modelos de aprendizaje automático o distribución de datos subyacente. Los ejemplos pueden estar presentes o no en los datos de entrada, y deben cumplir con una determinada condición que sea de interés para entender el modelo (como ser representativa del conjunto de datos o de una parte del mismo) [24]. Aun así, solo tiene sentido generarlos si las instancias en cuestión son comprensibles por humanos, ya que tienen su misma forma. Por otra parte, un **modelo subrogado** de un modelo complejo es un modelo más sencillo que imita sus predicciones. En el contexto de la IAX, el modelo subrogado será un modelo intrínsecamente explicable, como una regresión lineal o árbol de decisión, que aproxima las predicciones de un modelo "caja negra" [22]. Además, las **reglas** son un conjunto de condiciones que deben cumplir los valores de las diferentes variables para obtener un determinado resultado. El modelo más sencillo que explicar mediante reglas es un árbol de decisión, cuya estructura permite su extracción directa [24]. Otra posibilidad de explicar modelos es mediante **conceptos**. Un concepto es una representación atómica de algún elemento del problema que se quiere resolver y que permite explicar el modelo a través de él. Esta representación se aprende durante el propio proceso de aprendizaje [25]. De manera similar, los métodos basados en **características aprendidas** son aquellos que muestren las características aprendidas de manera explícita. Se suelen aplicar sobre redes de neuronas, y para que mejoren su explicabilidad se debe localizar aquellas unidades de la red que se generen características interpretables, entre todas las posibles [26]. Un último tipo de resultados incluyen aquellas que desvelan la **relación entre pares de variables**, o las diferentes variables y el resultado del modelo, generalmente mediante visualizaciones o métricas de relación entre características.

2.2.1.4. Funcionamiento

Finalmente, el funcionamiento de una técnica hace referencia al mecanismo interno que la propia técnica usa para generar la explicación. En cuanto a su funcionamiento, las técnicas se pueden basar en perturbaciones u oclusiones, ejemplos, meta-explicaciones y aprovechamiento o modificación de la arquitectura.

Las técnicas basadas en **permutaciones u oclusiones** proponen modificar los valores de las instancias presentes, generar e incluir instancias nuevas, o eliminar instancias anteriores para generar las explicaciones deseadas. Algunas técnicas como LIME generan perturbaciones aleatorias uniformemente distribuidas de una instancia

(a explicar) en su espacio de variables, evalúan las nuevas instancias con el modelo original, y con ello componen una nueva base de datos con la que entrenan un modelo subrogado más sencillo. Otras técnicas, sin embargo, perturban mínimamente las instancias y las evalúan en el modelo original con el objetivo de determinar aquellas que generen un resultado significativamente diferente. Estas serían contraejemplos del modelo. Además, técnicas como LOCO [27] o MCR [28] proponen calcular la importancia de características de las variables individuales midiendo el incremento en el error de predicción al permutar sus valores.

Por otra parte, las técnicas que generan explicaciones en forma de **ejemplos** emplean algoritmos de búsqueda en el espacio de variables de los datos. Dependiendo de las restricciones de este espacio, el coste computacional de este tipo de técnicas es mayor o menor, por lo que suelen emplear heurísticas. Además, el tipo de técnicas que extraen una explicación de otra explicación previamente generada funcionan mediante **meta-explicaciones**. El caso más común son aquellas técnicas que generan modelos subrogados de los que posteriormente extraen otra explicación de diferente tipo. Por ejemplo, técnicas como LIME utilizan un modelo subrogado para determinar la importancia de las características en el modelo al evaluar una instancia concreta, mientras que otras como LORE utilizan un modelo subrogado para generar reglas que señalan posibles cambios que puede sufrir una instancia sin que cambie el resultado de su predicción. Para modelos basados en NNs, también se pueden emplear técnicas que aprovechen las particularidades de su estructura o la modifiquen para generar explicaciones. Las más comunes utilizan los gradientes.

2.2.2. Métodos de importancia de características

La clasificación de los principales métodos de importancia de características post-hoc en función de la taxonomía presentada se observa en la Tabla 2.1. En su mayoría, se tratan de métodos locales. Además, se diferencia claramente entre aquellos agnósticos al modelo (y utilizados en el desarrollo de estre trabajo) y específicos a CNNs o NNs.

Cuadro 2.1: Clasificación de los principales métodos de importancia de características post-hoc.

Técnica	**Aplicabilidad**	**Alcance**	**Funcionamiento**	**Referencias**
LIME, Shapley Values, SHAP	Agnóstico	Local	Perturbaciones, Meta-Explicación	[29], [30], [31]
MCR	Agnóstico	Global	Perturbaciones	[28]
LOCO	Agnóstico	Local	Perturbaciones	[27]
VanillaGradient, Grad-CAM	Específico (CNNs)	Local	Aprovechamiento de la estructura	[32], [33]
SmoothGrad	Específico (CNNs)	Local	Perturbaciones, aprovechamiento de la estructura	[34]
DTD, LRP, IG, DeepLIFT	Específico (NNs)	Local	Aprovechamiento de la estructura, Perturbaciones	[35], [36], [37], [38]

Algunos de ellos, primero generan modelo subrogado local para aproximar las predicciones de un modelo "caja negra". A lo largo de esta sección, sea f el modelo de predicción original, g el modelo subrogado local, $f(x)$ la predicción de una entrada x que toma valores en el espacio de variables original, y $g(z)$ la estimación de $f(z)$ tal que $g(z) \approx f(z)$ siempre que $z \approx x$.Además, si el método de importancia de características genera un modelo de explicación g que toma valores en un espacio de características simplificado (binario), cumple con la Ecuación 2.1. Muchos métodos como LIME [29] o SHAP [31] se ajustan a esta definición.

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2.1}$$

Lundberg y Lee denominan a este tipo de métodos «métodos de importancia de características aditivos». Ahora, existe una función de mapeo h_x que revierte las entradas simplificadas x' al espacio de variables original mediante el mapeo $x = h_x(x')$. Dependiendo del espacio de variables de entrada, se utiliza un mapeo u otro. Por ejemplo, dados vectores de 0s y 1s en un espacio de características textuales o palabras (donde 0 significa que la palabra está ausente en un ejemplo y 1 que está presente), h_x mapea los 1s en el número de ocurrencias de la palabra en el ejemplo y mantiene los 0s. En cualquier caso, $g(z') \approx f(h_x(z'))$ siempre que $z' \approx x'$, y se puede expresar en función de los efectos ϕ_i de cada variable. Los efectos ϕ_i son los valores de importancia atribuidos a cada variable, de tal manera que la suma de todos ellos aproxima la predicción $f(x)$ del modelo original.

A continuación, se desarrollarán en mayor profundidad los métodos de importancia de características aditivos LIME, Shapley Values y SHAP, y comentarán brevemente los demás presentes en la Tabla 2.1.

2.2.2.1. LIME

LIME [29] utiliza una aproximación local del modelo original para explicar las predicciones individuales. Esta técnica genera un modelo lineal explicable que cumple con la Ecuación 2.1.

Para ello, primero genera una nueva base de datos con n instancias x' que toman valores en un espacio de variables simplificado (reversibles mediante el mapeo $x = h_x(x')$). Estas se generan como perturbaciones aleatorias uniformemente distribuidas de la instancia original x en el espacio de variables simplificado. A continuación, las nuevas instancias x' se ponderan en función a su proximidad con la instancia x (mediante, por ejemplo, un kernel gaussiano) y se utilizan para entrenar el modelo subrogado interpretable.

Para encontrar los valores ϕ_i, LIME minimiza la función objetivo de la Ecuación 2.2. Puesto que g satisface la Ecuación 2.1 y L es un error cuadrático, esta ecuación se puede resolver utilizando regresión lineal penalizada. Así, los valores de importancia de las características se aproximan mediante los valores ϕ_i.

$$\xi = \arg\min_{g \in \mathcal{G}} \; L(f, g, \pi_{x'}) + \Omega(g) \tag{2.2}$$

LIME es un método agnóstico del modelo, ya que en cualquier caso lo aproxima mediante un modelo subrogado para interpretar las predicciones, y se utiliza para explicar instancias individuales (es local). De esta manera, mientras es capaz de proporcionar explicaciones más precisas para cada instancia, es sensible al kernel y perturbaciones generadas, lo que supone que puede dar lugar a explicaciones muy diferentes o inconsistentes para instancias similares [29].

2.2.2.2. Shapley Values

Los Shapley Values se basa en la teoría económica del juego para estimar la aportación de los diferentes miembros de una coalición al resultado de la misma [30]. Esta teoría parte del reconocimiento de que en un juego cooperativo, los jugadores contribuyen de manera diferente en al resultado, y aboga por la redistribución justa de los beneficios en función de dicha contribución. De esta manera, los Shapley Values calculan la aportación de cada jugador al resultado total considerando todas las posibles subcoaliciones de jugadores y sus resultados individuales. En el ámbito del aprendizaje automático, los Shapley Values se utilizan para estimar la contribución de las diferentes variables en las predicciones del modelo.

Matemáticamente, dados N jugadores y una función de valor v, el valor Shapley ϕ_i del jugador i es una media ponderada de las contribuciones marginales de i en todos los subconjuntos S de N que contienen a i. Esta media se divide entre el número de subconjuntos de N (es decir, $|N|!$). Además, para cada subconjunto S que contenga a i, la contribución marginal se calcula mediante una diferencia entre el valor de v con i incluido y el valor de v sin i, ponderada por el número de formas en que se pueden formar estos subconjuntos [39]. Todo esto se refleja en la Ecuación 2.3.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(v(S \cup \{i\}) - v(S)) \tag{2.3}$$

Estos valores satisfacen las propiedades de eficiencia, simetría, «dummy player» y linealidad, mediante las cuales se asegura la presencia de una solución única al problema de atribución [31].

2.2.2.3. SHAP

Lundberg y Lee proponen los valores SHAP (SHapley Additive exPlanation) [31] como medida unificada de los valores de importancia de características. Estos calculan el valor de importancia de una característica como el cambio en la predicción esperada del modelo al condicionar con dicha característica. Así, explican cómo llegar desde el valor base $E[f(z)]$, que se predeciría si no conociéramos ninguna característica, hasta la salida actual $f(x)$. En su versión más sencilla, se calculan los efectos ϕ_i generados en un solo orden. Sin embargo, cuando el modelo es no lineal o las características de entrada no son independientes, el orden en el que se añaden las características al valor de la esperanza base para calcular los efectos ϕ_i importa, y los valores SHAP surgen del promedio de los efectos ϕ_i en todos los órdenes posibles.

Su cálculo exacto supone un reto matemático y computacional, pero existen numerosas aproximaciones inspiradas en métodos de importancia de características existentes. Una de ellas es KernelSHAP, que utiliza algunas ideas de LIME para calcular

valores de importancia de características que satisfagan las propiedades deseables de los Shapley Values. Para ello, al igual que LIME aproximan un modelo subrogado local alrededor de una instancia para estimar los valores de importancia, pero en este caso evitan el uso de heurísticas para poder recuperar los Shapley Values a partir de la Ecuación 2.2.

KernelSHAP es un método agnóstico, pero existen otras aproximaciones como TreeSHAP o DeepSHAP que son específicas para explicar las predicciones de modelos basados en árboles y redes profundas, respectivamente. Todos estos métodos generan explicaciones locales.

2.2.2.4. Otros

MCR y LOCO
MCR (Model Class Reliance) [28] evalúa la importancia de variables al medir el aumento en el error del modelo al permutar valores de las variables y recalculando el modelo. Es un método global que proporciona una visión general del modelo sin necesidad de reentrenamiento, pero puede no ser ideal con variables correlacionadas.

Por otra parte, LOCO (Leave-One-Covariate-Out) [27] estima la importancia de variables al dejar una variable fuera del modelo y comparar el rendimiento con el modelo completo. Este enfoque es local y requiere reentrenamiento, pero ofrece una evaluación más precisa de la importancia individual de las características.

Vanilla Gradient y SmoothGrad
Otras técnicas como Vanilla Gradient (o Saliency Maps) [32] y SmoothGrad [34] utilizan los gradientes de una red neuronal convolucional (CNN) para identificar los píxeles cuyo cambio influiría en la predicción del modelo. Vanilla Gradient utiliza los gradientes de la capa anterior a la clasificación (Softmax), retropropagándolos hasta la capa de entrada para aproximar la importancia de cada píxel de entrada respecto a la salida. Sin embargo, los mapas de calor obtenidos pueden ser ruidosos.

SmoothGrad mejora este enfoque mediante la generación de múltiples perturbaciones de la imagen original con ruido gaussiano. Luego, se retropropagan los gradientes de estas imágenes perturbadas y se promedian. A pesar de esto, ambas técnicas producen resultados menos robustos y precisos debido a que no cumplen ciertos criterios deseables y son sensibles a la saturación del gradiente.

DTD, LRP y DeepLIFT
Por otra parte, existen métodos basados en gradientes como DTD, LRP y DeepLIFT que generan valores de importancia para las características de manera que su suma en cada punto de la red coincida con la diferencia entre el resultado de una entrada específica y una de referencia. Estas explicaciones son consistentes con las características de entrada y robustas frente al ruido.

DTD (Deep Taylor Decomposition) [35] descompone el resultado de una red neuronal en valores de importancia, conocidos como relevance values", mediante una expansión de Taylor. La suma de estos valores coincide con la predicción de la red.

LRP (Layer-wise Relevance Propagation) [36] descompone la función de predicción en la suma de valores de importancia capa por capa, de manera más simple y específica para redes neuronales.

DeepLIFT [38] estima la importancia de las características de una entrada específica directamente como la diferencia entre los gradientes de esa entrada y los de una entrada de referencia. En el caso de imágenes, la entrada de referencia puede ser una imagen en blanco o una versión borrosa de la imagen original.

IG
Finalmente, IG (Integrated Gradients) [37] aproxima la importancia de las características mediante el cálculo de la integral de los gradientes a lo largo de un camino entre la clase de interés y la capa de entrada. Al igual que DTD, LRP, DeepLIFT e IR, conservan la relevancia total de las entradas, y puesto a que está únicamente basado en gradientes, es invariante a la implementación.

2.2.3. Aplicaciones

Medicina y biomedicina
En el ámbito de la investigación médica y biomédica, los SIA están suponiendo un creciente apoyo en el estudio de estructuras celulares, diagnosis de patologías y descubrimiento de nuevos fármacos. La mayoría de los sistemas utilizados están especializados en análisis de imágenes, y tratan con imágenes médicas y biomédicas a partir de las cuales se generan resultados. A pesar de ser prometedores, estos modelos no proporcionan suficiente evidencia científica con sus predicciones punto-a-punto para ofrecer un diagnóstico médico fiable. Así, en numerosas aplicaciones como en la clasificación de heridas crónicas mediante redes de neuronas convolucionales (CNNs) [40] o en la diagnosis de cáncer mediante otro tipo de DNNs [41], se han utilizado métodos como LIME para generar explicaciones visuales (en forma de mapas de calor) que identifiquen las regiones relevantes implicadas en la detección. Por otra parte, las explicaciones pueden ser una herramienta prometedora para el descubrimiento de conocimiento en este ámbito. Al explicar sistemas que exceden las capacidades humanas en determinadas tareas, se pueden aprender nuevas estrategias para abordarlas, y obtener una ventaja competitiva [6].

Salud
En el ámbito de la salud, los requisitos de intepretabilidad son aún mayores, ya que la toma de decisiones es aún más crítica. Por ejemplo, la predicción de la esperanza de vida de un paciente para la gestión de recursos de un hospital requiere de un entendimiento mayor del SIA que la simple predicción del coste de un procedimiento médico. En este ámbito, SHAP ha sido ampliamente utilizado para proveer explicaciones de modelos de admisión en hospitales [42], estimación de la calidad de vida [43], complicación de cirugías [44], oncología [45] y análisis de factores de riesgo implicados en la mortalidad en hospitales [46].

Ciberseguridad
La ciberseguridad consiste en el uso de procedimientos y tecnologías para la defensa de datos, aplicaciones, redes y sistemas ante potenciales ataques cibernéticos [47]. Con este propósito, se han desarrollado e implementado una gran variedad de SIA en este ámbito, aunque no exentos de las vulnerabilidades propias de estos sistemas. En este contexto, IAX se convierte en una herramienta imprescindible para desvelar las vulnerabilidades y errores de los SIA, ofreciendo la posibilidad de corregirlos durante su desarrollo y prevenir que se produzcan en situaciones críticas [6]. Métodos de importancia de características como LIME han sido utilizados para determinar las características más relevantes de los ataques de huella digital a sitios web [48]

o determinar los términos más utilizados en las peticiones que atentan contra la privacidad de determinadas aplicaciones [49]. Además, en este ámbito se ha desarrollado un método de IAX similar a LIME llamado LEMNA (Local Explanation Method using Nonlinear Approximation), optimizado para aplicaciones de seguridad basadas en el aprendizaje profundo como el reconocimiento de malware en archivos PDF y la ingeniería inversa de binarios [50].

Finanzas y legislación
En finanzas, los SIA se utilizan fundamentalmente para la predicción de valores financieros, como el precio de acciones o los beneficios de una compañía [51], y la evaluación de riesgos de activos [52]. En estos casos, LIME se puede emplear para simplificar modelos de predicción complejos y generar explicaciones locales de las predicciones. Esta técnica ha sido utilizada por Agarwal et al. [53] para explicar las predicciones de los valores de mercado generadas por un modelo AdaBoost. Con el mismo propósito, los autores también utilizaron SHAP, obteniendo valores de importancia de cada variable en la predicción generada.

En el dominio legal, uno de los principales desafíos de la IAX es la detección de sesgos en las decisiones judiciales para garantizar que se tomen de manera justa con respecto a determinados individuos y grupos [54]. Las aplicaciones desarrolladas en este ámbito suelen ser de alto riesgo, ya que las decisiones que generan tienen consecuencias significativas para los individuos y pueden atentar contra sus derechos fundamentales. Estas aplicaciones a menudo se basan en información textual en formato natural, con la que se entrenan CNNs para realizar predicciones y utilizan métodos de IAX como LIME o SHAP para determinar la contribución de cada palabra del texto en la predicción final [55].

