**Universidad Internacional de La Rioja (UNIR)**

**ESIT**

**Máster Universitario en Inteligencia Artificial**

# Comparativa de modelos de Machine Learning interpretables para la predicción de riesgo crediticio

**Trabajo Fin de Máster**

**Presentado por:** Ortega Martín, Sonia

**Director/a:** Ferragut Amengual, Antoni Manel

Ciudad: Madrid
Fecha: 22/07/2021

# Resumen

Los sistemas de Inteligencia Artificial son ampliamente utilizados e impactan en la vida diaria de las personas de forma creciente, en gran medida debido al avance en Machine Learning y su alta precisión. Los modelos suelen ser opacos y dificultan el entendimiento de su lógica. Su interpretabilidad se hace más necesaria, sobre todo en dominios de alto riesgo. Este trabajo se ha centrado en el dominio financiero, en la tarea predicción de riesgo crediticio, sobre el conjunto de datos de solicitudes de crédito con garantía hipotecaria (HELOC). Se han planteado distintos modelos y aplicado técnicas para obtener modelos interpretables. Se han definido métricas de interpretabilidad, que permiten la comparativa de modelos atendiendo a criterios de interpretabilidad y no únicamente de precisión. Los resultados obtenidos muestran que la elección de un modelo no solo debe estar basado en la precisión, sino que debe mantener un equilibrio entre precisión e interpretabilidad.

**Palabras Clave:** Inteligencia Artificial, Machine Learning, Interpretabilidad, Métricas de Interpretabilidad, Riesgo crediticio.

# Abstract

Artificial Intelligence systems are widely used and are increasingly impacting people's daily lives, largely due to advances in Machine Learning and their high accuracy. Models are often opaque and make it difficult to understand their logic. Their interpretability becomes more necessary, especially in high-risk domains. This work has focused on the financial domain, in the task of credit risk prediction, on the data set of home equity loan applications (HELOC). Different models have been proposed and techniques have been applied to obtain interpretable models. Interpretability metrics have been defined, which allow the comparison of models based on interpretability criteria and not only on accuracy. The results obtained show that the choice of a model should not only be based on accuracy, but should also maintain a balance between accuracy and interpretability.

**Keywords:** Artificial Intelligence, Machine Learning, Interpretability, Interpretability metrics, Credit risk.

# Índice de contenidos

# Índice de tablas

# Índice de figuras

# 1. Introducción

En esta sección se da una visión de la importancia que está alcanzando la interpretabilidad en los modelos de Inteligencia Artificial (IA) y los motivos por los que es cada vez más necesaria, en concreto para el dominio de alto riesgo como es el financiero. Cuando se utilizan modelos para predecir, por ejemplo, si una persona debe recibir un préstamo, es importante verificar que el modelo cumple con las normas éticas, garantizando que el modelo sea justo. El objetivo de este trabajo será realizar una comparativa de modelos, entrenados para la tarea de predicción del riesgo crediticio, atendiendo no solo a la precisión sino a la interpretabilidad de estos. Para realizarlo se parte de un conjunto de datos sobre las solicitudes de línea de crédito con garantía hipotecaria (HELOC) y se seleccionan distintos modelos que se ajustarán para obtener la mayor precisión. Los modelos elegidos serán tanto interpretables como no interpretables o de caja negra. A estos últimos será necesario aplicar técnicas que permitan obtener modelos interpretables. Para comparar los modelos se utilizarán las métricas de precisión existentes y se aportarán métricas de interpretabilidad que permitan comparar los modelos atendiendo no solo a criterios de precisión, sino también de interpretabilidad. Como resultado se ha podido comprobar que la elección del mejor modelo no siempre se debe basar en la precisión, sino que, en ciertos ámbitos es preferible renunciar a cierta precisión para poder mejorar en la interpretación del modelo.

## 1.1 Motivación

En los últimos años los avances en Machine Learning (ML) y, en particular, el avance en Deep Learning (DL), han propiciado el uso de modelos en distintas áreas, debido a su alta predicción. Cada vez es más usual que se utilicen en tareas cotidianas como reconocimiento de objetos en imágenes, transcripción de voz a texto, traducción entre idiomas, etc.

El aumento del uso de modelos de ML implica que cada vez es mayor el número de personas afectadas por la implantación de estos sistemas, haciéndose esencial la interpretabilidad sobre todo en dominios de alto riesgo, donde el coste de realizar una predicción incorrecta sea muy alto. En dominios de aplicación como transporte, seguridad, medicina, justicia penal o finanzas, los enfoques de Inteligencia Artificial Explicable (XAI) tienen alto potencial. No solo es necesario que el modelo tenga un alto nivel de predicción, también es importante saber los motivos por los que el modelo llegó a dicha predicción, proporcionando información que permita confiar en el resultado obtenido. Debe existir una compensación entre precisión e interpretabilidad.

Este trabajo se centra en el dominio financiero y en concreto en la predicción del riesgo crediticio. Se estudiarán y aplicarán varias técnicas que permitan obtener un modelo interpretable, cuya información sustente el resultado predictivo del mismo. Los modelos obtenidos se compararán en función de la precisión e interpretación.

Como primer paso, para intentar comprender el significado de interpretabilidad, se hace un resumen de las definiciones dadas por algunos autores. Debido al carácter subjetivo del término, no hay una definición estándar y globalmente aceptada para la interpretabilidad y los términos interpretable y explicable se utilizan a veces indistintamente, aunque ciertos autores los consideran diferentes.

Según (Doshi-Velez & Kim, 2017) a través de la interpretabilidad del sistema se puede explicar su razonamiento de forma comprensible al ser humano y así se podría verificar si es sólido. En la misma línea (Adadi & Berrada, 2018) considera que "La explicabilidad está estrechamente relacionada con el concepto de interpretabilidad: los sistemas interpretables son explicables si sus operaciones pueden ser entendidas por el ser humano."

Para (Gilpin et al., 2018) las explicaciones deben ser completas e interpretables. La completitud permite describir el comportamiento del sistema de forma precisa en el mayor número de situaciones. La explicación es interpretable si describe el sistema de forma que los humanos lo puedan entender, con descripciones sencillas y vocabulario acorde con el interlocutor. Según (Zhou et al., 2021) la evaluación de la calidad de las explicaciones tiene como objetivo evaluar hasta qué punto se satisfacen dos características, fidelidad e interpretabilidad. Además, considera que la explicación es un término inherentemente subjetivo, que su calidad está sujeta al contexto: los usuarios, la explicación en sí misma y el tipo de información que les interesa a los usuarios.

(Linardatos et al., 2021) asocia la interpretabilidad a la capacidad de identificar la relación existente entre las entradas y salidas del sistema, mientras que la explicabilidad se asocia a la lógica interna del sistema de aprendizaje automático, concluyendo que la interpretabilidad y explicabilidad no se implican mutuamente.

En el contexto de aprendizaje automático (Murdoch et al., 2019) considera la interpretación como la extracción de conocimiento del dominio a través de un modelo que ha aprendido las relaciones contenidas en los datos. El aprendizaje automático interpretable proporciona un conocimiento relevante sobre el problema del dominio en cuestión.