Educación
Por otra parte, la aplicación de los SIA en la educación han demostrado diferentes beneficios como el apoyo a la instrucción, los sistemas de aprendizaje personalizados y los sistemas de evaluación automatizados [56]. En este ámbito, el sesgo de estos sistemas es un riesgo prominente. Así, se pueden emplear técnicas como SHAP para explicar las predicciones del abandono de los estudiantes [57] u explicaciones globales para fomentar el aprendizaje de estudiantes de enfermería [58].

2.3. Desafíos y validación en IAX

2.3.1. Desafíos en IAX

Mientras que existen indicadores de rendimiento para evaluar la precisión de las predicciones de los modelos de aprendizaje automático, otros criterios como la explicabilidad pueden no ser tan fácilmente cuantificables. Es conocido que diferentes técnicas pueden proporcionar diferentes explicaciones para un mismo problema dado, incluso utilizando el mismo conjunto de datos y el mismo modelo (e hiperparámetros) [7]. En concreto, diferentes métodos post-hoc de importancia de características pueden identificar diferentes características de entrada como importantes en las predicciones de la red, lo que posteriormente puede derivar en diferentes conclusiones científicas [31]. En este contexto, características como la adecuación, robustez, la transferibilidad de las explicaciones generadas a datos no observados y la usabilidad, son elementos clave que hay que tener en cuenta para poder verificar los resultados obtenidos.

2.3.2. Métricas de validación

En términos generales, podemos dividir los tipos de validación y las métricas empleadas para lograrlas en dos grandes bloques. El primer bloque incluye el conocimiento del «ground truth» o verdad absoluta que se espera obtener, lo que permite validar las explicaciones de manera supervisada. Este define la adecuación como la consonancia entre las explicaciones obtenidas y las esperadas. El segundo bloque se aplica cuando no se dispone del ground truth, por lo que la validación de las explicaciones se basa en determinar y valorar un conjunto de condiciones que deben cumplirse para que las explicaciones sean válidas para una determinada tarea de predicción. Algunos ejemplos de condiciones deseadas son la fidelidad, la robustez, la transferabilidad, y la usabilidad. Esta clasificación se observa en la Figura 2.2.

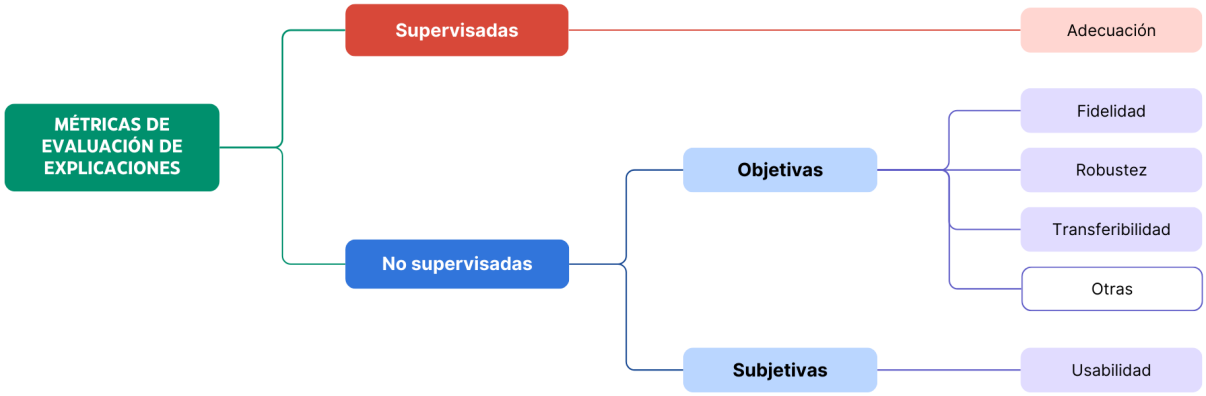


Figura 2.2: Clasificación de métricas de evaluación de explicaciones

2.3.2.1. Métricas de adecuación y «ground truth»

La **adecuación** de una explicación sólo puede medirse cuando se dispone del denominado «ground truth» o verdad absoluta que se espera obtener (es decir una etiqueta/s con la explicación esperada dada una instancia o un modelo). Este caso asume que esas etiquetas se conocen a priori (porque existe en la literatura o se ha modelado ese conocimiento previamente) o existe la posibilidad de obtenerlas bajo demanda (a partir de expertos o consenso de personas no expertas). De esta manera, las explicaciones se evalúan de manera supervisada.

Dependiendo del tipo de explicación generada, el ground truth requerido y la manera de evaluarlo será diferente. En imágenes, por ejemplo, se puede utilizar aprendizaje semi-supervisado para determinar si las explicaciones en forma de saliency maps son relevantes para localizar un objeto en una imagen. Para ello, proponen métricas como SSR [59] o Average drop [5]. Por otra parte, para explicaciones en forma de importancia de características se puede verificar hasta qué punto la explicación incluye el ground truth, a nivel de instancia o de modelo, mediante métricas como el coeficiente de Jaccard (que mide si la característica del ground truth está o no incluida en la explicación) o el coeficiente de Spearman y NDCG [60] (que también consideran el orden de importancia de las características, si se dispone de él).

Aun así, la obtención del ground truth experto puede ser una tarea difícil, ya sea por la escasez de literatura disponible o la limitada capacidad para generar nuevo conocimiento experto bajo demanda. Además, puesto a que el ground truth solo refleja

el conocimiento actual del mundo, este tipo de evaluación inevitablemente penaliza el descubrimiento de nuevo conocimiento a partir de las explicaciones. Por otra parte, obtener conocimiento no experto es más sencillo pero menos fiable, por lo que requiere técnicas adicionales que garanticen la robustez de las explicaciones [61].

2.3.2.2. Métricas no supervisadas

En el caso de que no exista un ground truth, se identifican las características deseadas en una explicación, y utilizan métricas no supervisadas para medir el grado de cumplimiento de estas. Las métricas más comunmente evaluadas son la fidelidad, la robustez, la transferabilidad, y la usabilidad.

Métricas de evaluación objetiva
Existen problemáticas, como la falta de causalidad con el modelo o datos de los que provienen, o las discrepancias entre explicaciones para modelos o datos similares, que requieren una evaluación cuantitativa. Esta determina la calidad de las explicaciones desde un punto de vista objetivo, que en ocasiones se pasa por alto ante la presencia de una evaluación cualitativa [62].

Por ejemplo, la causalidad entre las explicaciones el modelo o datos de los que provienen se conoce como **fidelidad** («faithfulness»), y se preocupa por que las explicaciones sean fielmente representativas del proceso de toma de decisiones subyacente. Esta se puede evaluar sistemáticamente mediante métodos como los «sanity checks» [61], que proponen aleatorizar los parámetros del modelo («Model Parameter Randomization Test») o conjunto de datos («Data Randomization Test») y medir el impacto que esto tiene en las explicaciones. Idóneamente, las explicaciones generadas tras aleatorizar el modelo (o datos) difieren significativamente de las originales, lo que implica que las explicaciones originales dependen del modelo (o datos) y ofrecen información sobre los mismos. Además, para explicaciones en forma de importancia de características, la fidelidad también mide que las características identificadas como más importantes tengan una influencia significativa en el modelo. En este caso, se pueden emplear técnicas como ROAR [63] para comparar la degradación del modelo reentrenado después de eliminar características supuestamente informativas y características seleccionadas al azar. Idóneamente, la degradación del modelo en el primer caso es mayor que en el segundo.

Por otra parte, la **robustez** se define como el grado en que las explicaciones generadas en condiciones similares son igualmente similares. Esta se puede considerar entre explicaciones para un mismo modelo y método de IAX, o para diferentes modelos y métodos, pero mismo problema de predicción. En el primer caso, se espera que datos de entrada similares generen explicaciones similares, de manera que estas sean robustas a pequeñas perturbaciones y generalizables a datos futuros. Esto se puede evaluar mediante la función de Lipschitz [7] o sensibilidad media [64], que cuantifican la estabilidad de las explicaciones frente a cambios en los datos de entrada. En el segundo caso, se espera que diferentes modelos de ML o técnicas de IAX generen explicaciones similares para el mismo problema de predicción, de manera que reflejen el conocimiento presente en los datos en vez de las particularidades de los métodos por los que se generaron. Para evaluar la robustez en este contexto, y para explicaciones en forma de importancia de características, se pueden utilizar métricas como NDCG (Normalized Discounted Cumulative Gain) o el coeficiente de correlación de Pearson que considera el orden de las características.

Otras métricas mediante que evaluar las explicaciones son la complejidad, monotonía, diversidad o viabilidad. Estas tienen más sentido para determinado tipo que explicaciones que de otras, hasta el punto de que existen autores como Jianlong Zhou et al. [65] que organizan las técnicas cualitativas de validación en base al tipo de explicación. Por ejemplo, para explicaciones en forma de modelo subrogado, tiene sentido evaluar su calidad (que se puede entender como la calidad de dicho modelo, por ejemplo, en términos de accuracy), su tamaño o complejidad; mientras que, para explicaciones en forma de importancia de características, resulta más relevante evaluar su monotonía (la correspondencia entre el orden de importancia de las características y el orden de influencia real en el modelo), y fidelidad mencionada. Además, las explicaciones en forma de ejemplos se podrían evaluar en base a su diversidad o viabilidad (que considera que las explicaciones finales sean factibles o prácticas en el contexto en el que se generaron) [65].

Métricas de evaluación subjetiva
La validez de las explicaciones se puede considerar también desde un punto de vista psicológico, ya que la explicabilidad está ligada a la confianza, transparencia y privacidad, y en realidad los humanos son los consumidores finales de las explicaciones[5]. Desde este punto de vista, para que una explicación sea válida debe ser interpretable, es decir, comprensible por parte de los sujetos para los que fue generada. En algunos contextos, esta cualidad se conoce como **usabilidad**. Además, autores como Hui-wen et al. [66] definen métricas subjetivas adicionales, como la satisfacción del usuario, su concentración o su rendimiento, que pueden resultar relevantes en diversos contextos. Todas ellas se deben evaluar mediante métodos cualitativos (como "casos de estudio" o cuestionarios), ya que requieren interacción del usuario.

2.3.3. Métricas de similitud

Como se ha comentado anteriormente, algunas de las métricas para validar las explicaciones en forma de importancia de características se apoyan en métricas de similitud. En este contexto, para que las explicaciones sean similares, o bien sus valores de importancia para las diferentes características son similares, o bien el orden las características según su importancia es similar. Para evaluar el primer caso, se utiliza error cuadrático medio o RMSE («Root Mean Squared Error») [67], mientras que para el segundo se utiliza el NDCG («Normalized Discounted Cumulative Gain») [60].

2.3.3.1. RMSE

El RMSE calcula similitud entre las explicaciones basándose en la diferencia absoluta entre los valores de importancia de las características. Este se calcula como la raíz cuadrada de la media de los cuadrados de las diferencias entre los valores de importancia atribuidos de una explicación *a* ($imp\_value\_a\_i$) y una explicación *b* ($imp\_value\_b\_i$). Así, un RMSE bajo indica una mayor similitud entre la importancia de características según las diferentes explicaciones. Para calcularlo, se utiliza la expresión de la Ecuación 2.4.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{2.4}$$

2.3.3.2. NDCG

El NDCG también evalúa la similitud entre explicaciones en términos de la importancia de características. Además, esta métrica considera que los elementos que ocupan las posiciones más altas en un ranking deben recibir más crédito cuando se clasifican correctamente que los elementos en las posiciones más bajas. En relación con las explicaciones, es cierto que las características menos importantes probablemente no sean tan informativas sobre el modelo como las más importantes, y es de especial interés que las características más importantes mantengan el mismo orden en diferentes explicaciones. De esta manera, el NDCG ayuda a mitigar el ruido introducido por las características de menor importancia.

Para calcularlo, primero se determina el DCG («Discounted Cumulative Gain»), que consiste en la suma de los elementos clasificados correctamente dividida por el logaritmo de su posición en el ranking (Ecuación 2.5). Aquí, i representa la posición de una característica en el ranking, imp_value_i es su valor de importancia y N es el número total de características. Así, cuanto menor sea la posición de una característica en el ranking, menor será su importancia y menos contribuirá a la suma acumulada de las diferencias.

$$\text{DCG} = \sum_{i=1}^{N} \frac{imp\_value_i}{\log_2(i+1)} \tag{2.5}$$

Para comparar este valor entre diferentes rankings, se normaliza dividiéndolo por el ranking "ideal" o IDCG («Ideal Discounted Cumulative Gain») para los datos, obteniendo un resultado entre 0 y 1 (Ecuación 2.6).

$$\text{NDCG} = \frac{DCG}{IDCG} \tag{2.6}$$

2.3.4. Metodologías en IAX

Hasta ahora, una estrategia común de evaluación ha consistido en mostrar ejemplos de explicaciones individuales y comprobar que parezcan razonables [68], o que pasen una primera prueba de "validez aparente" [69]. Aun así, muchos autores argumentan que depender de tales pruebas anecdóticas es insuficiente, e incluso puede ser "engañoso" [61].

Algunos autores proponen una metodología para evaluar y validar las técnicas de IAX independientemente del problema de predicción subyacente. Por ejemplo, Liu et al. [70] proponen el uso de conjuntos de datos sintéticos, de los que se conoce la distribución real de los datos, y para los que se puede calcular exactamente la distribución condicional sobre cualquier conjunto de características y las diferentes métricas de evaluación. Aun así, los métodos y medidas óptimos para evaluarlas y validarlas podrían depender del dominio de la aplicación, el tipo de explicación, el tipo de datos, los conocimientos previos del usuario y el propósito de las explicaciones, por lo que aunque una determinada técnica obtenga buenos resultados sobre conjuntos de datos sintéticos, no tendría por qué hacerlo para todos los problemas de predicción.

Además, existe una conexión inherente entre la evaluación de la precisión predictiva del modelo "caja negra" y la calidad de las explicaciones generadas sobre el mismo.

Aunque ambos fueran independientes, en algunos casos no se pueden distinguir. Por ejemplo, Samek et al. [71] comentan que en los mapas de calor que explican algoritmos de visión por computadora: "la calidad del mapa de calor no solo depende de los algoritmos utilizados para calcularlo, sino también del rendimiento del clasificador, cuya eficiencia depende en gran medida del modelo que se utilice, y de la cantidad y calidad de los datos de entrenamiento disponibles".

Otras iniciativas proponen utilizar un conjunto de métricas para evaluar las explicaciones generadas por diferentes técnicas de IAX y así comparar su rendimiento con respecto a la tarea subyacente, como es el caso de Bommer et al. [72]. Sin embargo, esta propuesta no ofrece indicaciones claras sobre cómo integrar las diferentes pruebas realizadas en una metodología de explicabilidad que permita validar las explicaciones generadas y orientar la selección de las mejores explicaciones para una tarea de clasificación.

Por otra parte, Jin et al. [73] proponen directrices clínicas de IAX («Clinical XAI Guidelines») fundamentadas en perspectivas clínicas y técnicas, que respaldan la selección y el diseño de técnicas de IAX clínicamente viables para tareas de imágenes médicas. En ellas se definen una serie de criterios imprescindibles en el dominio, que las explicaciones deben superar de manera incremental para ser apropiadas para su uso clínico. Sin embargo, este tipo de metodologías solo son aplicables a un ámbito específico y no tienen un propósito general.

Capítulo 3

Desarrollo

En este capítulo, se presentará la metodología de evaluación propuesta y los escenarios de prueba sobre los que se ejemplificará. Para ello, en la Sección 3.1 se justificará la selección de métricas y pruebas de evaluación implementadas, en la Sección 3.2 se integrarán en una metodolgía de explicabilidad y comentará su utilización para los diferentes propósitos planteados, y finalmente, en la Sección 3.3 se presentarán las bases de datos, modelos y técnicas de explicabilidad utilizadas y ofrecerán detalles sobre su implementación.

3.1. Selección de métricas y pruebas

Para evaluar las explicaciones generadas para un conjunto de datos y modelo, así como para guiar la selección entre diferentes explicaciones, se propone evaluar su fidelidad y robustez. Estas métricas son ampliamente reconocidas en la literatura y se pueden calcular de manera objetiva para diferentes tipos de explicaciones. Además, al no requerir conocimiento previo, son más flexibles y pueden calcularse fácilmente para las explicaciones generadas para cualquier problema de predicción.

La fidelidad de las explicaciones asegura que sean representativas del proceso subyacente de toma de decisiones y no aleatorias. Para evaluar la fidelidad de las explicaciones en forma de importancia de características, se utilizarán métodos como sanity checks y ROAR. Estos métodos tienen una implementación consolidada y son comunmnete utilizados para la evaluación de explicaciones. Además, con pequeñas modificaciones, podrían ser empleados también sobre otros tipos de explicaciones.

Por otra parte, la robustez de las explicaciones mide su consistencia bajo condiciones similares. Esta se puede evaluar desde diferentes perspectivas: en un contexto de multiplicidad predictiva y bajo condiciones fijas. En el primer caso, se propone comparar las explicaciones de diferentes modelos de ML y técnicas de IAX entre sí, mientras que en el segundo se propone comparar las explicaciones del mismo modelo de ML y técnica de IAX generadas sobre datos distribuidos de manera ligeramente diferente. En ambos casos, se aprovecha la morfología de las explicaciones en forma de importancia de características y utilizan métricas de similitud para calcular su robustez de manera intuitiva.

A continuación, se detalla la implementación de las diferentes pruebas de evaluación de fidelidad y robustez.

3.1.1. Evaluación de fidelidad

3.1.1.1. Sanity checks

Los «sanity checks» [61] son un conjunto de pruebas de aleatorización diseñadas para evaluar la robustez y la validez de las explicaciones generadas por un modelo. Estas alteran de manera aleatoria algunos aspectos clave del modelo o conjunto de datos, para generar nuevas explicaciones y compararlas con las originales. La idea central es que, si las explicaciones dependen efectivamente del modelo y los datos subyacentes, entonces alterar aleatoriamente estos elementos debería producir explicaciones significativamente diferentes de las originales y en mayor magnitud de lo que difieren las explicaciones originales entre sí (en diferentes iteraciones de la misma técnica de IAX sobre el mismo modelo y datos). Esto ayuda a determinar si las explicaciones son adecuadas para interpretar los resultados de la tarea de predicción.

En el artículo original, se proponen dos pruebas de aleatorización diferentes: para aleatorizar las etiquetas del conjunto de datos («Data Randomization Test») o los parámetros del modelo («Model Parameter Randomization Test»). La primera compara las explicaciones originales con aquellas generadas a partir de datos con etiquetas aleatorias. En ambos casos, se utiliza el mismo modelo (del mismo tipo y con los mismos parámetros) y misma técnica de IAX para generar las explicaciones. Además, la segunda compara las explicaciones generadas sobre el modelo y datos originales con aquellas generadas sobre un modelo inicializado aleatoriamente (del mismo tipo pero con diferentes parámetros, aleatorios). Así, los parámetros de este segundo modelo difieren de los del modelo. Esta prueba fue diseñada originalmente para redes neuronales que permiten la inicialización aleatoria de sus pesos, pero no es directamente aplicable a otros modelos como SVM (en el que sus parámetros se inicializa a 0 [74]) o XGBoost (que no acepta ningún tipo de inicialización). Por este motivo, se ha redefinido la prueba de aleatorización de los parámetros del modelo para que pueda realizarse también sobre modelos de tipo SVM y XGBoost. En este caso, se comparan las explicaciones originales con aquellas generadas sobre un modelo entrenado sobre datos con etiquetas aleatorias (los mismos que se utilizaron para la primera prueba). Así, los datos de los que se generan las predicciones y explicaciones son los originales, pero el modelo que realiza las predicciones ha sido entrenado con datos aleatorizados, y por lo tanto sus parámetros difieren de los del modelo original.

Para cada prueba, se calcula la similitud entre las explicaciones originales y las alteradas para 10 subconjuntos de los datos de manera independiente, obteniendo así una medida de similitud más robusta y representativa. La significancia de estas diferencias se evalúa mediante un conjunto de pruebas estadísticas. En primer lugar, se utiliza la prueba de Shapiro-Wilk [75] para contrastar si podemos aproximar la distribución con una distribución normal (con la media y la varianza de la muestra), y, de ser así, se emplea la prueba t de Welch [76] para comparar si la diferencia entre las distribuciones es relevante. En caso contrario, se utiliza la prueba U de Mann-Whitney [77]. Como hipótesis nula, se plantea que las distribuciones contrastadas son iguales. Para todas las pruebas, se establece un umbral de significancia de 0.05, de manera que un p-valor inferior a 0.05 indicará que las distribuciones no cumplen con la hipótesis nula, y existe una diferencia significativa entre ambas.

Para que un método IAX pase los sanity checks, debe generar explicaciones tales que la similitud entre las explicaciones originales y las alteradas sea significativamente diferente de la similitud entre las explicaciones originales consigo mismas («baseli-

ne»). De esta manera, un p-valor inferior a 0.05 demostrará que las explicaciones alteradas son significativamente diferentes de las originales, confirmando así que las explicaciones dependen del modelo y datos subyacentes.

3.1.1.2. RemOve And Retrain - ROAR

ROAR [63] es una técnica empleada para comprobar que las características más importantes según las técnicas de IAX de importancia de características tengan una influencia significativa en el modelo. La motivación detrás de ROAR es asegurar que las explicaciones proporcionadas por el modelo, en forma de importancia de características, sean realmente útiles para interpretar los resultados de la tarea de predicción. Si las características más importantes realmente influyen en el modelo, su eliminación debería degradar el rendimiento del modelo de manera significativa en comparación con la eliminación de características seleccionadas al azar.

Para ello, reentrena el modelo con diferentes subconjuntos de características y se compara la degradación (en términos de accuracy) del modelo reentrenado después de eliminar características supuestamente informativas y características seleccionadas al azar. Las características más informativas, según un método de importancia de características, son aquellas con mayor valor de importancia. Para ROAR, se ordenan en orden de importancia de mayor a menor, y se elimina un porcentaje fijo de las características más importantes en cada iteración, de manera incremental, hasta acabar con el conjunto vacío. En cada iteración también se reentrena y reevalúa el modelo sobre el nuevo subconjunto de características reducido. Simultáneamente, se elimina un porcentaje fijo de características al azar y reentrena y reevalúa el mismo modelo sobre este otro nuevo subconjunto de características. Finalmente, se compara la degradación de ambos modelos entrenados con diferentes porcentajes de características, seleccionadas de manera diferente, para evaluar la fidelidad de las explicaciones.

Se ha considerado realizar un experimento complementario para secundar las conclusiones del original, en el que se agrega un porcentaje fijo de las características hasta acabar con todas ellas, y reentrena y reevaluar el modelo en cada iteración. En este caso, la adición de características importantes debería mejorar el rendimiento del modelo de manera significativa en comparación con la adición de características seleccionadas al azar.

La degradación para cada k% de características se denomina "k-degradación" [63]. De manera similar, llamaremos la mejora para cada k% de características "k-mejora". Los valores de k seleccionados van desde 0 hasta 95, con un paso de 5. Además, la eliminación de una característica se ha implementado como el reemplazo de los valores de la característica con el valor medio de esa característica, y, por el contrario, la adición de una característica se ha implementado como la inclusión de los valores de características originales en el conjunto de características "vacío". En realidad, este conjunto de características no está vacío, sino compuesto por los valores medios de cada característica. Al igual que para los sanity checks, la degradación y mejora se calculan para 10 subconjuntos de los datos de manera independiente, y así obtener una medida más robusta y representativa de la población total. Para realizar las comparaciones descritas, primero se realiza el promedio de la degradación y mejora sobre todos los subconjuntos.

La condición de cumplimiento para ROAR, en el caso de la "k-degradación", es que la degradación del modelo al eliminar las características más importantes debe ser significativamente mayor que al eliminar características al azar. En el caso de "k-mejora", se espera que la mejora del modelo reentrenado al añadir las características más importantes debe ser significativamente mayor que al añadir características al azar. Esto indicará que cada característica informativa adicional es al menos más informativa que una característica seleccionada al azar, validando así que las explicaciones en forma de importancia de características son adecuadas para explicar los resultados de la tarea de predicción. En la práctica, se espera que esta degradación o mejora sea destacable para la mayoría de valores de k.

3.1.2. Evaluación de robustez

3.1.2.1. Robustez a multiplicidad predictiva

En el aprendizaje automático, "la posibilidad de que un problema de predicción admita múltiples modelos alternativos que funcionan casi igual de bien"se conoce como multiplicidad predictiva [78]. Normalmente, la selección del modelo se basa en su desempeño (en términos de precisión, por ejemplo), y en un escenario de multiplicidad predictiva, no existe un argumento claro para seleccionar un modelo frente a otro. Además, cada modelo generará predicciones ligeramente diferentes a partir de los mismos datos y, por ende, las explicaciones de la tarea de predicción también diferirán. Esta problemática es reconocida por Breiman, quien plantea: "si se pueden ajustar múltiples modelos competidores, cada uno de los cuales proporciona una explicación diferente del proceso generador de datos, ¿cómo podemos saber cuál explicación es correcta?"[79]. En la literatura, este dilema se conoce como el «efecto Rashomon», haciendo referencia a la película "Rashomon", en la que los testimonios discordantes de los testigos se asemejan a las explicaciones discordantes de los modelos competidores [79]. Por lo tanto, bajo la multiplicidad predictiva, se limitan las conclusiones que se pueden extraer de un único modelo y sus explicaciones, al menos hasta que se pueda descartar esta multiplicidad.

Para evaluar las explicaciones bajo la multiplicidad predictiva, una de las pruebas necesarias es medir la similitud entre las explicaciones de los posibles modelos para el problema de predicción. Esta similitud se calcula a nivel de característica, considerando el valor de importancia de una determinada característica en los diferentes modelos. Para ello, se enfrentan los vectores de importancia de características para cada 2 modelos en un diagrama de dispersión, donde cada punto i del gráfico tiene como coordenadas el valor de importancia para la característica i según un modelo A y otro B. Si los puntos se alinean con la recta de pendiente 1, las explicaciones son similares a nivel de característica para ambos modelos considerados.

Además, dicha similitud se mide de manera empírica mediante el error cuadrático medio o RMSE («Root Mean Squared Error») [67], que toma en cuenta las diferencias entre los valores de importancia de las características para ambos modelos. También se utiliza el NDCG («Normalized Discounted Cumulative Gain») [60], que considera la magnitud de los valores de importancia al contabilizar sus diferencias. Ambos métodos permiten una evaluación cuantitativa de la similitud entre las explicaciones generadas por diferentes modelos, proporcionando una base para determinar la robustez de las explicaciones bajo la multiplicidad predictiva.

La condición de cumplimiento para esta prueba es que las explicaciones generadas por diferentes modelos deben mostrar una alta similitud en términos de las medidas de RMSE y NDCG. Si los valores de estas medidas son bajos, esto indicará que las explicaciones proporcionadas por los diferentes modelos son consistentes entre sí, validando así la robustez de las explicaciones a pesar de la existencia de multiplicidad predictiva.

3.1.2.2. Robustez a cambios en la distribución

La robustez de las explicaciones también se puede evaluar considerando las explicaciones generadas para un mismo modelo y problema de predicción. En el caso de utilizar una técnica de IAX local, que genera explicaciones a nivel de instancia que pueden ser inconsistentes en el contexto global, la magnitud de las inconsistencias puede ser un indicativo de la calidad de las explicaciones. Además, esta medida permite estimar hasta qué punto las explicaciones son transferibles a ejemplos futuros que, aunque provengan de la misma población, probablemente siguen una distribución ligeramente diferente. Esto se debe a la inevitable presencia del elemento temporal en una explicación, lo que significa que en el momento en que se genera una explicación ya es posible que la información haya cambiado lo suficiente como para que se vuelva obsoleta [80].

Para evaluar la robustez ante cambios en la distribución de los datos, se investiga hasta qué punto las explicaciones generadas por un mismo modelo son consistentes cuando se utilizan diferentes subconjuntos de entrenamiento y test del mismo conjunto de datos. robustas a cambios menores en la distribución de los datos. Esto incluye las variaciones entre diferentes subconjuntos de entrenamiento y prueba tomados del mismo conjunto de datos. Esto implica entrenar modelos sobre los distintos subconjuntos de entrenamiento y generar las explicaciones sobre los distintos subconjuntos de test, midiendo luego la similitud entre estas explicaciones. La similitud entre las explicaciones generadas a partir de diferentes subconjuntos proporciona una estimación de cómo las explicaciones podrían comportarse en el futuro frente a datos que siguen una distribución similar.

Para visualizar y cuantificar esta similitud entre explicaciones, se comparan los vectores de importancia de características para cada par de subconjuntos en un diagrama de dispersión para visualizar esta similitud. En este diagrama, cada punto i del gráfico tiene como coordenadas el valor de importancia para la característica i según el modelo entrenado con el subconjunto A y el entrenado con el subconjunto B. La alineación de los puntos con la recta de pendiente 1 indica que las explicaciones son similares a nivel de característica para ambos modelos considerados. Además de la visualización, esta similitud se evalúa empíricamente utilizando el RMSE y el NDCG, que cuantifican las diferencias entre los valores de importancia de características para cada par de subconjuntos y proporcionan una evaluación cuantitativa de la consistencia de las explicaciones frente a cambios en la distribución de los datos.

La condición de cumplimiento para esta prueba es que las explicaciones generadas por el mismo modelo, pero entrenado en diferentes subconjuntos de datos, deben mostrar una alta similitud en términos de las medidas de RMSE y NDCG. Esto demostrará que las explicaciones son robustas y consistentes ante variaciones menores en la distribución de los datos, lo cual es fundamental para su aplicabilidad en escenarios futuros.

3.2. Desarrollo de la metodología

A continuación, se integran las pruebas definidas en una metodología de explicabilidad que sirva tanto para garantizar la generación de explicaciones adecuadas como para guiar la selección de las mejores explicaciones para una tarea de clasificación. Esta metodología aparece representada en la Figura 3.1.

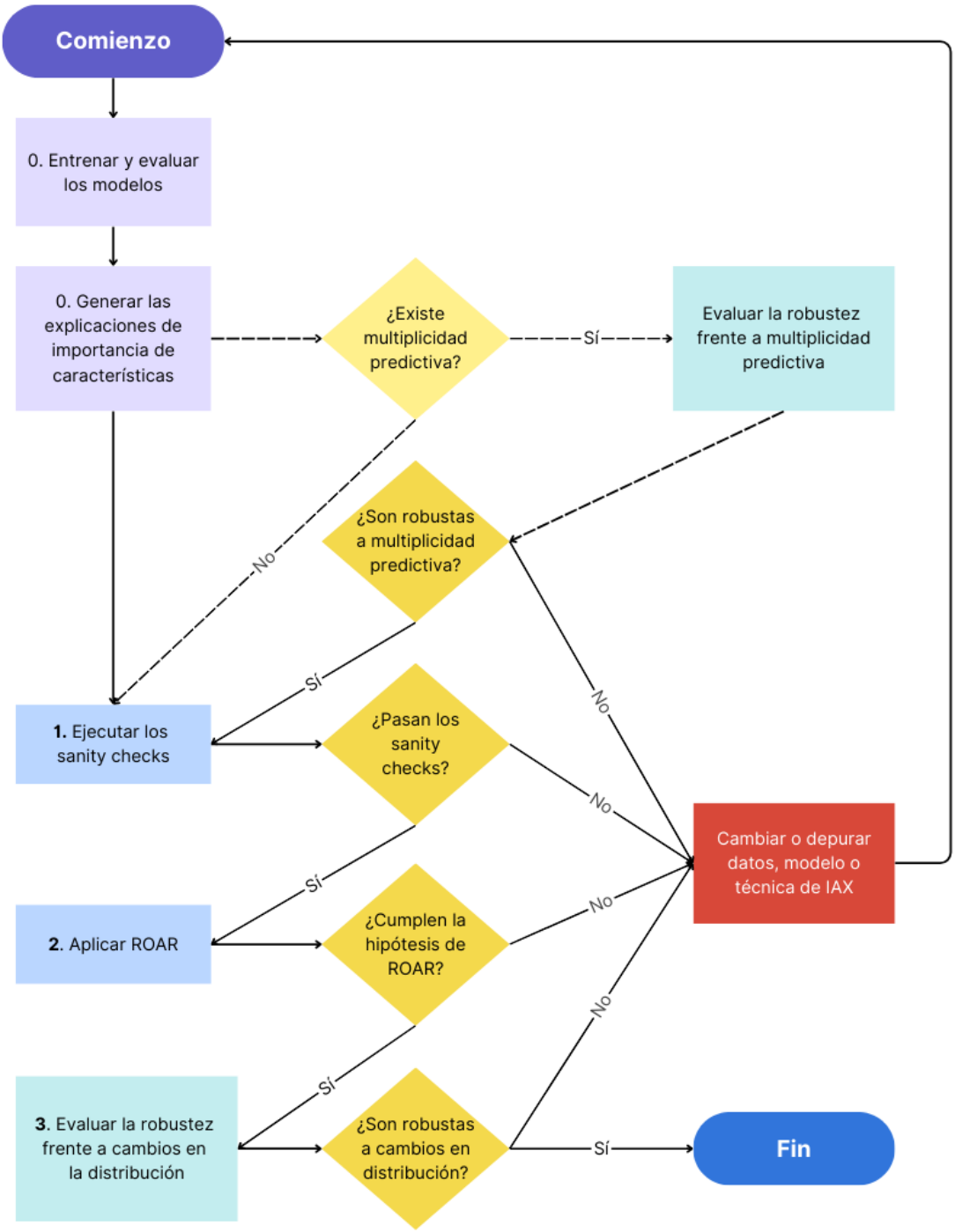


Figura 3.1: Metodología de evaluación de explicaciones en forma de importancia de características

Como se observa en la figura, esta se ejecuta sobre las explicaciones en forma de importancia de características generadas a partir de un modelo y técnica de IAX para un problema de clasificación. En primer lugar, en la etapa de entrenamiento del modelo, se propone considerar si nos encontramos en un escenario de multiplicidad predictiva. Para ello, se necesitaría entrenar diferentes modelos para la misma tarea de clasificación, lo que puede no ser computacionalmente posible para determinados contextos. De esta manera, aunque evaluar esta situación y en su caso ejecutar las pruebas de robustez a multiplicidad predictiva ofrecería mayor garantía acerca de la validez de las explicaciones, no resulta imprescindible para la validación o selección de las explicaciones mediante la metodología propuesta. Una vez que se supere (o descarte) esta primera etapa, las explicaciones primero superar las pruebas de evaluación de fidelidad, y en caso de hacerlo, posteriormente obtener resultados aceptables en las pruebas de evaluación de robustez ante cambios en la distribución. En el caso de no superar alguna de las pruebas, es recomendable regenerar las explicaciones para la tarea de clasificación, posiblemente alterando algunos aspectos del modelo, datos o técnica de IAX de importancia de características empleada.

Por otra parte, si se desea seleccionar las mejores explicaciones para un problema de clasificación, se deben ejecutar las pruebas de evaluación de fidelidad y robustez y observar aquellas explicaciones para las que se obtienen mejores resultados consistentemente. En el caso de contar con múltiples explicaciones prometedoras, se deben elegir aquellas que mejores resultados obtengan para las pruebas de evaluación que sean de mayor interés para la tarea de clasificación. Además, en caso de que ninguna de las explicaciones generadas no superen las pruebas de evaluación de fidelidad, no se recomienda hacer una selección, sino generar explicaciones alternativas.

3.3. Selección de escenarios de prueba

Finalmente, se ejemplifica la metodología en práctica sobre 2 conjuntos de datos con diferentes características, 3 tipos de modelos y 2 técnicas de explicabilidad de importancia de características. A continuación, se detallan las características y criterios de selección de los mismos.

3.3.1. Conjuntos de datos

Para poner en práctica la metodología, se han elegido 2 conjuntos de datos con características significativamente diferentes, y provenientes también de ámbitos diferentes. La primera contiene los valores m/z de pacientes con Covid-19, con sus correpondientes intensidades en el rango de masas de 5 a 20 kDA, obtenidos mediante la técnica MALDI-TOF [81]. Con ella, se pretende resolver la tarea de clasificación de diferenciar los pacientes con Covid-19 de los pacientes sin Covid-19. Esta está compuesta por 191 variables y 275 instancias, y nos referiremos a ella como *covid19* a lo largo del documento. La segunda se trata del conjunto de datos conocido como "Adult" o "Census Income" del "UCI Machine Learning Repository" [82]. Esta contiene 14 variables de información censal de los individuos, con 48.842 instancias, y plantea la tarea de clasificación de predecir si los ingresos de un determinado individuo superan los $50,000 anuales. En este documento, nos referiremos a esta base de datos como *census-income*.