Según (European Commission. Joint Research Centre., 2020), con la interpretabilidad se puede comprender el mecanismo interno del sistema a la vez que se demuestra si se ajusta

a las especificaciones y cumple con las normas éticas. Esto estaría alineado con los cuatro factores que (Adadi & Berrada, 2018) considera para la necesidad de la explicabilidad, mostrados en la Figura 1:

- Explicar para justificar los resultados y decisiones, garantizando que no se tomen por error y que cumplen con la legislación.
- Explicar para controlar, pudiendo detectar vulnerabilidades que permitan corregir rápidamente los errores.
- Explicar para mejorar, realizando una mejora continua de los modelos.
- Explicar para descubrir, aprender nuevos hechos y adquirir conocimiento.

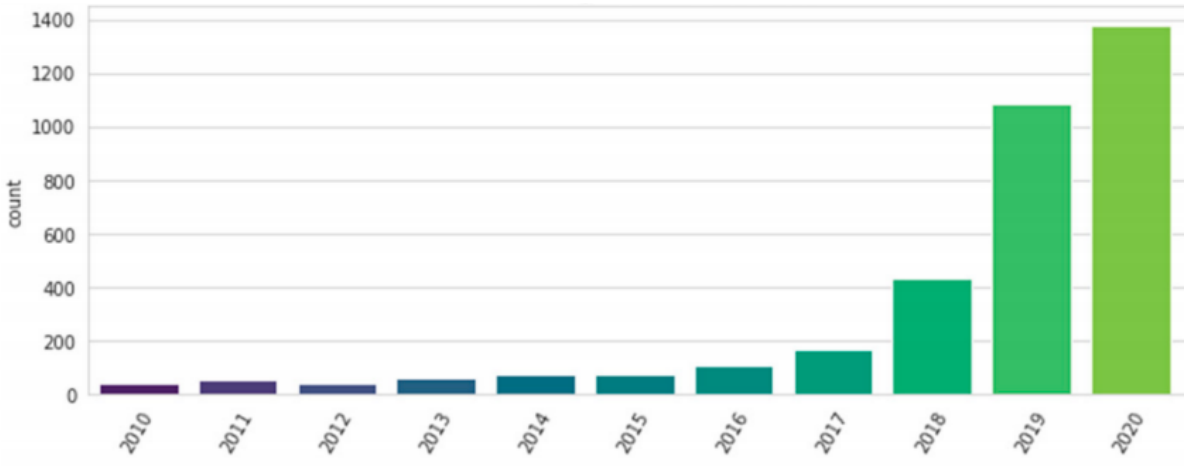

Fuente: Elaboración propia a partir de (Adadi & Berrada, 2018).

Figura 1. Factores para la necesidad de la Explicabilidad

La necesidad de comprender el comportamiento interno de los sistemas no es solo por seguridad sino por razones legales. La aplicación del Reglamento General de Protección de Datos (GDPR) (Regulation (EU) 2016/679, 2016) implica que los sistemas de Inteligencia Artificial deben adaptarse para cumplir con ciertos requisitos, incluido también el de explicación. Los artículos 13 (apartado 2 f) y 14 (apartado 2g) indican que, en los casos de existencia de decisiones automatizadas referidas en el artículo 22, se deberá facilitar información significativa sobre la lógica aplicada y sus consecuencias.

Se considera que la interpretabilidad permite comprender la relación existente entre las entradas y salidas presentándose de forma sencilla y entendible por el humano y así comprobar la equidad de los modelos e identificar y corregir posibles sesgos. Con la interpretabilidad se fomenta la confianza, una mayor credibilidad en el modelo y el cumplimiento de la legislación, que puede llevar a la elección de los modelos interpretables frente a los que no lo son, sobre todo en aquellos dominios de alto riesgo.

Todo lo expuesto anteriormente permite tener una idea de la importancia que ha tomado la interpretabilidad en los sistemas de Inteligencia Artificial. Eso ha llevado a un creciente número de publicaciones de investigación sobre interpretación de modelos de ML, como se muestra en la Figura 2, donde se puede ver el aumento de publicaciones experimentado desde 2018 hasta 2020.




Fuente: Adaptación a partir de (Zhou et al., 2021). Información basada en Scopus.com hasta diciembre de 2020.

Figura 2. Número de publicaciones de investigación sobre explicaciones de ML

Debido al creciente número de técnicas y métodos, a lo largo de la literatura se han realizado clasificaciones para ayudar a la elección de los más adecuados para la interpretabilidad del problema concreto, como se explicará en la sección 2. La elección del enfoque de interpretabilidad deberá tener en cuenta características esenciales como la naturaleza del problema, el contexto y la audiencia según (Murdoch et al., 2019). Las expectativas y deseos de las partes interesadas influyen en la evaluación de la calidad de las explicaciones. Es difícil medir y cuantificar en qué medida la explicación logra el objetivo, debido a que no existen métricas acordadas para la calidad de métodos de explicación. (Nguyen & Martínez, 2020) y (Zhou et al., 2021) coinciden en que no es posible definir una implementación de métricas de evaluación que se puedan aplicar a todos los métodos de explicación.

## 1.2 Planteamiento del trabajo

Los modelos de riesgo crediticio permiten a una entidad financiera predecir si un prestatario pudiera incumplir el reingreso del préstamo, lo que podría suponer pérdidas para la entidad, en función de la probabilidad de incumplimiento.

Los modelos estiman si se produce o no el incumplimiento a partir de características asociadas a cada prestatario y en función de todas aquellas relaciones aprendidas de los datos históricos relacionados.

Como se ha expuesto anteriormente, en dominios como el financiero, el objetivo de los modelos no debe ser exclusivamente la optimización de la precisión, también deben equilibrar esta precisión con la interpretabilidad.

Las métricas para determinar la precisión predictiva están ampliamente estudiadas y consensuadas. Los modelos que estiman si se produce o no el incumplimiento podrán evaluarse utilizando métodos como área bajo la curva ROC (AUROC).

En la literatura no se encuentra el mismo consenso sobre las métricas que permitan determinar la interpretabilidad de un modelo. Este trabajo se centra en el proceso expuesto por (Doshi-Velez & Kim, 2017) para definir y evaluar la interpretabilidad.

Como punto de partida se definen los principios generales sobre los que se basará la evaluación de la interpretabilidad del modelo. Para ello se da respuesta a tres puntos planteados por el autor:

- Necesidad de la interpretabilidad debida a la incompletitud de la formulación del problema.

  Considera que no en todos los casos es necesaria una interpretabilidad del modelo, argumentando que la necesidad de modelos interpretables está relacionada con la incompletitud a la hora de definir el problema.

  En este trabajo se pretenden modelos que sean justos en la decisión de conceder o no el crédito. Debido a la incompletitud de la definición, se necesita interpretabilidad que permita evaluar si el modelo no discrimina y sigue unas bases éticas.

- Nivel al que se realiza la evaluación de la interpretación.

  En su artículo, se presentan tres tipos de enfoques para la evaluación: basada en aplicaciones, basada en humanos y basada en la funcionalidad.

  Los dos primeros tipos de evaluaciones requieren experimentos humanos. La evaluación basada en aplicaciones se realizaría con expertos en el dominio en cuestión, para evaluar la calidad de la explicación en el contexto de la aplicación. La evaluación basada en humanos se realiza con humanos legos, para probar la calidad de la explicación respecto a nociones más generales.

  La evaluación basada en la funcionalidad utiliza una definición formal de interpretabilidad para medir la calidad de la explicación, sin requerir de experimentos con humanos.

Este trabajo se centra en el nivel de evaluación basado en la funcionalidad, que permita la comparación entre modelos y la elección de aquel que permita una mayor interpretabilidad y precisión para la tarea en cuestión.

- Factores relevantes para la interpretación.

  Estos factores son las dimensiones que permiten evaluar el rendimiento del modelo respecto a la tarea de aplicación real. Podrían ser factores relacionados con la tarea o bien relacionados con el método.

  Para este trabajo se utilizan los siguientes factores para la evaluación de los modelos:

  - Interpretabilidad local, que permita una justificación para una decisión específica.
  - Unidades básicas de explicación que se utilizan, identificando:
    - Cuántas unidades básicas contiene la explicación.
    - Si las unidades básicas son características en bruto o derivadas.
    - Cuál es la estructura de dichas unidades, si se mantiene una estructura jerárquica entre ellas.
    - Qué interacciones existen entre las unidades básicas.

En resumen, dada la tarea de predicción de riesgo crediticio, que debe realizarse de forma justa, se evalúa la interpretabilidad de los modelos que resuelven la tarea. La evaluación estará basada en la funcionalidad, que permita medir la calidad de la explicación, utilizando para ello los factores relevantes de explicabilidad local y características asociadas a las unidades básicas de explicación.

## 1.3 Estructura de la memoria

En la sección 2 se muestra el estado del arte en interpretabilidad y se hace un repaso a las clasificaciones de las técnicas de interpretabilidad más utilizados. Se revisan las aplicaciones de interpretabilidad en el dominio financiero y se explican los modelos y las técnicas de interpretabilidad local que se utilizarán en el trabajo. En la sección 3 se describen los objetivos del trabajo y la metodología utilizada. En la sección 4 se presentan los detalles del conjunto de datos utilizados, la arquitectura de los modelos, las explicaciones locales para cada uno de ellos y los criterios de medición de los modelos. En la sección 5 se realiza el desarrollo de la comparativa y los resultados obtenidos, los cuales serán discutidos y analizados en la sección 6. Finalmente, en la sección 7 se resumen los aspectos más importantes obtenidos del análisis y las líneas de trabajo futuro.

# 2. Contexto y estado del arte

A continuación, se presenta una revisión de la literatura más destacable en relación con la interpretabilidad y los enfoques existentes para la clasificación de los métodos y técnicas, las partes interesadas por la interpretabilidad, así como la forma de evaluar la interpretabilidad proporcionada por los modelos. También se analizarán estos enfoques sobre el dominio financiero, revisando los trabajos realizados en esta dirección. Se realizará un resumen de todo ello y cómo se aplicará al desarrollo de la comparativa. Finalmente se explican los detalles de los modelos, técnicas de interpretabilidad y métricas de precisión para la tarea de clasificación.

## 2.1. Interpretabilidad de los modelos

En los últimos años ha habido un aumento en la investigación sobre la interpretabilidad de los modelos, en parte debido a iniciativas como la lanzada por la Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA) de Estados Unidos (Gunning & Aha, 2019), donde se planteaban tres desafíos de investigación: producir modelos más explicables, diseñar interfaces de explicación y comprender los requisitos psicológicos para explicaciones más efectivas.

A medida que estas investigaciones aumentan también lo hacen las clasificaciones y organizaciones de las técnicas utilizadas. En este apartado, se analizan las taxonomías o clasificaciones más representativas, aunque no exhaustivas, de las técnicas para el aprendizaje automático interpretable.

Como indica (Linardatos et al., 2021), existen diferentes puntos de vista y aspectos a la hora de estudiar los métodos de interpretabilidad, por lo que la clasificación de las técnicas de interpretación de aprendizaje automático no debe ser unilateral. En la misma línea de atender a diferentes perspectivas, (Adadi & Berrada, 2018) realiza una revisión de las obras más relevantes en la literatura organizando su estudio en cuatro ejes principales: la taxonomía de los métodos de interpretabilidad, la medición y evaluación de las explicaciones, la figura del humano en el bucle y el equilibrio entre explicar y predecir.

En su gran mayoría los estudios realizados sobre interpretabilidad se centran en el primero de estos ejes, la búsqueda de métodos, procedimiento y estrategias para la explicación de los modelos y la realización de una clasificación de estos. En comparación con este eje, los otros tres mencionados son menos representativos en la literatura. A continuación, se desarrollan los tres primeros: taxonomía, la figura del humano en el bucle y medición y evaluación.

## I. Taxonomía

Los enfoques más aceptados para la clasificación de los métodos y procedimientos son los realizados en función del alcance, complejidad y dependencia del modelo.

- Alcance de las explicaciones

  Atendiendo al alcance de las explicaciones se identifican dos opciones: local o global.

  - Alcance Local

    Proporciona información sobre predicciones individuales del modelo, una explicación para la predicción de una única entrada. Muestra una porción limitada del comportamiento del sistema. Entre las formas se puede realizar a través del análisis de las características con más peso en la toma de decisión o bien utilizando explicaciones contrafácticas (Wachter et al., 2018), las cuales permiten conocer qué características son necesarias cambiar para modificar el resultado de la predicción, sin explicar el funcionamiento interno del modelo.

    (Ribeiro et al., 2016) propone LIME para una explicación local del modelo. (Lundberg & Lee, 2017) introdujeron una técnica de enfoque local con sólida base teórica llamada SHapley Additive exPlanation, SHAP.

  - Alcance Global

    Un alcance global proporciona información general del modelo. El objetivo es explicar la lógica de este que permita un conocimiento del razonamiento de todos los resultados. Se suelen utilizar para la toma de decisiones a nivel de población. La forma habitual de este enfoque se realiza a través de modelos simples e interpretables que aproximan a un modelo más complejo. La información producida por estos modelos puede requerir un conocimiento previo del dominio para su comprensión.

- Complejidad del modelo

  En relación con la complejidad del modelo utilizado para la predicción, se distinguen entre interpretable intrínseco y modelo de caja negra interpretable.

  - Interpretable intrínseco

    Los modelos interpretables intrínsecos no requieren de un procesado adicional para mostrar el funcionamiento del modelo y sus resultados. Debido a la

naturaleza de estos modelos son más fácilmente comprensibles por los humanos. También se les denomina modelos de caja blanca, intrínsecamente transparentes o auto-explicables.

En este grupo se incluirían los árboles de decisión, modelos basados en reglas y aproximaciones lineales (Freitas, 2014) (Ribeiro et al., 2016). La utilización de cada uno de ellos dependerá del campo en el que se apliquen. El modelo lineal no sería apropiado para tratar con numerosas características relevantes y datos no lineales. En estos casos, el modelo de árbol puede ser más adecuado que el modelo lineal. Los árboles de decisión presentan la información de forma gráfica, mientras que las reglas tienen una representación textual (si-entonces).