3.3.2. Modelos y métricas de evaluación

Además, se han generado explicaciones de 3 tipos de modelos entrenados sobre los diferentes conjuntos de datos. Estos son XGBoost, SVM y MLP, seleccionados por sus considerables diferencias arquitectónicas, así como por su implementación sencilla y usabilidad en una amplia gama de aplicaciones. Para cada conjunto de datos, se han entrenado 10 modelos de cada tipo, a partir de diferentes subconjuntos de entrenamiento y test. En total, se han entrenado 30 modelos por conjunto de datos. Estos modelos, entrenados con diferentes subconjuntos, ofrecen el soporte necesario para evaluar la robustez de las explicaciones ante cambios en la distribución descrita anteriormente. Además, aportan robustez a los resultados del resto de etapas de la metodología.

En todos los casos, los modelos se han entrenado con el 80 % de las instancias, y evaluado con el 20 % restante. Como función de optimización, se ha utilizado el área bajo la curva o AUROC («Area Under the Receiver Objective Characteristic»), ya que compara la tasa de verdaderos positivos y falsos positivos, y es invariante al ratio de clases del conjunto de datos. De esta manera, asegura que el modelo sea evaluado de manera justa y representativa independientemente de la distribución de clases en los datos.

Además, para identificar los mejores hiperparámetros, se utilizó validación cruzada con cv=10. Aun así, dado que los mejores hiperparámetros para cada tipo de modelo variaban según el subconjuntos de entrenamiento usado, se eligieron los hiperparámetros más populares entre todos los subconjuntos, y entrenaron los modelos finales con estos para todos los subconjuntos. Además, todos los modelos se inicializaron con una semilla, lo que garantiza su reproducibilidad.

3.3.3. Técnicas de explicabilidad

Para generar las explicaciones, se han utilizado 2 de las técnicas de IAX de importancia de características más maduras y utilizadas en diversos contextos: SHAP y LIME. Ambos métodos generan explicaciones locales; sin embargo, dado que nos interesa identificar tendencias y patrones generales en el comportamiento del modelo, se han promediado para conseguir explicaciones globales. Es importante tener en cuenta que, en ello, se puede perder información relevante sobre los casos particulares, lo que significa que las explicación globales serán solo aproximadas.

Las explicaciones se han generado sobre las instancias de test (en lugar de las de entrenamiento), ya que se ha considerado más interesante interpretar las predicciones de los modelos (y entender los motivos por los que fueron generadas) más que el funcionamiento de los mismos (lo que aprendieron durante el entrenamiento). Así, se pueden extraer conclusiones relevantes sobre la transferibilidad de las explicaciones a instancias futuras, mediante la evaluación de la robustez ante cambios en la distribución. Dichas explicaciones no pretenden ser las óptimas para el problema de clasificación, y por eso tampoco han sido generadas sobre instancias de validación, sino que pretenden dar soporte a la métodología planteada. Aun así, puesto a que se generan sobre distintos subconjuntos de test, se asemejan a las hipotéticas explicaciones de validación.

Capítulo 4

Resultados

En este capítulo, se presentarán y analizarán los resultados obtenidos para cada fase de la metodología planteada aplicándola a distintos conjuntos de datos. Debido a sus características significativamente diferentes, estos conjuntos de datos ilustran diversos escenarios y permiten una discusión de la metodología desde diferentes ángulos. En la subsección 4.1, se comentarán los resultados para cada conjunto de datos de forma individual, mientras que en la subsección 4.2, se discutirán las conclusiones derivadas de la implementación de la metodología en su conjunto.

4.1. Resultados

4.1.1. Conjunto de datos 1: *covid19*

4.1.1.1. Ajuste de modelos

En primer lugar, se ajustan y entrenan los modelos XGBoost, SVM y MLP sobre los 10 subconjuntos del conjunto de datos. Tras utilizar validación cruzada con cv=10 sobre cada subconjunto, se observa que los hiperparámetros más populares entre todos los subconjuntos son los siguientes:

- Para **XGBoost**: min_child_weight = 2, max_depth=5 y n_estimators=100.
- Para **SVM**: kernel='rbf', gamma=1e-3 y C=100.
- Para **MLP**: hidden_layer_sizes=(10,10,10,10,10), activation='relu', solver='lbfgs' y alpha=1e-5.

Así, se entrenaron los modelos con los hiperparámetros mencionados en cada subconjunto, obteniendo los valores medios y las desviaciones típicas que se muestran en la Tabla 4.1. En esta tabla se especifican el AUROC, la sensibilidad y la especificidad de la clase positiva para los diferentes modelos, evaluados en los subconjuntos de prueba. El AUROC fue la métrica optimizada durante el entrenamiento de los modelos, mientras que la sensibilidad y la especificidad reflejan otras cualidades de los modelos que no están recogidas en el AUROC.

Cuadro 4.1: Métricas para los modelos de *covid19*, evaluadas sobre los subconjuntos de test

	AUROC	Sensibilidad	Especificidad
XGBoost	0.8619 ± 0.07	0.9759 ± 0.03	0.7478 ± 0.15
SVM	0.9644 ± 0.06	0.9847 ± 0.01	0.9442 ± 0.11
MLP	0.9448 ± 0.04	0.9844 ± 0.01	0.9207 ± 0.08

Como se puede observar, los modelos SVM y MLP obtienen valores parecidos de media para las métricas seleccionadas, por lo que podemos considerar que nos encontramos ante una situación de multiplicidad predictiva. Además, aunque los modelos XGBoost obtengan valores ligeramente inferiores de media, también manifiestan mayor desviación típica para los diferentes subconjuntos de los datos, por lo que cabe imaginar que existen subconjuntos para los que dicho modelo se ajusta considerablemente mejor.

4.1.1.2. Evaluación de fidelidad

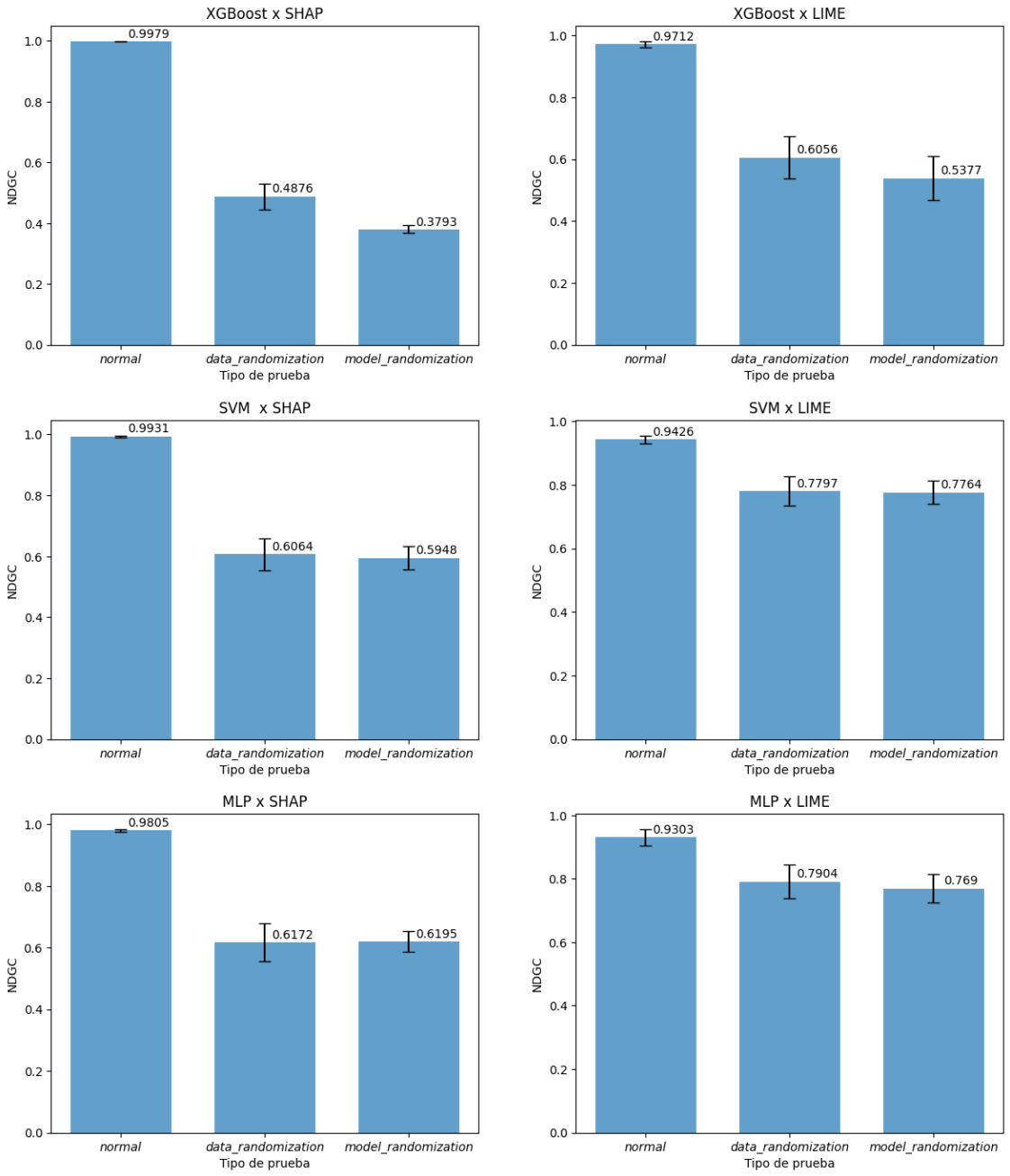
Sanity checks

A continuación, se aplicaron los sanity checks para evaluar la robustez de las explicaciones generadas a partir de diferentes modelos y técnicas de IAX sobre el primer conjunto de datos. En la Figura 4.1, se muestran los resultados donde se compara la similitud entre las explicaciones originales ("baseline"), y la similitud entre las explicaciones originales y las alteradas de ambas maneras (”data_randomization” y ”model_randomization”).

Para el ”baseline”, se calculó la similitud entre las explicaciones generadas a partir del modelo y datos originales consigo mismas. Para ello, se necesita al menos generar 2 conjuntos de explicaciones bajo las mismas condiciones, con las que calcular la similitud. Además, para el ”data_randomization” se calculó la similitud entre las explicaciones originales y las generadas tras aleatorizar las etiquetas del conjunto de datos, y para el ”model_randomization” la similitud entre las explicaciones originales y las generadas tras aleatorizar los parámetros del modelo. En todos los casos, se utilizó el NDCG medio como métrica de similitud entre las explicaciones globales.

Como se observa en la Figura 4.1, el ”baseline” muestra un valor medio superior que el ”data_randomization” y ”model_randomization” para todos los modelos y técnicas de IAX utilizadas. Además, la variabilidad entre las repeticiones es baja en todos los casos. Esta se calcula mediante la desviación típica y representa mediante el tamaño de una recta perpendicular a cada columna del histograma. Aun así, para determinar la significancia de estas diferencias, se utilizan las prueba t de Welch o la prueba U de Mann-Whitney (dependiendo del caso) y obtienen los p-valores de la Tabla 4.2. Todos los p-valores fueron inferiores al umbral de significancia de 0.05, lo que indica que las distribuciones contrastadas son significativamente diferentes. Por lo tanto, se superan satisfactoriamente el «Data Randomization Test» y «Model Parameter Randomization Test» para todos los modelos y técnicas de IAX en el primer conjunto de datos.

Por otra parte, se observaron diferencias significativas entre los modelos y técnicas de IAX. Por ejemplo, las discrepancias en las métricas de similitud fueron particularmente notables para el modelo XGBoost con SHAP, sugiriendo una mayor dependencia del modelo y los datos en las explicaciones generadas mediante esta combinación.



(a) SHAP (b) LIME

Figura 4.1: Sanity checks para explicaciones de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP de *covid19*, generadas con (a) SHAP y (b) LIME.

Cuadro 4.2: p-valor para los test de significancia de los sanity checks sobre las explicaciones de *covid-19*.

	SHAP			LIME		
	SVM	**XGBoost**	**MLP**	**SVM**	**XGBoost**	**MLP**
data_random* vs. *normal	1.83 E-04	1.83 E-04	2.66 E-08	9.44 E-07	4.39 E-08	8.13 E-06
model_random* vs. *normal	1.83 E-04	1.83 E-04	7.92 E-11	3.29 E-08	1.21 E-08	1.36 E-07

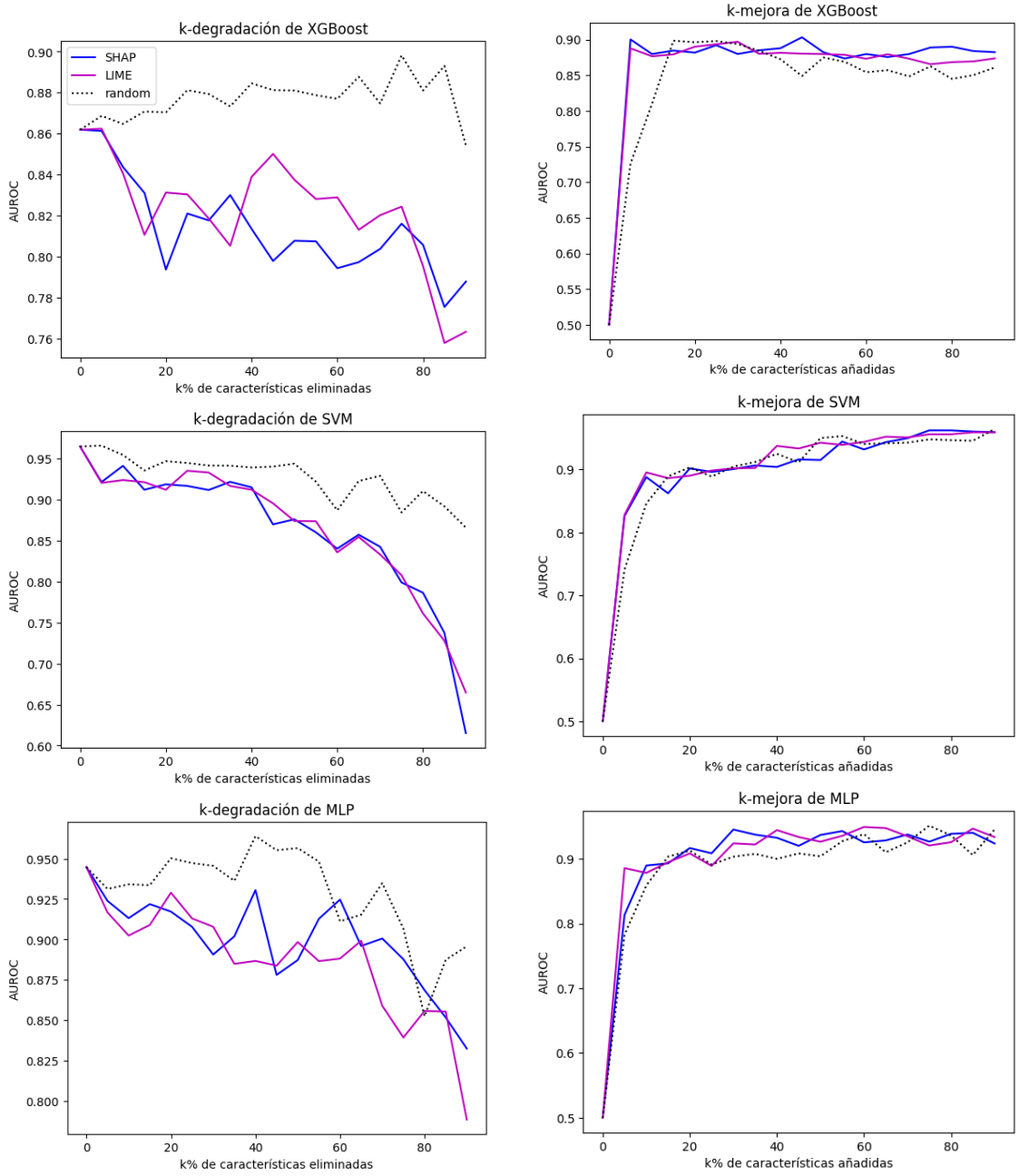
RemOve And Retrain - ROAR

La segunda técnica para evaluar la fidelidad de las explicaciones en forma de importancia de características es ROAR. Esta compara la degradación del modelo reentrenado (en términos de AUROC) después de eliminar características supuestamente informativas y características seleccionadas al azar, y de manera complementaria, la mejora del modelo reentrenado al añadir características supuestamente informativas y características seleccionadas al azar. Los resultados de k-degradación y k-mejora se presentan en la Figura 4.2. En cada gráfica, las líneas continuas azules y moradas representan el AUROC del modelo reentrenado en base a las explicaciones generadas mediante SHAP y LIME, respectivamente, mientras que las líneas punteadas muestran el AUROC con características aleatorias.

En términos de k-degradación, se observa que la degradación del modelo reentrenado utilizando las explicaciones es generalmente mayor que la del modelo reentrenado con características aleatorias para todos los valores de k (porcentaje de características eliminadas). Sin embargo, la degradación no sigue una tendencia monótonamente decreciente, sugiriendo que, aunque las explicaciones identifiquen características importantes más informativas que características seleccionadas al azar, estas no están ordenadas en base a su verdadera importancia. Esto puede deberse la existencia de redundancias entre las características, que perjudica el funcionamiento de SHAP y LIME al generar explicaciones.

Además, en términos de k-mejora, la mejora del modelo reentrenado considerando las explicaciones es comparable o ligeramente superior a la del modelo reentrenado con características aleatorias para la mayoría de valores de k. Nuevamente, esto puede ser consecuencia de la existencia de variables correlacionadas. Cabe destacar que la mejora al añadir el 5 y 10% de las características más importantes es notablemente superior en comparación con añadir un 5 y 10% de características aleatorias, lo que indica que las características identificadas como las más importantes son efectivamente informativas. Por lo tanto, aunque las explicaciones generadas por ambas técnicas pueden ser mejoradas, logran pasar las pruebas de k-degradación y k-mejora para todos los modelos, lo que sugiere que pueden ser representativas para la tarea de predicción.

Por otra parte, cabe destacar que las explicaciones generadas para XGBoost tienen un impacto significativamente mayor en la degradación de este modelo en comparación con las demás explicaciones generadas para sus respectivos modelos. Esto sugiere nuevamente que las explicaciones generadas para XGBoost son las más adecuadas para este problema de clasificación.