Estos modelos captan la lógica asociada a los datos, pero cuando las dimensiones son abrumadoras, las explicaciones podrían ser inmanejables para las personas.

En cierto sentido, los modelos intrínsecamente interpretables pueden tener un problema de precisión. Aun así, estudios como(Rudin, 2019) abogan por la utilización de este tipo de modelos, en especial en dominios de alto riesgo.

- Modelo de caja negra interpretable

Este enfoque parte del modelo utilizado para la predicción y utiliza procedimientos y técnicas que permiten extraer información explicativa del modelo. El modelo de base no es intrínsecamente interpretable, lo que se llama modelo de caja negra y suele ser de alta complejidad. También se les denomina modelos post-hoc.

(Guidotti et al., 2018) divide la explicación de caja negra, en función del enfoque de la explicación: explicación del modelo, explicación del resultado, e inspección del modelo.

El objetivo de la explicación del modelo es dar una visión global del comportamiento de la caja negra utilizando para ello un modelo interpretable. Este modelo se debe aproximar al comportamiento de la caja negra y además debe ser comprensible. Los modelos utilizados para tal fin pueden incluir árboles de decisión o clasificadores de reglas de decisión.

La explicación del resultado trata de dar una explicación de la predicción dada por una caja negra, en función de su instancia de entrada. Explica los motivos de la predicción, no la lógica interna de la caja negra.

Por último, el problema de la inspección del modelo sería un punto intermedio entre las anteriores. Proporciona una representación que permita entender una propiedad específica del modelo o de sus predicciones, sin ser necesaria una compresión global del modelo. La representación podría ser visual con gráficos o a través de literatura y texto.

- Dependencia del modelo

  Otro enfoque sería considerar si los métodos utilizados para la interpretación dependen o no del modelo inicial utilizado, clasificándose en específicos del modelo o métodos agnósticos al modelo.

  - Métodos específicos del modelo

    Estos métodos tienen en cuenta características propias del modelo que están analizando, por lo que no podría ser utilizado con otro tipo de modelo predictor.

  - Métodos agnósticos al modelo

    En los métodos agnósticos al modelo, las explicaciones no dependen del modelo de caja negra que se utilice. La información que proporcionan depende de la observación de entrada y salida. Funcionan para cualquier tipo de modelo. Debido a que son genéricos, pueden ser menos eficientes y explicativos que los específicos del modelo.

## II. Humano en el bucle

En relación con el eje identificado por (Adadi & Berrada, 2018) correspondiente a la necesidad de dar importancia a los humanos y cuáles son sus intereses respecto a la explicación, (Langer et al., 2021) enfatiza el papel de las partes interesadas sobre la explicabilidad y como se debe satisfacer sus desideratas de comprensión. Identifica varios grupos de interesados, junto con sus deseos a ser cubiertos con la interpretación.

- Usuarios que utilizan las recomendaciones para toma de decisiones, sus deseos para la explicabilidad serían la usabilidad y confianza. Un sistema es más utilizable si los resultados están acompañados de explicaciones. Si la información proporcionada por

las explicaciones está alineada con el conocimiento o experiencia generará confianza en el sistema.

- Desarrolladores que diseñan, programan y construyen los sistemas artificiales. Los principales intereses de este grupo son verificación y rendimiento. La explicabilidad del modelo permitirá analizar el funcionamiento y posibilitará la corrección de errores y ajustes que pueden llevar a la mejora en la predicción y funcionamiento del sistema.

- Partes afectadas por las decisiones de los sistemas, un grupo cada vez más amplio debido al gran crecimiento de decisiones automatizadas. Cuyo interés en la explicabilidad sería la equidad y la ética de los modelos. Los modelos pueden intensificar los sesgos, que podrían ser identificados con la explicación de los motivos por los que se llega a un resultado.

- Implantadores de un sistema cuyos intereses fundamentales serían que el sistema implantado por ellos sea aceptado y que cumpla con las legislaciones.

- Reguladores que estipulan las normas legales y éticas, para los cuales la explicabilidad de los modelos es una forma de facilitar la legalidad, ética y robustez de los sistemas.

### III. Medición y evaluación

Otro de los ejes a considerar de la interpretabilidad es el de la evaluación de efectividad de las explicaciones, comparando y validando para cuantificar la mejora que representan.

(Murdoch et al., 2019) establece un marco predictivo, descriptivo y relevante (PDR) para seleccionar y evaluar los métodos de interpretación. Entendiendo la predicción predictiva como la capacidad del modelo para aproximar las relaciones de los datos, y la predicción descriptiva como capacidad de las interpretaciones para explicar lo aprendido por el modelo. El marco considera que una interpretación confiable maximiza la predicción predictiva y descriptiva, además, esta interpretación es relevante para una audiencia particular de un dominio elegido. Las mejoras en la precisión predictiva son fáciles de medir con las métricas existentes, mientras que no hay un protocolo de evaluación estándar para evaluar las mejoras en precisión descriptiva o relevancia. En su trabajo indica direcciones a tal efecto como estudios de simulación o validación retrospectiva, a partir de hallazgos experimentales previos que se puedan tomar como verdad fundamental.

A partir de la línea establecida por (Doshi-Velez & Kim, 2017) con un enfoque de evaluación con tres categorías: basada en aplicaciones, basada en humanos y basada en funciones, (Nguyen & Martínez, 2020) se centran en la evaluación basada en funciones y consideran que las explicaciones poseen tres aspectos cuantitativos que pueden ser medidos objetivamente: fidelidad, sencillez y amplitud. Bajo su punto de vista, no sería posible definir métricas que puedan aplicar a todas las técnicas de interpretabilidad, debido a la naturaleza de las interpretaciones que dependen de las características utilizadas para la explicación y del método de la explicabilidad.

En un estudio posterior (Zhou et al., 2021) realiza una distinción entre evaluación centrada en humanos y evaluación basada en funciones. La primera agrupa las definidas por (Doshi-Velez & Kim, 2017) como evaluación basada en aplicaciones y basada en humanos, ya que ambas emplean experimentos con usuarios. En estos experimentos podría utilizarse tanto métricas cualitativas como cuantitativas para evaluar las cualidades de la explicación. En este tipo de evaluaciones es difícil la comparación de la calidad ya que no existen criterios consensuados. Por su parte, las evaluaciones basadas en funciones pueden proporcionar métricas cuantitativas sin necesidad de experimentos con humanos. Analizan la evaluación de la calidad de las explicaciones en relación con la medida en que satisfacen las propiedades de las explicaciones: interpretabilidad y fidelidad, así como de sus propiedades derivadas: claridad, amplitud y simplicidad para la interpretabilidad, y completitud y solidez para la fidelidad. Los tipos de explicaciones basadas en modelos y ejemplos se utilizan mayoritariamente para evaluar la simplicidad de la interpretabilidad y las explicaciones basadas en atribuciones se utilizan frecuentemente para evaluar la solidez de la fidelidad.

## 2.2. Interpretabilidad en el dominio financiero

Una vez explicada la relevancia que tiene la interpretabilidad en los modelos de aprendizaje automático y sus características más relevantes, se hace un repaso a su aplicación en el dominio financiero.

(Bracke et al., 2019) desarrolla un marco analítico para abordar el problema de la explicabilidad, mediante el estudio de las entradas y salidas. Identifica cinco tipos de explicaciones significativas, desde un nivel más individual por instancia hasta una interpretación más general de los modelos y los relaciona con las partes interesadas en la explicación. Utilizando dos modelos de ML, Regresión Logística y Gradient Tree Boosting (GTB), y la técnica para la explicabilidad Quantitative Input Influence (QII) (Datta et al., 2016),

hace una comparativa de ambos modelos para cada tipo de explicación significativa, lo que permite ver cómo se adaptan las explicaciones a cada uno.

(Bussmann et al., 2021) propone un modelo de ML explicable para la gestión de riesgo crediticio y, en particular, en la medición de los riesgos de créditos en plataformas peer to peer (P2P). El planteamiento en este caso sería realizar una selección de modelos en función de su precisión predictiva, empleando posteriormente una técnica de XAI que consigue la interpretabilidad. Se comparan los modelos de Regresión Logística y Gradient Boosting (XGBoost), por precisión selecciona este último y aplica valores Shapley para explicar la contribución de cada una de las variables explicativas.

En la misma línea de aplicación, (Moscato et al., 2021) propone un estudio comparativo de modelos de ML para la predicción de riesgo crediticio, con el objetivo de predecir si un préstamo se reembolsará en una plataforma P2P. El estudio selecciona los tres mejores modelos, en términos de precisión predictiva, a los que aplica técnicas de XAI, como LIME, Anchors, SHAP, BEEF y LORE. Los modelos junto a las técnicas de XAI se evalúan según el protocolo experimental descrito en (Ribeiro et al., 2016). Como resultado de la comparativa la técnica de LORE muestra los mejores resultados en los tres modelos seleccionados al combinar predicciones locales con el uso de explicaciones contrafactuales para mejorar la comprensión de la explicación.

(Munkhdalai et al., 2021) proponen un nuevo modelo parcialmente interpretable para la calificación crediticia, PIA-Soft, que combina modelos de red neuronal y regresión softmax. La parte lineal de regresión permite explicar la relación entre variables de entrada y salida. La red neuronal identifica la relación no lineal entre características. En el estudio se compararon diversos modelos, tanto de caja negra como el modelo Regresión Logística y utilizaron este último para comparar los coeficientes estimados del modelo PIA-Soft y comprobar que eran lógicamente consistentes con la regresión logística y la vida real.

En un enfoque diferente, utilizando modelos intrínsecamente interpretables, (C. Chen et al., 2018) presenta un modelo de riesgo aditivo en dos capas, que denominan Two-Layer Additive Risk Model, y una herramienta de visualización interactiva. El modelo globalmente interpretable permite el razonamiento basado en casos, análisis por importancia de las características y las restricciones de monotonicidad. En el estudio se compara el modelo de dos capas con otros modelos como Regresión Logística, RandomForest, Máquinas de Vector de Soporte (SVM) o Redes Neuronales, comprobando una mejor precisión frente a todos ellos, por lo que se demuestra que no siempre son necesarios lo modelos de caja negra para la evaluación de riesgo crediticio.

## 2.3. Conclusiones interpretabilidad

Como se ha analizado en los apartados anteriores la interpretabilidad de los modelos es importante y se hace necesaria en ciertos dominios, como el financiero, para cumplir la legislación vigente, como GDPR. Si bien se ha visto que no hay un único enfoque para la aplicación, ya que esta debe ajustarse a los objetivos de las partes interesadas. Este trabajo se centra en la interpretabilidad local de los modelos, que permitan dar una justificación de la predicción realizada por el modelo para una instancia. De esta forma, las partes afectadas en la decisión, los prestatarios, podrían conocer las causas de la decisión y comprobar que los modelos son justos en su tarea.

Se desea comparar cuál es el rendimiento de modelos de distintas arquitecturas no solo atendiendo a la precisión de estos sino también a la interpretabilidad. Para ello se utilizarán modelos interpretables intrínsecamente y modelos de caja negra. A estos últimos se aplicarán técnicas de XAI que permitan su interpretabilidad.

Para realizar la evaluación de la precisión de los modelos se utilizará la métrica de área bajo la curva ROC (AUROC). Para realizar la evaluación de la interpretabilidad se utilizará el enfoque propuesto por (Doshi-Velez & Kim, 2017) para la evaluación basada en funciones. Se emplearán para ello los factores relevantes de explicabilidad local, a partir de los cuales se generarán métricas para la evaluación de la interpretabilidad de los modelos.

## 2.4. Modelos interpretables intrínsecamente y de caja negra

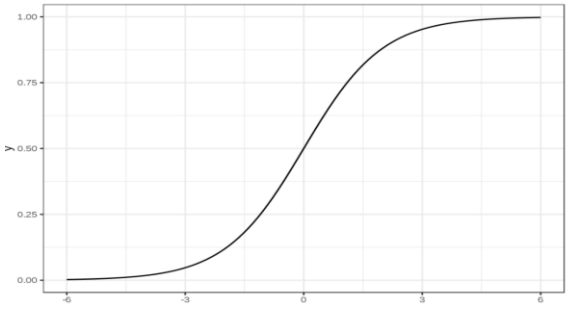
### 2.4.1. Regresión Logística

La Regresión Logística es una extensión del modelo de regresión lineal para problemas de clasificación, cuya salida modela las probabilidades de pertenencia a una clase.

El ajuste de los datos no se realiza a través de una línea recta o hiperplano, sino que utiliza la función logística para realizar dicho ajuste, dando una salida entre 0 y 1. La función logística se define como:

$$log(p) = \frac{1}{1 + exp(-p)}$$

La representación de dicha función corresponde con la Figura 3.




Fuente: (Molnar, 2021).

Figura 3. Representación de la función logística

Como se explica en (Molnar, 2021), en un modelo de Regresión Lineal, dada una instancia $x$, con el valor de sus $p$ características $(x_1, x_2, \cdots, x_P)$, la relación entre estas características y la predicción vendría dada por la fórmula:

$$\hat{y}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Los coeficientes $\beta_i$ corresponden al peso de la característica $i$.

En clasificación se desea obtener la probabilidad entre 0 y 1, por lo que se ajusta dicha ecuación a la ecuación logística:

$$P(y = 1) = \frac{1}{1 + exp\left(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)\right)} \quad (1)$$

Para la interpretación de los pesos en Regresión Logística se debe tener en cuenta que los pesos no influyen linealmente en la probabilidad obtenida como resultado. Reformulando la ecuación $(1)$ se obtiene:

$$log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

A la probabilidad del evento dividido por la probabilidad del no evento se denota como *odds* y al término en la función $log()$ como *log-odds*. Aplicado el logaritmo a dichos *odds*, se puede ver que el modelo de Regresión Logística es un modelo de Regresión Lineal para *log-odds*.

$$log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Para obtener cómo cambia la predicción en términos de la variación en una unidad de la característica se aplica a ambos lados la función $exp()$:

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp\ (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

Y se compara la ratio entre dos predicciones cuando se aumenta en una unidad la característica $x_j$:

$$\frac{odds_{x_j+1}}{odds} = \frac{exp\big(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j(x_j + 1) + \cdots + \beta_p x_p\big)}{exp\big(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_p x_p\big)}$$

Se simplifica aplicando la regla $\frac{exp(a)}{exp(b)} = exp(a-b)$ obteniendo:

$$\frac{odds_{x_j+1}}{odds} = exp\big(\beta_j(x_j + 1) - \beta_j x_j\big) = \exp\ (\beta_j)$$

Lo que significa que un cambio de una unidad en una característica cambia la ratio de *odds* en un factor de $\exp\ (\beta_j)$. Un cambio de $x_j$ en una unidad aumenta la relación de log-odds en el valor del peso correspondiente, cuando todas las otras características permanecen constantes.