(a) k-degradación (b) k-mejora

Figura 4.2: ROAR: (a) k-degradación y (b) k-mejora de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP de *covid19* según las explicaciones generadas con SHAP y LIME.

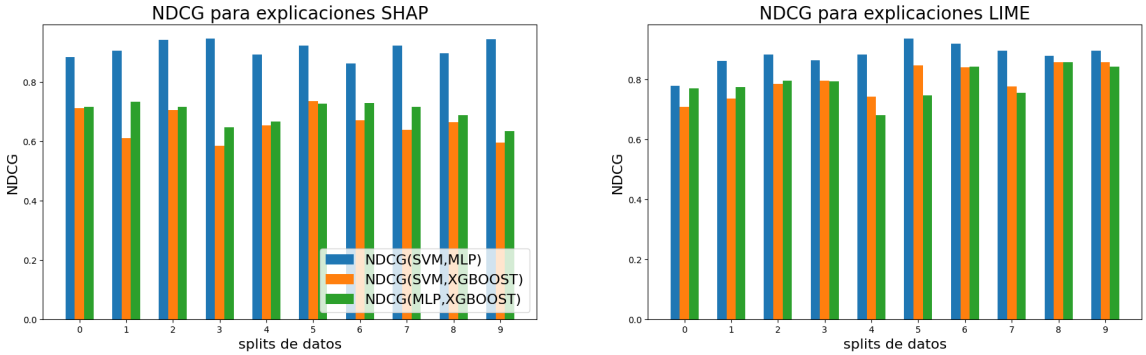
4.1.1.3. Evaluación de robustez

Una vez que las explicaciones han superado las pruebas de fidelidad, se procede a evaluar su robustez ante multiplicidad predictiva y cambios en la distribución de los datos. A continuación, se presentan los resultados para ambas pruebas.

Robustez a multiplicidad predictiva

Como se ha observado en la Sección 4.1.1.1, los modelos SVM, XGBoost y MLP obtienen valores parecidos para las métricas seleccionadas (especialmente SVM y MLP, que además obtienen valores ligeramente superiores). Ante multiplicidad predictiva, es importante evaluar la similitud entre las explicaciones generadas por diferentes modelos. Esta se evaluará para distintos subconjuntos de entrenamiento y test (splits), seleccionados de manera aleatoria y común para los diferentes modelos.

Primero, se calculó la similitud entre las explicaciones como el NDCG medio entre las explicaciones globales de cada par de modelos, y se muestra en la Figura 4.3. En esta, se observa una mayor similitud entre las explicaciones generadas por SVM y MLP para todos los subconjuntos, tanto para las explicaciones generadas mediante SHAP como para las generadas mediante LIME. Para las primeras, el NDCG tiene un valor medio de 0.9108 entre los modelos SVM y MLP, frente a un valor de 0.6563 y 0.6967 entre los modelos XGBoost y SVM, y XGBoost y MLP, respectivamente. Además, para las explicaciones generadas mediante LIME, tiene un valor medio de 0.8807 (frente a 0.7954 y 0.7868). Los valores de similitud para cada subconjunto (y su media) para todos los modelos y técnicas de IAX se pueden observar en detalle en las Tablas A.1 y A.2.

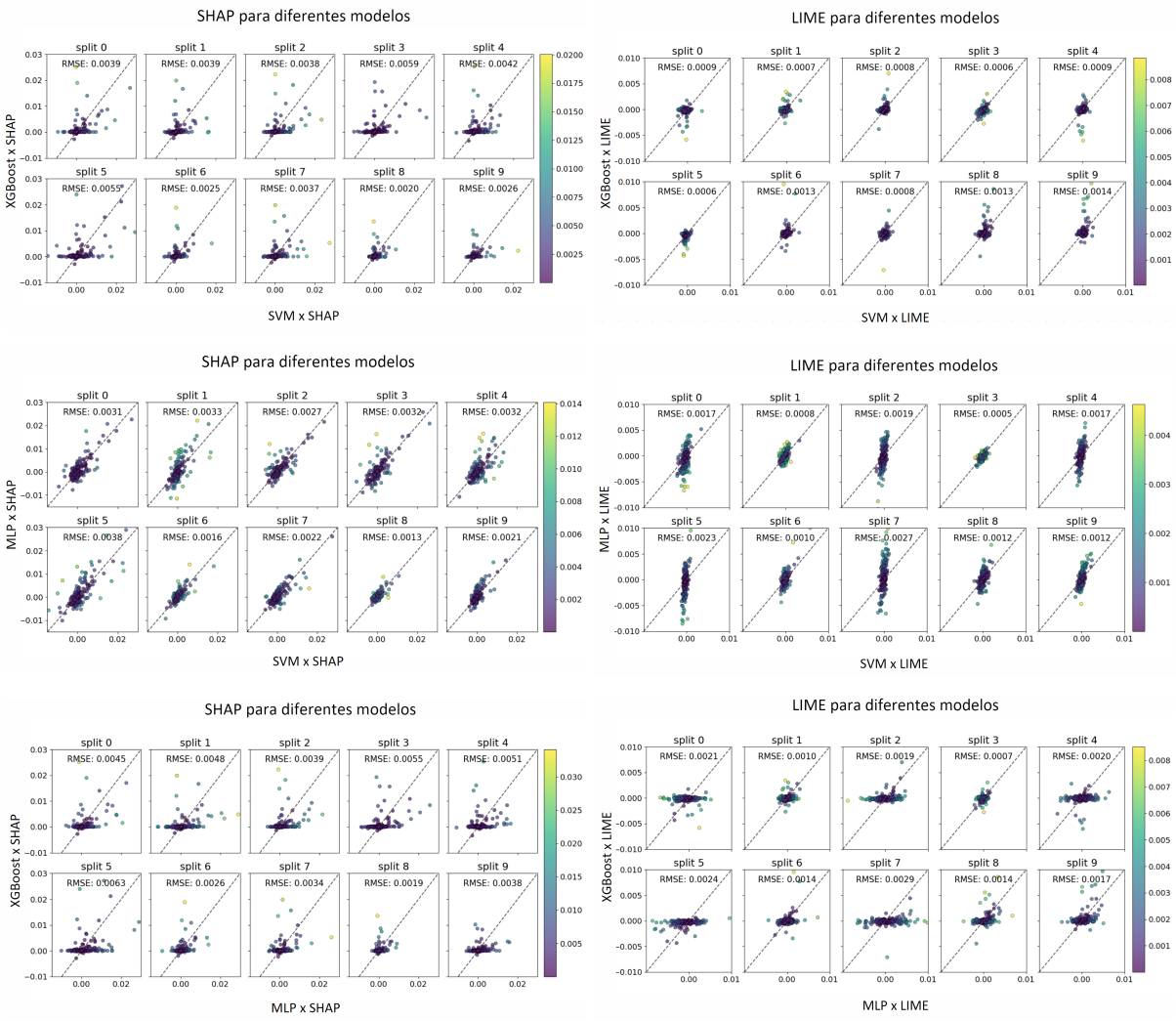


(a) SHAP (b) LIME

Figura 4.3: NDCG entre las explicaciones de los modelos SVM y XGBoost (azul), SVM y MLP (naranja), y MLP (verde) y XGBoost con diferentes subconjuntos (splits) de *covid19*, generadas con (a) SHAP y (b) LIME.

Además, se analizaron los diagramas de dispersión de los valores de importancia de características entre pares de modelos en la Figura 4.4. En general, las explicaciones de SHAP muestran una distribución más dispersa que las de LIME, lo cual se refleja en valores de RMSE más bajos para SHAP. Nuevamente, las explicaciones generadas por los modelos SVM y MLP muestran una mejor alineación con la línea x=y, que representa una similitud perfecta, en comparación con los demás pares de modelos.

Todo ello señala una mayor consistencia entre las explicaciones producidas por SVM y MLP, especialmente aquellas generadas mediante SHAP, lo que sugiere que podrían ser intercambiables para explicar la tarea de clasificación de manera efectiva. Aun así, los valores de similitud entre el resto de pares de modelos siguen siendo altos, lo que indica que las explicaciones proporcionadas por los diferentes modelos son bastante consistentes entre sí a pesar de la multiplicidad predictiva.



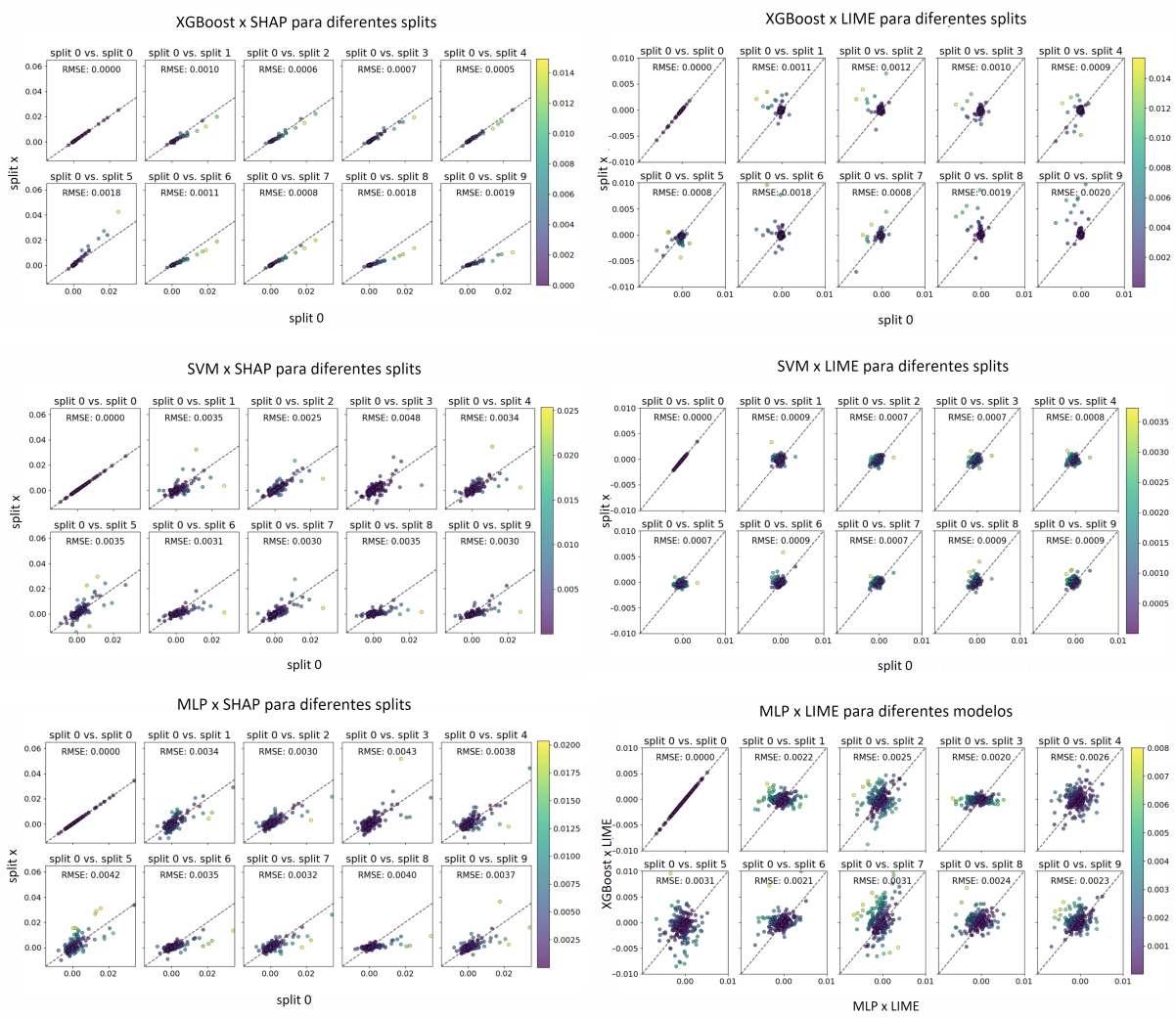
(a) SHAP (b) LIME

Figura 4.4: Diagramas de dispersión entre las explicaciones de los modelos (superior) SVM y XGBoost, (intermedio) SVM y MLP, y (inferior) MLP y XGBoost de *covid19*, generadas con (a) SHAP y (b) LIME.

Robustez a cambios en la distribución

Finalmente, se evalúa la robustez de las explicaciones generadas para un mismo modelo pero con subconjuntos de datos que siguen distribuciones ligeramente diferentes.

Primero, se analizaron los diagramas de dispersión de los valores de importancia de características generados a partir de diferentes subconjuntos de los datos. Con 10 subconjuntos distintos disponibles, se podrían realizar hasta 10! comparaciones distintas. En la Figura 4.5, se presentan unicamente los diagramas de dispersión que comparan las explicaciones generadas a partir del subconjunto 0 (split 0) con todos los demás (split x, donde x toma valores entre 0 y 9). En estos gráficos, los puntos de la primera gráfica están perfectamente alineados con la línea x=y para cada modelo y técnica de IAX considerados, ya que enfrentan explicaciones generadas a partid del mismo subconjunto.

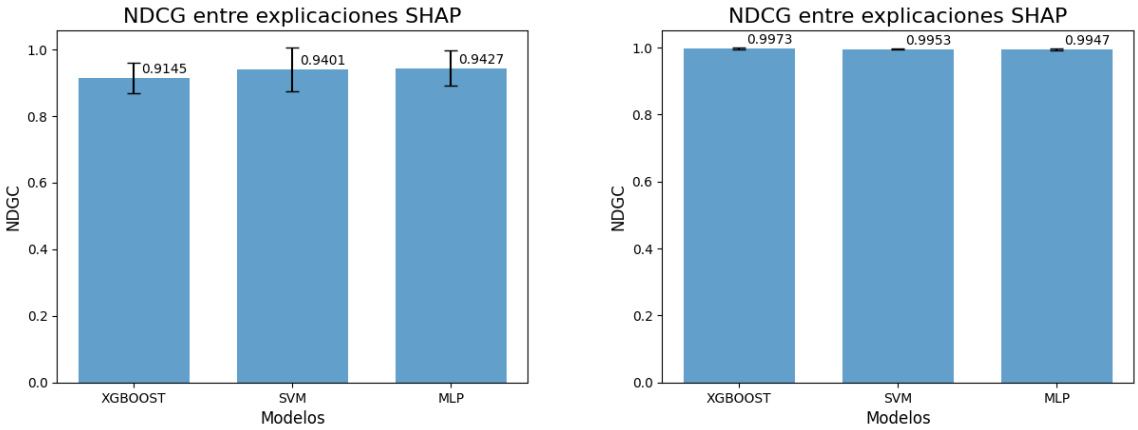


(a) SHAP (b) LIME

Figura 4.5: Diagramas de dispersión entre las explicaciones de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP con diferentes subconjuntos (splits) de *covid19*, generadas con (a) SHAP y (b) LIME.

Nuevamente, las explicaciones generadas mediante SHAP sobre modelos XGBoost muestran una mejor alineación con la línea x=y que las generadas sobre modelos SVM y MLP para todos los subconjuntos. Los valores de RMSE empíricamente apoyan este observación, ya que son inferiores en el primer caso que en los demás. Aun así, en el caso de LIME, las explicaciones que mejor se ajustan a la recta x=y son las generadas sobre SVM.

Además, se calculó la similitud entre las explicaciones para un mismo modelo como el NDCG medio entre las explicaciones globales para todos sus subconjuntos, y se muestra en la Figura 4.6. Los resultados sugieren que las explicaciones generadas por SHAP son más consistentes ante cambios en la distribución que las generadas por LIME, que presentan mayor variabilidad. Aun así, los valores de similitud en general son altos, lo que sugiere que las explicaciones generadas mediante las diferentes técnicas y modelos son consistentes entre sí a pesar de los cambios en la distribución.



(a) SHAP

(b) LIME

Figura 4.6: NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP de *covid19*, generadas con (a) SHAP y (b) LIME.

4.1.2. Conjunto de datos 2: *census-income*

4.1.2.1. Ajuste de modelos

Primero, se ajustan y entrenan los modelos SVM, XGBoost y MLP sobre los 10 subconjuntos del conjunto de datos. Se seleccionan los siguientes hiperparámetros:

- Para **XGBoost**: min_child_weight = 5, max_depth=10 y n_estimators=100.
- Para **SVM**: kernel='rbf', gamma=1e-3 y C=100.
- Para **MLP**: hidden_layer_sizes=(5, 2), activation='relu', solver='lbfgs' y alpha=1e-5.

Además, se obtienen los valores medios y las desviaciones típicas del AUROC, sensibilidad y especificidad mostrados en la Tabla 4.3. Nuevamente, nos encontramos ante una situación de multiplicidad predictiva, especialmente entre XGBoost y MLP. En este caso, los modelos de tipo SVM ajustados generan peores resultados para todas las métricas. Aun así, los 3 tipos de modelos generan resultados significativamente peores que para el problema de clasificación anterior.

Cuadro 4.3: Métricas para los modelos de *census-income*, evaluadas sobre los subconjuntos de test

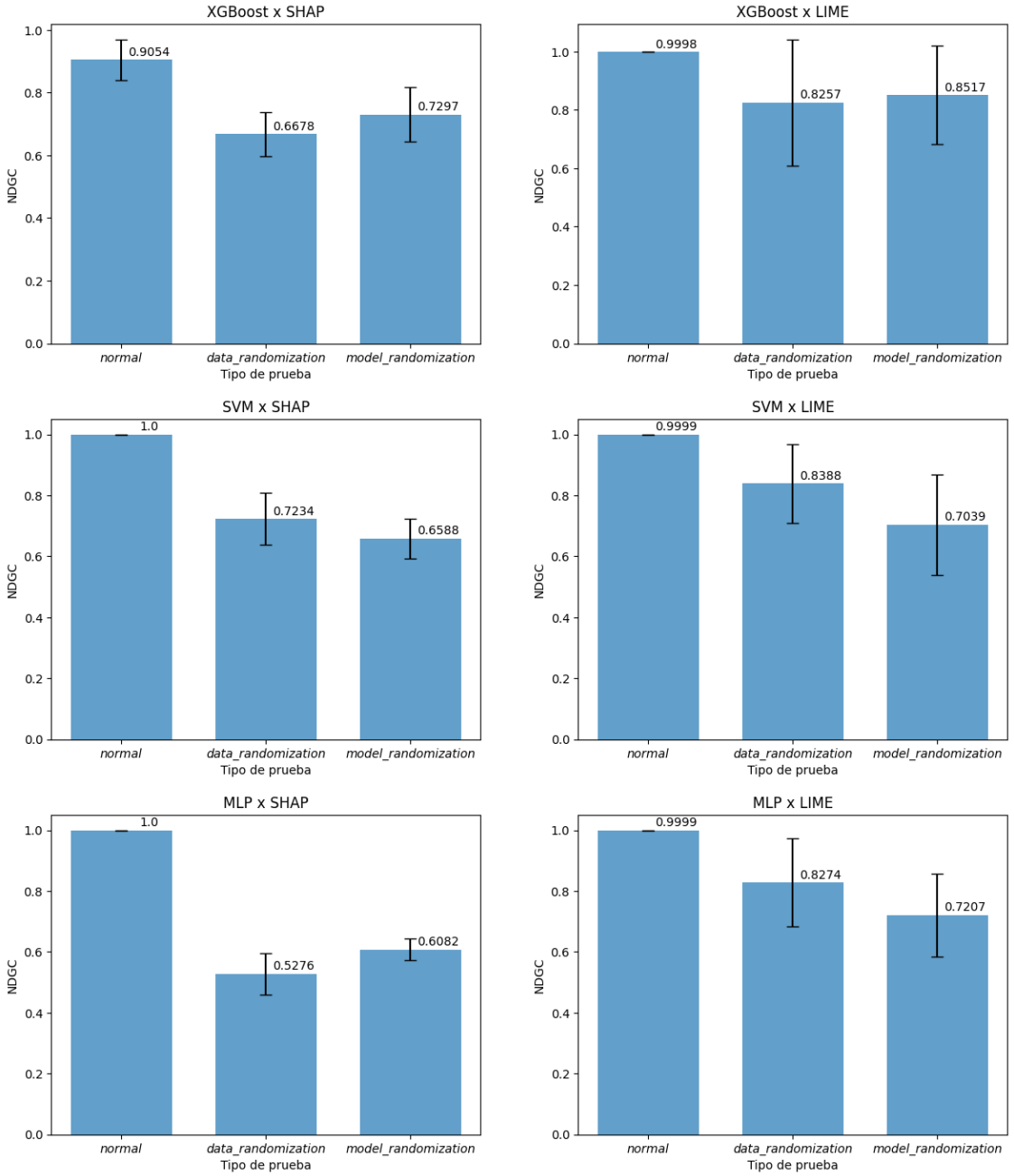
	AUROC	Sensibilidad	Especificidad
XGBoost	0.7756 ± 0.02	0.6113 ± 0.04	0.9399 ± 0.02
SVM	0.6974 ± 0.02	0.4302 ± 0.04	0.9646 ± 0.01
MLP	0.7540 ± 0.01	0.5853 ± 0.03	0.227 ± 0.01

4.1.2.2. Evaluación de fidelidad

Sanity checks

A continuación, se aplican los sanity checks, cuyos resultados se muestran en la Figura 4.7.

Aunque el "baseline" tiene un valor medio superior que el "data_randomization" y "model_randomization" para todos los modelos y técnicas de IAX utilizadas, la desviación típica parece significativa para la mayoría de casos, especialmente para las explicaciones generadas mediante LIME.



(a) SHAP (b) LIME

Figura 4.7: Sanity checks para explicaciones de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP de *census-income*, generadas con (a) SHAP y (b) LIME.

Para evaluar la significancia de las diferencias entre las distribuciones, se calcularon los p-valores detallados en la Tabla 4.4. Sorprendentemente, todos ellos son menores que el umbral de significancia establecido (0.05), lo que indica que las distribuciones contrastadas no son significativamente diferentes. Por lo tanto, se superan satisfactoriamente las pruebas de «Data Randomization Test» y «Model Parameter Randomization Test» para todos los modelos y técnicas de IAX. Sin embargo, es importante destacar que las explicaciones generadas para este conjunto de datos no son tan dependientes del modelo y los datos subyacentes como en el caso anterior, lo que las hace menos adecuadas para la tarea de clasificación. Esto se debe a que las diferencias entre las columnas son menores que para el conjunto de datos *covid19*, así como a los p-valores más altos obtenidos en las pruebas de significancia.

Cuadro 4.4: p-valor para los test de significancia de los sanity checks sobre las explicaciones de *census-income*.

	SHAP			LIME		
	SVM	**XGBoost**	**MLP**	**SVM**	**XGBoost**	**MLP**
data_random* vs. *normal	4.24 E-06	6.23 E-07	1.46 E-04	1.83 E-04	3.30 E-04	1.83 E-04
model_random* vs. *normal	7.35 E-08	1.60 E-04	1.46 E-04	1.83 E-04	1.83 E-04	1.83 E-04

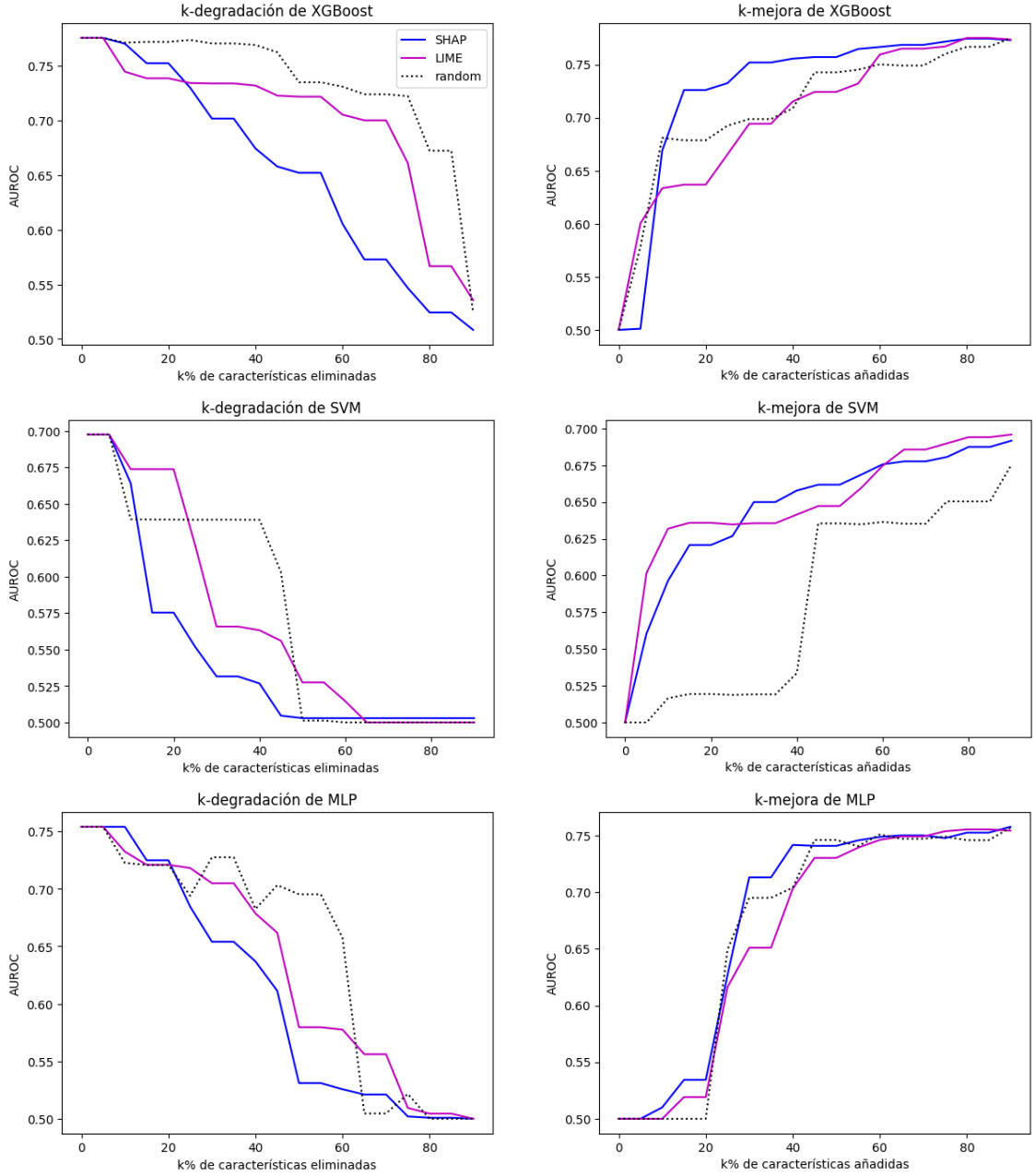
RemOve And Retrain - ROAR

A continuación, se aplica ROAR, y se muestra la k-degradación y k-mejora resultantes en la Figura 4.8.

En términos de k-degradación, la degradación de los modelos reentrenado considerando las explicaciones es generalmente mayor que la del modelo reentrenado con características aleatorias para la mayoría de los valores de k. Entre ellos destaca positivamente la degradacion de los modelos XGBoost reentrenados. Además, la degradación en todos los casos sigue una tendencia monótonamente decreciente, sugiriendo que las explicaciones son capaces de ordenar las características en base a su verdadera importancia.

Por otra parte, los resultados de k-mejora varían significativamente entre los diferentes modelos y técnicas. Cuando se consideran las explicaciones SHAP o LIME generadas sobre SVM, la mejora del modelo reentrenado es notablemente superior a la obtenida utilizando características aleatorias. Sin embargo, esta mejora no es tan prominente para las explicaciones generadas sobre los otros modelos, especialmente cuando se utiliza LIME.

Es importante destacar que los resultados de k-degradación y k-mejora para un mismo modelo y técnica muestran discrepancia significativa acerca de la fidelidad de las explicaciones. Esta discrepancia sugiere que las explicaciones no necesariamente contribuyen de la misma manera a la degradación y mejora del modelo, lo cual plantea interrogantes sobre pueden su representatividad de la tarea de clasificación. Por lo tanto, se podría considerar que las explicaciones generadas no pasan esta prueba, y especialmente aquellas generadas con LIME.



(a) k-degradación (b) k-mejora

Figura 4.8: ROAR: (a) k-degradación y (b) k-mejora de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP de *census-income* según las explicaciones generadas con SHAP y LIME.

4.1.2.3. Evaluación de robustez

A partir de los resultados anteriores, parece razonable considerar la posibilidad de descartar las explicaciones generadas, dado que muestran una baja dependencia del modelo y de los datos subyacentes, lo cual las hace menos idóneas para la tarea de clasificación. No obstante, resulta intrigante analizar cómo esta falta de fidelidad afecta a su robustez, por lo que procederemos con la metodología propuesta.

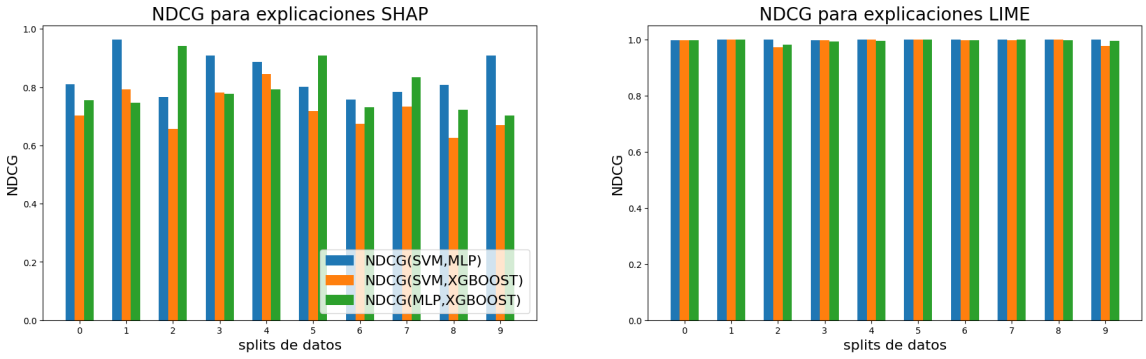
Robustez a multiplicidad predictiva

En primer lugar, se enfrentan los vectores de importancia de características para cada 2 modelos en los diagramas de dispersión de la Figura 4.10. Es importante recordar que, para este problema de clasificación, se observa multiplicidad predictiva, especialmente entre los modelos de tipo XGBoost y MLP.

En los diagramas de dispersión de la Figura 4.10, las explicaciones generadas por SHAP muestran una vez más una distribución menos compacta que las generadas por LIME. A pesar de esto, no se refleja en el valor del RMSE, ya que para la mayoría de explicaciones LIME se observa un punto distante y aislado que incrementa el valor de esta métrica. En este caso, no se observan diferencias entre la similitud de las explicaciones generadas por los diferentes pares de modelos.

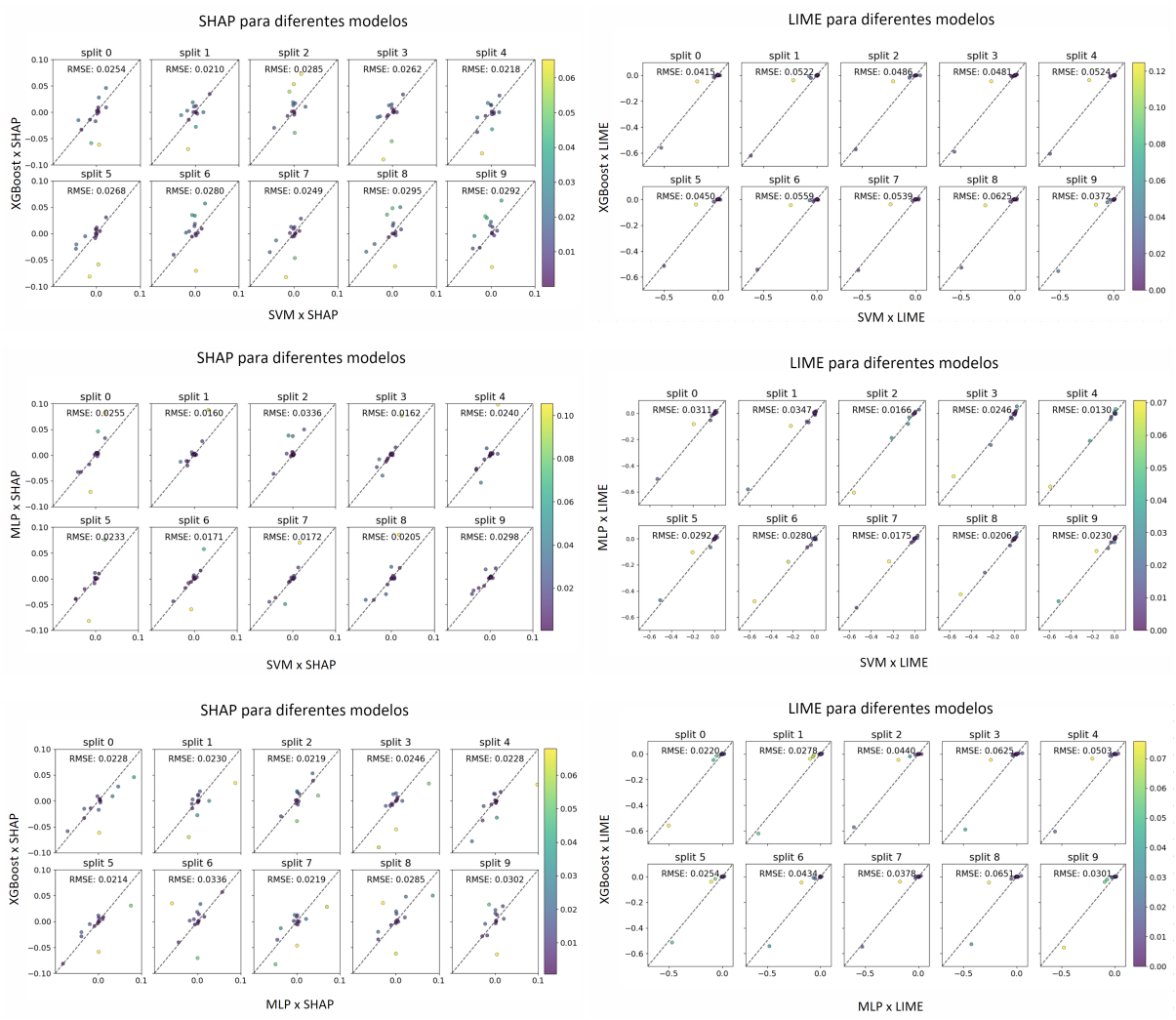
Por otra parte, el NDCG medio entre las explicaciones globales de cada par de modelos se representa en la Figura 4.9. Nuevamente, no se observan diferencias significativas entre la similitud de las explicaciones generadas por los diferentes pares de modelos, especialmente entre las generadas por LIME. De hecho, las explicaciones presentan una similitud alta entre casi todos los pares de modelos, con valores medios de 0.9929 entre las explicaciones generadas por SVM y XGBoost, 0.9985 entre SVM y MLP, y 0.9951 entre MLP y XGBoost. Esto puede parecer contradictorio, dado que según las métricas de la Tabla 4.3, no todos los modelos aprenden igual. En el caso de SHAP, la pareja de modelos para la cual las explicaciones son más similares varía para cada subconjuntos. No obstante, su NDCG es menor que en los casos anteriores, con valores medios de 0.7914, 0.8399 y 0.7201, respectivamente. Los valores de similitud para cada subconjunto (y su media) para todos los modelos y técnicas de IAX se pueden observar en detalle en las Tablas A.3 y A.4.

Por lo tanto, las explicaciones generadas por SHAP no parecen coincidir entre diferentes modelos, especialmente entre los modelos MLP y XGBoost, donde existe multiplicidad predictiva. Además, aunque las explicaciones generadas por LIME muestran una alta similitud, se sabe que la capacidad para explicar adecuadamente el problema de clasificación subyacente es cuestionable. Todo esto sugiere que esta prueba de evaluación de robustez por sí sola no resulta adecuada para validar las explicaciones generadas.



(a) SHAP (b) LIME

Figura 4.9: NDCG entre las explicaciones de los modelos SVM y XGBoost (azul), SVM y MLP (naranja), y MLP (verde) y XGBoost con diferentes subconjuntos (splits) de *census-income*, generadas con (a) SHAP y (b) LIME.