En función del tipo de dato de cada característica la interpretación sería:

- Característica numérica: si aumenta el valor de la variable $x_j$ en una unidad, las probabilidades estimadas cambian en un factor de $\exp\ (\beta_j)$.
- Característica categórica binaria: cambiar la variable $x_j$ de la categoría de referencia al otro valor, cambia las probabilidades estimadas por un factor de $\exp\ (\beta_j)$.
- Características categóricas con más de dos categorías: cada una de las características se transforma en características hot encoding, donde cada categoría de una característica tiene su propia columna categórica binaria, y su interpretación es equivalente a la interpretación de características categóricas binarias.

## 2.4.2. Árbol de decisión

Los modelos basados en árboles dividen los datos varias veces en función de ciertos valores de corte en las características. Las divisiones crean subconjuntos del conjunto de datos y cada instancia pertenece a un único subconjunto. Los subconjuntos finales se denominan nodos terminales o de hoja, al resto se denominan nodos internos.

Se puede establecer los criterios de corte para realizar divisiones, cuándo detener la división y el tamaño del subconjunto hoja, lo que dará lugar a generación de distintos árboles.

Matemáticamente la relación entre el resultado y las características se puede describir según (Molnar, 2021) como:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I_{\{x \in R_m\}}$$

Correspondiendo con $R_m$ los subconjuntos hoja donde caen cada una de las instancias. $I_{\{x \in R_m\}}$ es la función identidad que devuelve 1 si la instancia $x$ está en el subconjunto $R_m$ y 0 en otro caso. Si una instancia cae en un nodo hoja $R_l$, el resultado previsto es $\hat{y} = c_l$ donde $c_l$ es el promedio de todas las instancias de entrenamiento en el nodo hoja $R_l$.

Una medida para el criterio de corte es el índice Gini, que indica el grado de impureza de un nodo. Si todas las clases tienen la misma frecuencia dentro del nodo, este es impuro. Si solo hay una clase, es completamente puro. El índice de Gini se minimiza cuando los datos en los nodos tienen valores muy similares. El mejor punto de corte del nodo hace que se separe en dos subconjuntos lo más diferentes posibles respecto al objetivo. Una vez encontrado dicho punto de corte divide el nodo en dos nodos nuevos que se añaden al árbol. La división puede terminar en función del criterio de número de instancias en un nodo antes de la división o el número de instancias en los nodos terminales.

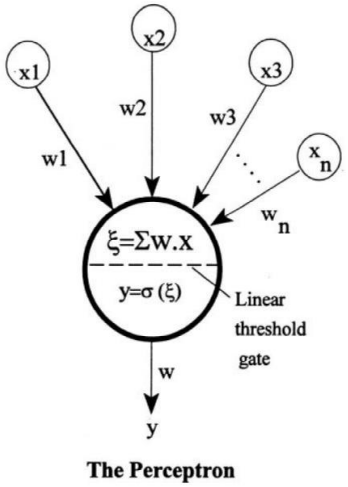
La interpretación del árbol parte del nodo raíz, siguiendo los nodos internos hasta llegar a la hoja, cuyo nodo indica el resultado predicho. La instancia que cae en un nodo hoja cumple todas las condiciones de corte de cada uno de los nodos desde la raíz hasta dicho nodo hoja.

Según (Freitas, 2014), la compresión de los árboles de decisión se ve facilitada por varios factores, como que un árbol de decisión normalmente contiene un subconjunto de características, en lugar de todas. El factor más importante para la comprensión es la estructura gráfica del árbol, fácil de entender y que proporciona información sobre la relevancia entre características debido a su representación jerárquica, más relevante cuanto más cercano a la raíz. Por el contrario, dado que se debe preservar esa estructura de árbol, en algunos casos puede llevar a contener atributos irrelevantes o a tener que replicar estructuras, esto puede hacer que la interpretación de los árboles sea menor.

### 2.4.3. Redes Neuronales

Las redes neuronales son modelos inspirados en el sistema nervioso humano. Sistemas de neuronas artificiales densamente conectados que son capaces de procesar información y de aproximar cualquier función continua con una precisión arbitraria.

A partir de los hallazgos en el funcionamiento de las neuronas biológicas McCulloh y Pitts modelaron el funcionamiento de neuronas artificiales simples. Una neurona recibe entradas, $x_i$, que son ponderadas con los pesos en función de su importancia. Estas entradas ponderadas, $w_i x_i$, son sumadas y se aplica una función de activación, $\sigma$, pasando a través de una puerta umbral lineal, para obtener una salida, $y$, como se puede ver en la Figura 4. Solo cuando la suma ponderada supera el umbral la neurona se activará.




Fuente: (Basheer & Hajmeer, 2000).

Figura 4. Diagrama descriptivo de una neurona artificial

Matemáticamente se puede expresar con la ecuación:

$$y(x) = \sigma\left(\sum_{i=1}^{n} w_i x_i\right)$$

Las características que diferencian los modelos neuronales es la función de activación $\sigma$, la tipología de la red y el algoritmo de entrenamiento.

La función de activación aplicada a cada neurona permite procesar la información y hacer que se propague por la red. Existen distintas funciones de activación como la función Rectified Linear Unit (ReLU) cuya fórmula sería:

$$\sigma(x) = max(0, x)$$

Esta función de activación es ampliamente utilizada en redes neuronales debido a sus ventajas: no saturación en valores mayores que 0, computacionalmente muy eficiente y aceleración la convergencia de descenso de gradientes del entrenamiento.

La tipología de la red o arquitectura describe el número de neuronas en el modelo, el número de capas y la forma en que están conectadas. La adición de capas intermedias permite resolver problemas no lineales, aumentando la complejidad de la red.

El algoritmo de entrenamiento permite la búsqueda especializada de pesos que permitan ajustarse mejor al conjunto de datos de entrada. El algoritmo de backpropagation es el algoritmo utilizado para entrenar redes neuronales. Se basa en la búsqueda de mínimos utilizando la técnica de descenso de gradiente. Cada iteración consta de dos pasos, el de activación hacia adelante para producir una solución y una propagación hacia atrás del error entre la solución producida y el valor real, actualizando el valor de los pesos para minimizar dicho error.

### 2.4.4. XGBoost

El nombre de XGBoost proviene de Extreme Gradient Boosting (T. Chen & Guestrin, 2016), basado en la propuesta de (Friedman, 2001) de Gradient Boosting, donde los modelos se generan de forma secuencial, utilizando cada nuevo modelo información de los anteriores, lo que permite generar modelos más precisos a partir de otros más simples.