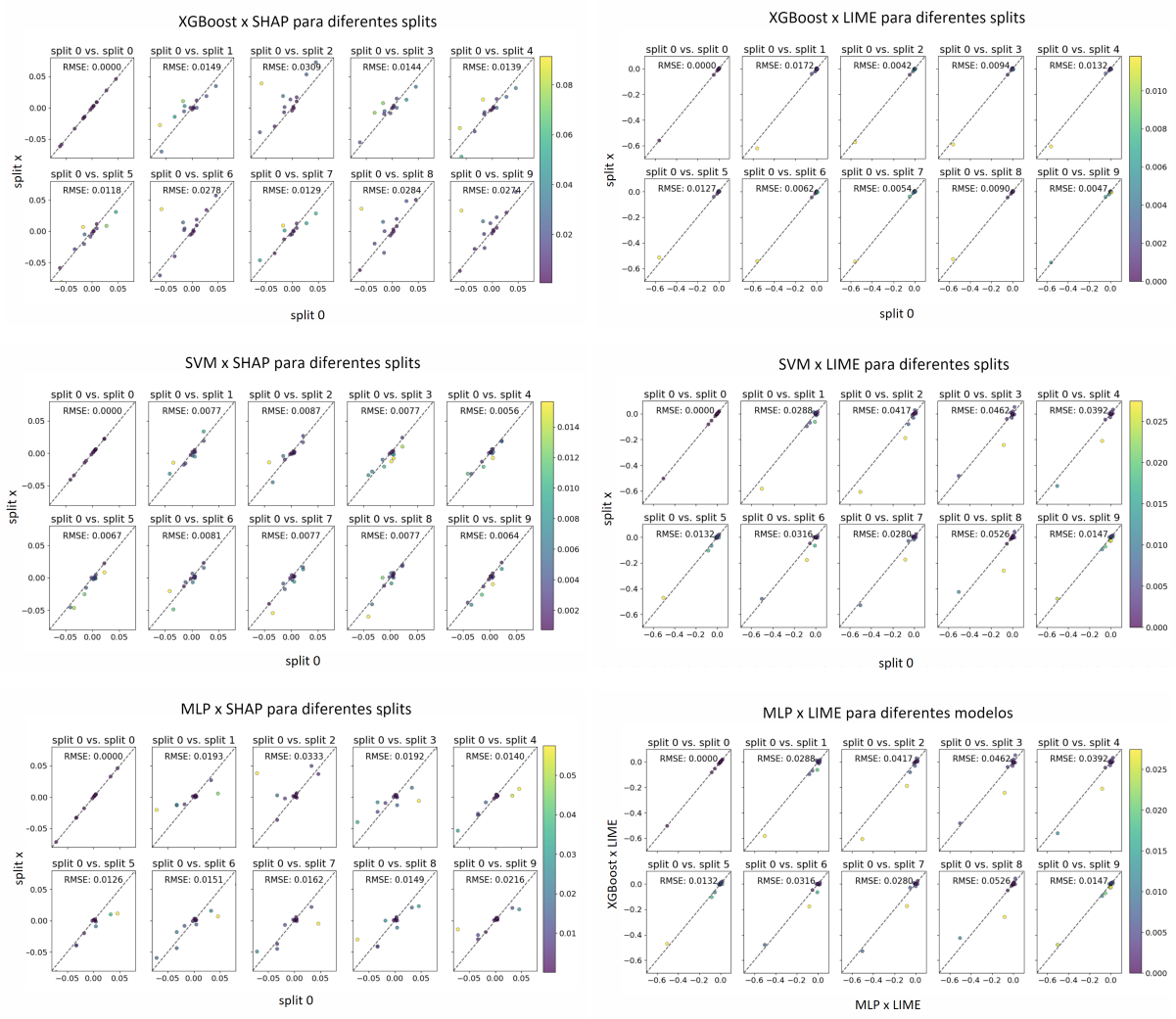
(a) SHAP (b) LIME

Figura 4.10: Diagramas de dispersión entre las explicaciones de los modelos (superior) SVM y XGBoost, (intermedio) SVM y MLP, y (inferior) MLP y XGBoost de *census-income*, generadas con (a) SHAP y (b) LIME.

Robustez a cambios en la distribución

Finalmente, se evalúa la robustez de las explicaciones generadas para un mismo modelo pero con subconjuntos de datos que siguen distribuciones ligeramente diferentes.

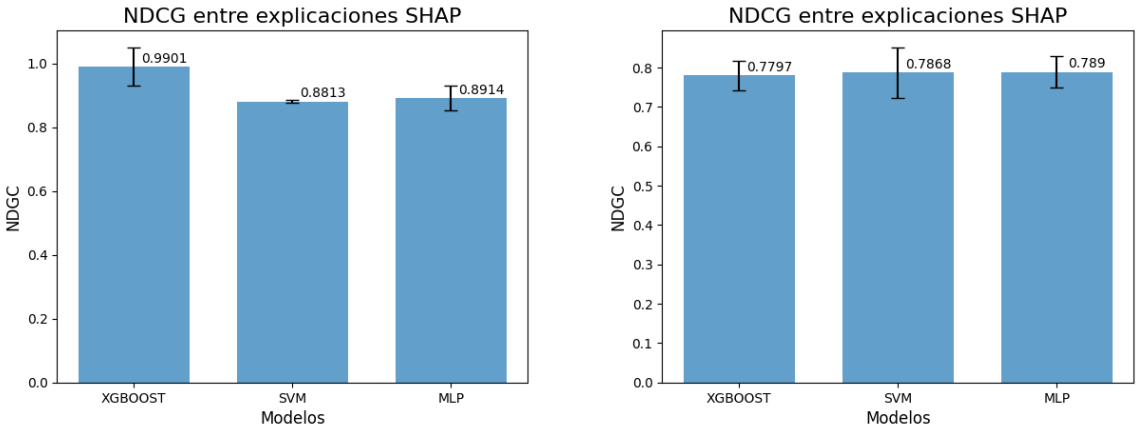
En la Figura 4.11, se observan los diagramas de dispersión que comparan las explicaciones generadas a partir del subconjunto 0 (split 0) con todos los demás (split x). Ahora, las explicaciones generadas mediante SHAP sobre modelos SVM se ajustan mejor a la recta x=y que las generadas sobre el resto de modelos para la mayoría de los subconjuntos. Además, las explicaciones generadas mediante LIME sobre modelos XGBoost son las más similares entre sí de las generadas mediante esta técnica de IAX. Estas observaciones se respaldan con los valores de RMSE. No obstante, a simple vista, las diferencias entre los modelos no parecen ser demasiado significativas.



(a) SHAP (b) LIME

Figura 4.11: Diagramas de dispersión de las explicaciones de los modelos (superior) XGBoost, (intermedio) SVM y (inferior) MLP con diferentes subconjuntos (splits) de *census-income*, generadas con (a) SHAP y (b) LIME.

Por otra parte, en la Figura 4.12 se presenta el NDCG medio de las explicaciones globales para cada modelo. De nuevo, cabe destacar que las explicaciones generadas por LIME muestran una similitud casi perfecta, a pesar de que se sabe su capacidad para explicar adecuadamente el problema de clasificación subyacente es cuestionable. Por otro lado, las explicaciones generadas por SHAP muestran valores altos de NDCG, subrayando así la consistencia de las explicaciones obtenidas mediante esta técnica. Todo esto nuevamente sugiere que esta prueba de evaluación de robustez por sí sola no resulta adecuada para validar las explicaciones generadas.



(a) SHAP (b) LIME

Figura 4.12: NDCG medio entre las explicaciones de los modelos XGBoost, SVM y MLP de *census-income*, generadas con (a) SHAP y (b) LIME.

4.2. Discusión y limitaciones

A partir de los resultados anteriores, se pueden realizar numerosas observaciones. En primer lugar, todas las explicaciones generadas sobre el conjunto de datos de *covid19* parecen adecuadas para todos los modelos y técnicas de IAX, ya que superan las pruebas de evaluación de fidelidad y robustez con éxito. No obstante, los resultados obtenidos en pruebas como ROAR sugieren la existencia de redundancias en las datos que perjudican el funcionamiento de SHAP y LIME al generar explicaciones.

En general, las explicaciones generadas mediante SHAP obtienen mejores resultados en todas las pruebas que las generadas mediante LIME, lo que sugiere que SHAP podría ser una mejor opción para explicar la tarea de clasificación. Entre los diferentes modelos, XGBoost genera explicaciones más similares entre sí de acuerdo con la prueba de evaluación de robustez para las explicaciones de un mismo modelo, y supera los sanity checks y ROAR con mayor facilidad que los demás modelos. No obstante, los modelos SVM y MLP obtienen mejores resultados según las métricas de evaluación y sus explicaciones son más similares entre sí de acuerdo con la prueba de evaluación de robustez ante multiplicidad predictiva, además de obtener resultados competentes en el resto de las pruebas. Por tanto, las explicaciones generadas mediante SHAP a partir de los tres tipos de modelo son adecuadas para la tarea de clasificación, y su selección depende del objetivo específico de las mismas.

Por otra parte, las explicaciones generadas sobre el conjunto de datos de *census-income* también superan los sanity checks, aunque con menos éxito que para el caso anterior. Esto se debe a que la diferencia entre el "baseline" y demás pruebas es menor y presentan una desviación típica alta. Además, no superan la segunda prueba de evaluación de fidelidad, ya que los resultados de la k-degradación y k-mejora para las explicaciones generadas a partir de un mismo modelo y técnica no concuerdan, y existen numerosos valores de k para los que no se cumple la condición establecida por ROAR. Esto sugiere que las explicaciones son mejorables y deben mejorarse antes de su utilización, posiblemente alterando aspectos de los modelos, datos o técnicas de IAX de importancia de características empleadas.

Asimismo, al ejecutar las pruebas de evaluación de robustez sobre explicaciones poco adecuadas, se obtienen resultados que pueden llevar a confusión. En este caso, el NDCG entre las explicaciones en un contexto de multiplicidad predictiva y bajo condiciones fijas es mayor que para el caso anterior, especialmente para las explicaciones generadas mediante LIME. Sin embargo, dado que estas explicaciones no cumplen con la métrica de fidelidad, su alta concordancia es irrelevante, ya que no reflejan la realidad de los datos y el modelo subyacentes. Por ello, es crucial evaluar la fidelidad de las explicaciones antes de su robustez.

Las diferencias entre los resultados para los diferentes conjuntos de datos también se pueden deber a sus características. Estos sugieren que SHAP genera explicaciones más robustas para conjuntos de datos dispersos (o conjuntos de datos de alta dimensión donde la densidad de instancias es baja), mientras que LIME lo hace para conjuntos de datos densos [83].

Como limitaciones, cabe destacar que el criterio de evaluación utilizado en las pruebas de robustez no es determinista, y lo que se considera una "alta similitud en términos de las medidas de RMSE y NDCG"depende en gran medida de aquel que lo determine. Esto se debe a que no se han podido realizar los experimentos necesarios para establecer un umbral de decisión preciso. Además, en el caso de que varias explicaciones superen las pruebas y no haya una unequivocamente victoriosas en todas ellas (como ocurre para el conjunto de datos *covid19*), no se proporciona un criterio único para la selección de las mejores explicaciones. Así, su selección depende del objetivo específico de las mismas, para cada problema de predicción.

Capítulo 5

Conclusiones y Trabajo Futuro

En este capítulo, se concluirá el trabajo realizado. En la subsección 4.2 se extraerán conclusiones acerca de la metodología planteada, así como su impacto en relación a los Objetivos de Desarrollo Sostenible (ODS). Finalmente, en la subsección 5.2, se comentarán posibles futuras líneas de investigación que podrían complementar el trabajo realizado.

5.1. Conclusiones

Se puede concluir que la metodología propuesta, a través de la aplicación de las pruebas de fidelidad y robustez, es efectiva para evaluar las explicaciones generadas para un problema de clasificación. Esta metodología permite discernir entre explicaciones que pueden fomentar la explicabilidad del modelo y aquellas que no, como se ha observado en los diferentes casos prácticos expuestos. Además, propone considerar la multiplicidad predictiva como un ángulo adicional para evaluar la robustez, abordando un problema común que suele ser aislado en este tipo de evaluaciones. Ante la posibilidad de generar explicaciones mediante diferentes técnicas y modelos, esta metodología también facilita la selección de las explicaciones más adecuadas para el problema en cuestión.

De esta manera, simplifica la tarea evaluación y validación de la explicabilidad de los SIA, fomenta su aceptación en sectores críticos como la sanidad, la justicia y la movilidad, donde la confianza y la transparencia son fundamentales, y supone una iniciativa realista para acometer las recientes políticas y normativas, ayudando a conformar un marco regulador más robusto y adaptado a las necesidades actuales. Esta, a su vez, se podría integrar en metodologías más grandes que gestionen el ciclo de vida completo de los SIA, como es el caso de la basada en «quality gates» [84]. Todo ello, facilita la innovación responsable y el desarrollo sostenible en la economía digital, de acuerdo con las iniciativas de la Comisión Europea.

5.1.1. Objetivos de Desarrollo Sostenible

Además, este proyecto se alinea estrechamente con los Objetivos de Desarrollo Sostenible (ODS) establecidos por la Organización de las Naciones Unidas. Al promover la transparencia y la confianza de los SIA, facilita su desarrollo y contribuye así a varios objetivos centrados en la creación de infraestructuras y sistemas más seguros y

sostenibles. En concreto, está directamente relacionado con el ODS 9, que promueve la innovación y la infraestructura sostenible. Asimismo, al fomentar un uso ético y seguro de la IA, se apoya el ODS 16, que busca promover sociedades justas, pacíficas e inclusivas. La capacidad de los SIA para tomar decisiones explicables y justificables puede mejorar la eficiencia y la responsabilidad en áreas como la sanidad (ODS 3), la justicia (ODS 16), y la educación (ODS 4). Además, al asegurar que los SIA sean desarrollados y utilizados de manera ética, este proyecto también promueve la igualdad de género (ODS 5) y la reducción de las desigualdades (ODS 10), al mitigar los sesgos y discriminaciones que pueden surgir de los algoritmos de IA [85, 86].

5.2. Trabajo futuro

Una primera oportunidad para extender y mejorar este trabajo sería aplicar la metodología propuesta a una mayor variedad de conjuntos de datos, lo que permitiría refinarla y utilizarla para extraer conclusiones más precisas. Por ejemplo, este enfoque podría ayudar a establecer umbrales de decisión informados en las pruebas de robustez, determinando si las explicaciones son realmente robustas en diferentes escenarios, en comparación con lo que se espera o podría ser habitual.

Además, aunque las explicaciones basadas en la importancia de características son uno de los métodos más utilizados para mejorar la explicabilidad de los modelos de aprendizaje automático, no son los únicos. Otros métodos de IAX post-hoc, como aquellos que generan ejemplos o reglas, también se beneficiarían de una metodología similar para ser validados. Conceptualmente, la metodología propuesta podría reutilizarse para validar este tipo de explicaciones, ya que las métricas de fidelidad y robustez son relevantes para diferentes tipos de explicaciones y se pueden definir según la morfología de cada una. No obstante, otra línea de trabajo futuro prodría centrarse en adaptar las pruebas utilizadas para su aplicación a otros métodos de IAX post-hoc o, idealmente, desarrollar una metodología que pudiera aplicarse universalmente a todos los métodos de explicabilidad.

Finalmente, el futuro de la IAX se beneficiaría del desarrollo de nuevos modelos intrínsecamente interpretables, que permitan a los humanos comprender sus decisiones y resultados directamente, sin necesidad de técnicas auxiliares. Aunque esta iniciativa representa un desafío considerable, debido a la mencionada relación de compromiso entre explicabilidad y precisión de los modelos, es la forma más eficiente de abordar el problema de la explicabilidad y cumplir con la normativa de la Comisión Europea. Es por ello, que la principal línea de trabajo futuro en este campo es aquella centrada en el desarrollo y validación de estos nuevos modelos, posiblemente reutilizando aspectos de esta metodología.

Bibliografía

[1] Rich Caruana et al. «Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission». En: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, págs. 1721-1730. DOI: 10.1145/2783258.2788613.

[2] «State v. Loomis». En: *Harvard Law Review* 130.5 (mar. de 2017). URL: https://harvardlawreview.org/print/vol-130/state-v-loomis.

[3] Ryan Randazzo. *What went wrong with Uber's Volvo in fatal crash? Experts shocked by technology failure*. Recuperado Mayo 28, 2023. Mar. de 2018. URL: https://www.azcentral.com/story/money/business/tech/2018/03/22/what-went-wrong-ubers-volvo-fatal-crash-experts-shocked-technology-failure/448487002/.

[4] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending certain Union legislative acts*. Inf. téc. Document 52021PC0206, 2021/16, Bruselas: European Commission, Abril de 2021.

[5] Amina Adadi y Mohammed Berrada. «Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)». En: *IEEE Access* 6 (2018), págs. 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

[6] Raquel González-Alday et al. «A Scoping Review on the Progress, Applicability, and Future of Explainable Artificial Intelligence in Medicine». En: *Applied Sciences* 13.19 (2023), pág. 10778. DOI: 10.3390/app131910778.

[7] David Alvarez-Melis y Tommi S. Jaakkola. «On the robustness of interpretability methods». En: *arXiv preprint arXiv:1806.08049* (2018). DOI: 10.48550/arXiv.1806.08049.

[8] Denise M. Rousseau et al. «Not so different after all: A cross-discipline view of trust». En: *The Academy of Management Review* 23 (jul. de 1998), pág. 395.

[9] Erik Erikson. *Childhood and Society*. New York, NY: Norton & Company, Inc., 1950, pág. 221.

[10] Erico Tjoa y Cuntai Guan. «A survey on explainable artificial intelligence (xai): Toward medical xai». En: *IEEE Transactions on Neural Networks and Learning Systems*. 2020, págs. 1-21. DOI: 10.1109/TNNLS.2020.3027314.

[11] Jennifer Weeks. *Bridging the Human-AI Gap: The Indispensable Role of Behavioral Science*. Recuperado Febrero 6, 2024. Diciembre de 2023. URL: https://beworks.com/blog/bridging-the-human-ai-gap-the-indispensable-role-of-behavioral-science/.

[12] Gaia Molinaro. *When AI meets behavioural science*. Recuperado Febrero 6, 2024. Enero de 2022. URL: https://medium.com/behaviouralucl/when-ai-meets-behavioural-science-54703fb61d3.