Matemáticamente para un conjunto de datos $D = \{(x_i, y_i)\}$ un modelo de conjunto de árboles usa $K$ funciones aditivas para predecir la salida:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon F$$

Donde $F$ es el espacio de árboles de regresión, también conocidos por CART. Definiendo la función objetivo a minimizar como

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_i) \quad (2)$$

Se necesitan aprender las funciones $f_i$, cada una con una estructura de árbol y puntuaciones en las hojas. Para aprender estas funciones como parámetros se utiliza una técnica aditiva, agregando cada vez un nuevo árbol:

$$\hat{y}_i^{(0)} = 0$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$

...

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Con lo que la ecuación $(2)$ quedaría

$$\mathcal{L} = \sum_i l\left(\hat{y}_i^{(t-1)} + f_t(x_i), y_i\right) + \sum_k \Omega(f_i)$$

Utilizando el error cuadrático medio (MSE) como función de pérdida quedaría

$$\mathcal{L} = \sum_i [2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2] + \sum_k \Omega(f_i)$$

Utilizando la aproximación de segundo orden y eliminando los términos constantes se puede obtener el objetivo simplificado

$$\mathcal{L} = \sum_i [g_i\, f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \sum_k \Omega(f_i)$$

Donde $g_i$ y $h_i$ son los estadísticos de primer y segundo orden en la función de pérdida.

Considerando el término regularizador $\Omega(f_i)$, que controla la complejidad del modelo para evitar sobreajuste, como $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{J=1}^{T} w_j^2$ y las funciones $f_t(x) = w_{q(x)}$, donde $w \in \mathbb{R}^T$ es el vector de puntuaciones en las hojas, $q: \mathbb{R}^m \to \{1, \dots, T\}$ la función que asigna cada punto a la hoja correspondiente y $T$ el número de hojas. Se puede simplificar la ecuación como

$$\mathcal{L} = \sum_i [G_j\, w_j + \frac{1}{2}(H_j + \lambda) w_j^2\,] + \gamma T$$

Donde el mejor $w_j$ se obtiene para $w_j^* = -\frac{G_j}{H_j + \lambda}$ quedando la función

$$\mathcal{L}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

Esta función se puede utilizar como función de puntuación para medir la calidad de la estructura del árbol, una medida de impureza del árbol de decisión, que además tiene en cuenta la complejidad del modelo.

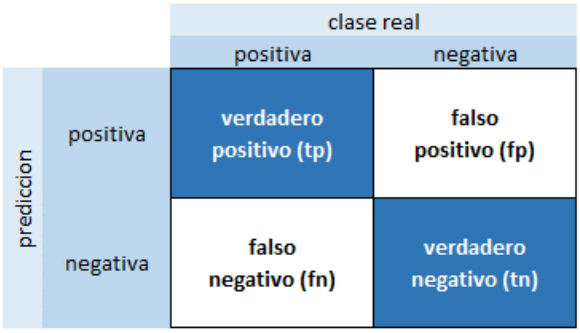
No es posible enumerar todos los árboles posibles y elegir el mejor de ellos, y en su lugar se parte de una sola hoja y se añaden ramas de forma iterativa, hasta que no se pueda mejorar más el modelo.

Uno de los puntos de éxito de XGBoost es su escalabilidad, además, la computación paralela y distribuida permite acelerar el aprendizaje y explorar más modelos de forma más rápida.

## 2.5. Métricas de precisión.

Existe una gran diversidad de medidas de evaluación de la precisión para la tarea de clasificación, utilizadas para medir cada modelo y poder realizar comparativas entre ellos. La mayoría de las medidas de evaluación se pueden expresar en función de la matriz de confusión. Dicha matriz contiene una división de los ejemplos en función de su clase real y la predicción. Para los casos de clases binarias la matriz sería como se muestra en la Figura 5, donde se puede identificar los distintos tipos de predicciones:

- Verdadero Positivo ($tp$): la clase real es positiva y la predicción es positiva.
- Falso Positivo ($fp$): la clase real es negativa y la predicción es positiva.
- Verdadero Negativo ($tn$): la clase real es negativa y la predicción es negativa.
- Falso Negativo ($fn$): la clase real es positiva y la predicción es negativa.




Fuente: Inteligencia Artificial Avanzada. (Benítez et al., 2013).

Figura 5. Matriz de confusión

La medida de exactitud, $accuracy$, correspondiente a las instancias clasificadas correctamente respecto al total de ellas, puede no ser precisa al no considerar la importancia de los falsos positivos frente a falsos negativos.

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

Otras medidas como la sensibilidad, $recall$, y la especificidad, $especificity$, evalúan la exactitud de las muestras positivas y las negativas respectivamente.

$$recall = \frac{tp}{tp + fn} \qquad especificity = \frac{tn}{fp + tn}$$

La precisión, $precision$, corresponde a los ejemplos positivos bien clasificados sobre el total de ejemplos con precisión positiva.

$$precision = \frac{tp}{tp + fp}$$

La métrica $F1$ combina precisión y sensibilidad, utilizando la media armónica, simplificando en una sola métrica el rendimiento del modelo.

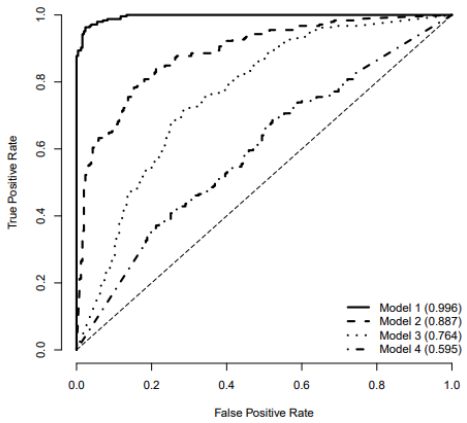
$$F1 = \frac{2 * precision * recall}{precision + recall} = \frac{2 * tp}{2 * fp + fp + fn}$$

Una medida comúnmente utilizada es la curva de características operativas (ROC) que representa la relación entre la tasa de verdaderos positivos ($tpr$) y la de falsos positivos ($fpr$), para un modelo de predicción, para un umbral o punto de corte determinado.

$$tpr = recall = \frac{tp}{tp + fn} \qquad fpr = \frac{fp}{fp + tn}$$

El eje vertical de la curva representa la proporción de verdaderos positivos y el eje horizontal los falsos positivos. Los puntos a lo largo de la curva indican la ratio de los verdaderos positivos ($tp$) a medida que se incrementan los valores de los falso positivos ($fp$). En la Figura 6 se muestra un ejemplo de gráfica de la curva ROC.

La curva ROC ideal coincide con el eje vertical, y los mejores modelos son aquellos cuya curva esté más ajustada a dicho eje. Se utiliza como métrica el área bajo la curva ROC (AUROC), valor entre 0 y 1, que indica cuánto es capaz el modelo de distinguir entre clases, cuanto mayor sea el valor de AUROC mejor será el modelo.




Fuente: (Reina, 2018).