[13] Evan Nesterak. *Imagining the next decade of behavioral science*. Recuperado Febrero 6, 2024. Enero de 2020. URL: https://behavioralscientist.org/imagining-the-next-decade-future-of-behavioral-science/.

[14] Thomas Rojat et al. «Explainable artificial intelligence (xai) on timeseries data: A survey». En: (2021). DOI: 10.48550/arXiv.2104.00950.

[15] Zachary C. Lipton. «The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery». En: *Queue* 16.3 (2018), págs. 31-57. DOI: 10.1145/3236386.3241340.

[16] Sajid Ali et al. «Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence». En: *Information Fusion* 99 (2023), pág. 101805. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805.

[17] Andrew Saxe, Stephanie Nelli y Christopher Summerfield. «If deep learning is the answer, what is the question?» En: *Nature Reviews Neuroscience* 22 (2021), págs. 55-67. DOI: 10.1038/s41583-020-00395-8.

[18] Zewen Li et al. «A survey of convolutional neural networks: analysis, applications, and prospects». En: *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 33. 12. 2022, págs. 6999-7019. DOI: 10.1109/TNNLS.2021.3084827.

[19] Iqbal H. Sarker. «Machine learning: Algorithms, real-world applications and research directions». En: *SN Computer Science* 2 (2021), págs. 1-21. DOI: 10.1007/s42979-021-00592-x.

[20] Wojciech Samek y Klaus-Robert Müller. «Towards Explainable Artificial Intelligence». En: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019, págs. 5-22. DOI: 10.1007/978-3-030-28954-6_1.

[21] Kacper Sokol y Peter Flach. «Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches». En: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020)*. New York, NY, USA: Association for Computing Machinery, 2020, págs. 56-67. DOI: 10.1145/3351095.3372870.

[22] Timo Speith. «A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods». En: jun. de 2022, págs. 2239-2250. DOI: 10.1145/3531146.3534639.

[23] Pantelis Linardatos, Vasilis Papastefanopoulos y Sotiris Kotsiantis. «Explainable AI: A review of machine learning interpretability methods». En: *Entropy* 23.1 (2021), pág. 18. DOI: 10.3390/e23010018.

[24] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2.ª ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[25] Been Kim et al. «Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)». En: *International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR. 2018, págs. 2668-2677. DOI: https://doi.org/10.48550/arXiv.1711.11279.

[26] Anh Nguyen et al. «Synthesizing the preferred inputs for neurons in neural networks via deep generator networks». En: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, págs. 3387-3395. DOI: 10.48550/arXiv.1605.09304.

[27] Jing Lei et al. «Distribution-free Predictive Inference for Regression». En: *Journal of the American Statistical Association* 113.523 (2018), págs. 1094-1111. DOI: 10.48550/arXiv.1604.04173.

[28] Aaron Fisher y Cynthia Rudin y Francesca Dominici. «All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously». En: *Journal of Machine Learning Research* 20.177 (2019), págs. 1-81. DOI: 10.48550/arXiv.1801.01489.

[29] Krystian Safjan. *LIME - Understanding How This Method for Explainable AI Works*. 2023. URL: https://safjan.com/how-the-lime-method-for-explainable-ai-works/.

[30] Alvin E. Roth, ed. *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988. DOI: 10.1017/CBO9780511528446.

[31] Scott M. Lundberg y Su-In Lee. «A unified approach to interpreting model predictions». En: *Advances in Neural Information Processing Systems*. Vol. 30. 2017. DOI: 10.48550/arXiv.1705.07874.

[32] Karen Simonyan, Andrea Vedaldi y Andrew Zisserman. «Deep inside convolutional networks: Visualising image classification models and saliency maps». En: *arXiv preprint arXiv:1312.6034* (2013). DOI: 10.48550/arXiv.1312.6034.

[33] Ramprasaath R. Selvaraju et al. «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. DOI: 10.1007/s11263-019-01228-7.

[34] Daniel Smilkov et al. *Smoothgrad: removing noise by adding noise*. 2017. DOI: 10.48550/arXiv.1706.03825.

[35] Grégoire Montavon et al. «Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition». En: *Pattern Recognition* 65 (2017), págs. 211-222. DOI: 10.1016/j.patcog.2016.11.008.

[36] Sebastian Bach et al. «On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation». En: *PLOS ONE* 10.7 (2015). DOI: 10.1371/journal.pone.0130140.

[37] Mukund Sundararajan, Ankur Taly y Qiqi Yan. «Axiomatic Attribution for Deep Networks». En: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. 2017, págs. 3319-3328. DOI: 10.48550/arXiv.1703.01365.

[38] Avanti Shrikumar, Peyton Greenside y Anshul Kundaje. «Learning Important Features Through Propagating Activation Differences». En: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. 2017, págs. 3145-3153. DOI: 10.48550/arXiv.1704.02685.

[39] Tanisha. *Summary of "A unified approach to interpreting model predictions"*. Jun. de 2022. URL: https://medium.com/@tanisha21180/summary-of-a-unified-approach-to-interpreting-model-predictions-eb9e059a9a2a.

[40] Salih Sarp et al. «The enlightening role of explainable artificial intelligence in chronic wound classification». En: *Electronics* 10.12 (2021), pág. 1406. DOI: 10.3390/electronics10121406.

[41] Zizhao Zhang et al. «Pathologist-level interpretable whole-slide cancer diagnosis with deep learning». En: *Nature Machine Intelligence* 1.5 (2019), págs. 236-245. DOI: 10.1038/s42256-019-0052-1.

[42] Christopher Duckworth et al. «Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19». En: *Scientific Reports* 11.1 (2021), págs. 1-10. DOI: 10.1038/s41598-021-02481-y.

[43] Anna Markella Antoniadi et al. «Prediction of caregiver quality of life in amyotrophic lateral sclerosis using explainable machine learning». En: *Scientific Reports* 11.1 (2021), págs. 1-13. DOI: 10.1038/s41598-021-91632-2.

[44] Xian Zeng et al. «Explainable machine-learning predictions for complications after pediatric congenital heart surgery». En: *Scientific Reports* 11.1 (2021), págs. 1-11. DOI: 10.1038/s41598-021-96721-w.

[45] Majid Farhadloo et al. «Samcnet: towards a spatially explainable AI approach for classifying MXIF oncology data». En: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022, págs. 2860-2870. DOI: 10.1145/3534678.3539168.

[46] Zhengyu Jiang et al. «An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission». En: *Computers in Biology and Medicine* 204 (2021), pág. 106040. DOI: 10.1016/j.cmpb.2021.106040.

[47] Dipankar Dasgupta, Zahid Akhtar y Sajib Sen. «Machine learning in cybersecurity: a comprehensive survey». En: *Journal of Defense Modeling and Simulation* 19.1 (2022), págs. 57-106. DOI: 10.1177/1548512920951275.

[48] Berk Gulmezoglu. «XAI-based microarchitectural side-channel analysis for website fingerprinting attacks and defenses». En: *IEEE Transactions on Dependable and Secure Computing*. Vol. 20. 2021, pág. 10. DOI: 10.1109/TDSC.2021.3117145.

[49] Johannes Feichtner y Stefan Gruber. «Understanding privacy awareness in Android app descriptions using deep learning». En: *CODASPY '20: Tenth ACM Conference on Data and Application Security and Privacy*. 2020, págs. 203-214. DOI: 10.1145/3374664.3375730.

[50] Wenbo Guo et al. «Lemna: explaining deep learning based security applications». En: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, págs. 364-379. DOI: 10.1145/3243734.3243792.

[51] Salvatore Carta et al. «Explainable AI for financial forecasting». En: *International Conference on Machine Learning, Optimization, and Data Science*. 2021, págs. 51-69. DOI: 10.1007/978-3-030-95470-3_5.

[52] M. Chromik et al. «I think I get your point, AI! The illusion of explanatory depth in explainable AI». En: *26th International Conference on Intelligent User Interfaces*. 2021, págs. 307-317.

[53] Abhishek Agarwal et al. «Machine learning based explainable financial forecasting». En: *2022 4th International Conference on Computer Communication and the Internet (ICCCI)*. IEEE, 2022, págs. 34-38. DOI: 10.1109/ICCCI55554.2022.9850272.

[54] Richard A. Berk y Justin Bleich. «Statistical procedures for forecasting criminal behavior: a comparative assessment». En: *Criminology & Public Policy* 12 (2013), pág. 513. DOI: 10.1111/1745-9133.12047.

[55] Dina Mardaoui y Damien Garreau. «An analysis of LIME for text data». En: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, págs. 3493-3501. DOI: 10.48550/arXiv.2010.12487.

[56] Selin Akgun y Christine Greenhow. «Artificial intelligence in education: addressing ethical challenges in K-12 settings». En: *AI Ethics* 20 (2021), págs. 1-10. DOI: 10.1007/s43681-021-00096-7.

[57] Elvis Melo et al. «On the use of explainable artificial intelligence to evaluate school dropout». En: *Education Sciences* 12.12 (2022), pág. 845. DOI: 10.3390/educsci12120845.

[58] Gloria Milena Fernandez-Nieto et al. «Storytelling with learner data: guiding student reflection on multimodal team data». En: *IEEE Transactions on Learning Technologies* 14.5 (2021), págs. 695-708. DOI: 0.1109/TLT.2021.3131842.

[59] Piotr Dabkowski y Yarin Gal. «Real time image saliency for black box classifiers». En: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., 2017, págs. 6970-6979. DOI: 10.48550/arXiv.1705.07857.

[60] Arize AI. *Normalized discounted cumulative gain (NDCG): A pragmatic guide*. Recuperado Enero 11, 2024. Enero de 2023. URL: https://arize.com/blog-course/ndcg/.

[61] Julius Adebayo et al. «Sanity checks for saliency maps». En: *Advances in Neural Information Processing Systems*. Vol. 31. 2018. DOI: 10.48550/arXiv.1810.03292.

[62] Md Abdul Kadir, Amir Mosavi y Daniel Sonntag. «Assessing XAI: Unveiling Evaluation Metrics for Local Explanation, Taxonomies, Key Concepts, and Practical Applications». En: *arXiv preprint arXiv:2301.2989* (2023). DOI: 10.31224/2989.

[63] Sara Hooker et al. «A Benchmark for Interpretability Methods in Deep Neural Networks». En: *Advances in Neural Information Processing Systems*. Vol. 32. 2019. DOI: 10.48550/arXiv.1806.10758.

[64] Chih-Kuan Yeh et al. «On the (in)fidelity and sensitivity for explanations». En: *Advances in Neural Information Processing Systems*. Vol. 32. 2019. DOI: 10.48550/arXiv.1901.09392.

[65] Jianlong Zhou et al. «Evaluating the quality of machine learning explanations: A survey on methods and metrics». En: *Electronics* 10.5 (2021), pág. 593. DOI: 10.3390/electronics10050593.

[66] Janet Hui-wen Hsiao et al. «Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI)». En: *arXiv preprint arXiv:2108.01737* (2021). DOI: 10.48550/arXiv.2108.01737.

[67] Gareth James et al. *An Introduction to Statistical Learning: With Applications in Python*. Springer, 2023.

[68] W. James Murdoch et al. «Definitions, methods, and applications in interpretable machine learning». En: *Proceedings of the National Academy of Sciences of the United States of America* 116 (44 2019). DOI: 10.1073/pnas.1900653116.

[69] Finale Doshi-Velez y Been Kim. «Considerations for Evaluation and Generalization in Interpretable Machine Learning». En: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018. DOI: 10.1007/978-3-319-98131-4_1.

[70] Yang Liu et al. «Synthetic Benchmarks for Scientific Research in Explainable Machine Learning». En: *arXiv preprint arXiv:2106.12543* (2021). DOI: 10.48550/arXiv.2106.12543.

[71] Wojciech Samek et al. «Evaluating the visualization of what a deep neural network has learned». En: *IEEE Transactions on Neural Networks and Learning Systems*. Vol. 28. 11. 2017. DOI: 10.1109/TNNLS.2016.2599820.

[72] Philine Bommer et al. «Finding the right XAI method – A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science». En: *Artificial Intelligence for the Earth Systems* 3.3 (2024).

[73] Weina Jin et al. «Guidelines and evaluation of clinical explainable AI in medical image analysis». En: *Medical Image Analysis* 84 (feb. de 2023), pág. 102684. DOI: 10.1016/j.media.2022.102684.

[74] Corinna Cortes y Vladimir Vapnik. «Support-vector networks». En: *Machine Learning* 20 (1995), págs. 273-297. DOI: 10.1007/BF00994018.

[75] S. S. Shapiro y M. B. Wilk. «An Analysis of Variance Test for Normality (Complete Samples)». En: *Biometrika* 52 (Diciembre de 1965), págs. 591-611. DOI: 10.1093/biomet/52.3-4.591.

[76] Zhenqiu Laura Lu y Ke-Hai Yuan. «Welch's t test». En: *Encyclopedia of Research Design* (Enero de 2010), págs. 1620-1623. DOI: 10.13140/RG.2.1.3057.9607.

[77] Nadim Nachar. «The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution». En: *Tutorials in Quantitative Methods for Psychology* 4 (mar. de 2008). DOI: 10.20982/tqmp.04.1.p013.

[78] Charles T. Marx, Flavio du Pin Calmon y Berk Ustun. «Predictive multiplicity in classification». En: *International Conference on Machine Learning*. Nov. de 2020, págs. 6765-6774. DOI: 10.48550/arXiv.1909.06677.

[79] Leo Breiman. «Statistical modeling: The two cultures». En: *Statistical science* 16.3 (2001), págs. 199-231. DOI: 10.1214/ss/1009213726.

[80] Saumitra Mishra et al. «A Survey on the Robustness of Feature Importance and Counterfactual Explanations». En: *arXiv preprint arXiv:2111.00358* (2021). DOI: 10.48550/arXiv.2111.00358.

[81] Meritxell Deulofeu et al. «Detection of SARS-CoV-2 infection in human nasopharyngeal samples by combining MALDI-TOF MS and artificial intelligence». En: *Front Med (Lausanne)* 8 (Abril de 2021). PMID: 33869258; PMCID: PMC8047105, pág. 661358. DOI: 10.3389/fmed.2021.661358.

[82] UCI Machine Learning Repository. *Adult*. Recuperado Febrero 6, 2024. URL: https://archive.ics.uci.edu/dataset/2/adult.

[83] Alba María Lopez González y Esteban García-Cuesta. «On the transferability of local model-agnostic explanations of machine learning models to unseen data». En: *IEEE International Conference on Evolving and Adaptive Intelligent Systems (IEEE EAIS 2024)*. 2024.

[84] Miriam Elia y Bernhard Bauer. «A Methodology Based on Quality Gates for Certifiable AI in Medicine: Towards a Reliable Application of Metrics in Machine Learning». En: *18th International Conference on Software Technologies*. Jul. de 2023, págs. 486-493. DOI: 10.5220/0012121300003538.

[85] United Nations. *Objetivos y Metas de Desarrollo Sostenible - Desarrollo Sostenible*. Recuperado Junio 22, 2024. URL: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/.

[86] *Objetivos - agenda2030*. Recuperado Junio 22, 2024. URL: https://www.agenda2030.gob.es/objetivos/home.htm.

Apéndice A

Anexo

A.1. Valores de NDCG entre las explicaciones de diferentes modelos para todos los subconjuntos

Cuadro A.1: NDCG entre las explicaciones SHAP de los modelos SVM y XGBoost, SVM y MLP, y MLP y XGBoost para los diferentes subconjuntos (splits) de *covid19*.

modelo/ NDCG de split	0	1	2	3	4	5	6	7	8	9	**medio**
SVM XGBoost	0.7109	0.6098	0.7039	0.5847	0.6521	0.7349	0.6699	0.6389	0.6637	0.6637	**0.6563**
SVM MLP	0.8831	0.9045	0.9411	0.9458	0.8908	0.9209	0.8612	0.9214	0.8962	0.9427	**0.9108**
MLP XGBoost	0.7148	0.7327	0.7148	0.6467	0.6668	0.7264	0.7290	0.7144	0.6874	0.6338	**0.6967**

Cuadro A.2: NDCG entre las explicaciones LIME de los modelos SVM y XGBoost, SVM y MLP, y MLP y XGBoost para los diferentes subconjuntos (splits) de *covid19*.

modelo/ NDCG de split	0	1	2	3	4	5	6	7	8	9	**medio**
SVM y XGBoost	0.7091	0.7365	0.7856	0.7968	0.7432	0.8471	0.8413	0.7768	0.8584	0.8589	**0.7954**
SVM y MLP	0.7787	0.8619	0.8848	0.8656	0.8845	0.9377	0.9198	0.8971	0.8795	0.8968	**0.8807**
MLP y XGBoost	0.7708	0.7752	0.7961	0.7943	0.6806	0.7478	0.8438	0.7560	0.8590	0.8444	**0.7868**

Cuadro A.3: NDCG entre las explicaciones SHAP de los modelos SVM y XGBoost, SVM y MLP, y MLP y XGBoost para los diferentes subconjuntos (splits) de *census-income*.

modelo/ NDCG de split	0	1	2	3	4	5	6	7	8	9	**medio**
SVM y XGBoost	0.7037	0.7928	0.6560	0.7809	0.8459	0.7174	0.6749	0.7338	0.6253	0.6703	**0.7914**
SVM y MLP	0.8099	0.9641	0.7662	0.9087	0.8879	0.8024	0.7564	0.7848	0.8088	0.9097	**0.8399**
MLP y XGBoost	0.7547	0.7476	0.9423	0.7767	0.7935	0.9087	0.7313	0.8335	0.7224	0.7032	**0.7201**

Cuadro A.4: NDCG entre las explicaciones LIME de los modelos SVM y XGBoost, SVM y MLP, y MLP y XGBoost para los diferentes subconjuntos (splits) de *census-income*.

modelo/ NDCG de split	0	1	2	3	4	5	6	7	8	9	**medio**
SVM y XGBoost	0.9975	0.9982	0.9729	0.9966	0.9982	0.9985	0.9959	0.9965	0.9990	0.9762	**0.9929**
SVM y MLP	0.9970	0.9984	0.9996	0.9975	0.9983	0.9986	0.9984	0.9987	0.9997	0.9988	**0.9985**
MLP y XGBoost	0.9967	0.9992	0.9809	0.9923	0.9955	0.9986	0.9968	0.9987	0.9972	0.9949	**0.9951**

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Fri Jun 28 17:18:48 CEST 2024
	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)