Figura 6. Curva ROC

## 2.6.Técnicas XAI interpretabilidad local

Se analizarán con más detalle las técnicas de XAI, agnósticas del modelo para interpretabilidad local, que permitan una justificación de la predicción realizada para una instancia. Las técnicas agnósticas o independientes del modelo tienen la ventaja de la flexibilidad, ya que se pueden utilizar sobre cualquier modelo de ML. La interpretabilidad local parte de una instancia individual y analiza por qué el modelo tomó una decisión específica.

En un modelo simple lineal, la explicación se obtiene del propio modelo, ya que dada una instancia $x$, con el valor de sus $p$ características $(x_1, x_2, \cdots, x_P)$, la contribución de cada característica a la predicción vendría dada por la fórmula:

$$\hat{y}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Los coeficientes $\beta_i$ corresponden al peso de la característica $i$.

La contribución $\phi_i$ de la característica $i$ a la predicción $\hat{y}(x)$ viene dada por la ecuación:

$$\phi_i(\hat{y}) = \beta_i x_i - E(\beta_i x_i)$$

Donde $E(\beta_i x_i)$ es la estimación del efecto medio de la característica $i$.

Si se suman todas las contribuciones para una instancia, el resultado sería:

$$\sum_{i=1}^{p} \phi_i(\hat{y}) = \sum_{i=1}^{p} \left(\beta_j x_j - E\left(\beta_j x_i\right)\right) = \hat{y}(x) - E\left(\hat{y}(x)\right)$$

Correspondiente al valor predicho para la instancia $x$ menos el valor promedio (Molnar, 2021).

Para modelos más complejos como redes de neuronas no se puede obtener directamente estos pesos, por lo que se debe adoptar otros métodos para obtenerlos.

En su trabajo (Lundberg & Lee, 2017) mostraron cómo varios métodos de explicación utilizaban un mismo modelo de explicación, entendiendo este modelo de explicación como una aproximación interpretable del modelo original. Definieron los métodos de atribución de características aditivas a aquellos métodos que utilizan un modelo de explicación que sea una función lineal de variables binarias:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (3)$$

Donde $z' \in \{0,1\}^M$, $M$ el número de características de entrada simplificado y $\phi_i \epsilon \mathbb{R}$.

Si el modelo de explicación $g$ sigue esta estructura, el método atribuye a cada característica $i$ un efecto $\phi_i$, y la suma de los efectos de todas las atribuciones de características aproxima a la salida $\hat{y}$ para una instancia $x$. Usa los coeficientes $\phi_i$ como importancia de la característica para explicar el modelo original.

Entre los métodos de atribución de características aditivas se encuentran LIME y estimación del valor de Shapley clásico. Adicionalmente (Lundberg & Lee, 2017) propusieron un enfoque unificado para mejorar estos modelos que denominaron valores SHAP (SHapley Additive exPlanation).

### 2.6.1. LIME

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), es un método independiente del modelo, que interpreta predicciones individuales basándose en una aproximación local del modelo de caja negra, a partir de una predicción dada. El modelo de explicación lineal local se ajusta a la ecuación $(3)$ y por tanto es un método de atribución de características aditivas.

La idea principal de LIME es utilizar modelos sustitutos locales que se aproximen a las predicciones del modelo de caja negra. El objetivo es comprender por qué el modelo original hizo una predicción para una instancia concreta.

La explicación producida por LIME se obtiene como:

$$\varepsilon(\mathrm{x}) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (4)$$

$G$ es la clase de modelos potencialmente interpretables, como los árboles de decisión, modelos lineales, o reglas. $\Omega(g)$ es la medida de complejidad de los modelos (opuesto a la interpretabilidad). $f$ es el modelo que se explica y $\pi_x$ es la medida de proximidad a la instancia de interés $x$.

Partiendo de esta instancia $x$ genera un conjunto de datos alrededor de la misma, ponderándolas por $\pi_x$. Para este nuevo conjunto de datos ajusta el modelo sustituto ponderado e interpretable, optimizando la ecuación $(4)$ para obtener la explicación $\varepsilon(\mathrm{x})$.

LIME presenta una explicación que es localmente fiel al modelo de caja negra, pero no tiene que ser una buena aproximación global del modelo.

El principal problema de LIME es la inestabilidad de sus explicaciones, debido a la generación del conjunto de datos alrededor de la instancia $x$ en función de la medida de proximidad $\pi_x$.

Cómo elegir este valor de cercanía es un punto importante y depende de la curvatura local del modelo $f$.

## 2.6.2. Estimación clásica del valor de Shapley

Los valores de Shapley (Shapley, 2016) es un método de la teoría de los juegos cooperativos que permite determinar cómo distribuir equitativamente la predicción entre las características, asignando a cada una su contribución. Este valor representa el efecto que tiene sobre el modelo la inclusión de esa característica.

Para calcularlo se parte de un conjunto F el conjunto de todas las características, siendo $i$ una de esas características a analizar. Para cada subconjunto de características sin la característica $i$, $S \subseteq F \setminus \{i\}$, se entrena el modelo con el conjunto $S \cup \{i\}$, es decir con el conjunto de características S y la característica $i$ presente, $f_{S\cup\{i\}}$. Se entrena también el modelo con el conjunto de características S sin la característica $i$, $f_S$, y se comparan las previsiones de ambos modelos para la instancia. Estos cálculos se repiten para todos los posibles subconjuntos $S$. Los valores de Shapley son un promedio ponderado de todas las posibles diferencias.

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right]$$

Considerando $\phi_0 = f_\emptyset\,(\emptyset)$, se obtendría una fórmula que se ajusta a $(3)$, correspondiente a un método de atribución de características aditivas.

(Shapley, 2016) demostró que los valores de Shapley satisfacen tres axiomas denominados de simetría, eficiencia y ley de agregación (*symmetry*, *efficiency* y *law of aggregation*):

- Simetría (*Symmetry*): si dos características $i$ y $k$ contribuyen por igual a todas las coaliciones, las contribuciones deben ser iguales para ambas características.
- Eficiencia (*Efficiency*): la suma de las contribuciones de características deben ser la diferencia entre la predicción para $x$ y el promedio. Además, si una característica que no cambia el valor predicho, para cualquier coalición a la que se agregue, debe tener el valor de Shapley igual a 0 (dummy).
- Ley de agregación (*law of aggregation*): para modelos combinados, se puede calcular el valor Shapley para cada modelo y obtener el valor del modelo combinado promediando dichos valores Shapley.

Existiendo una única solución que satisface estos tres axiomas y cumple con la definición (3), que son los valores de Shapley.

### 2.6.3. SHAP

El método de explicación SHAP (Lundberg & Lee, 2017) calcula los valores de Shapley a partir de la teoría de juegos de coalición.

En su trabajo se describen tres propiedades deseables que se cumplen en los métodos de atribución de características aditivas: precisión local, ausencia y consistencia (*local accuracy*, *missingness*, *consistency*), similares a los axiomas que cumplían los valores de Shapley, y demuestra que solo hay un método de atribución de características aditivas que cumpla las propiedades. Proponen un enfoque unificado que mejora los métodos anteriores, evitando que violen involuntariamente estas tres propiedades.

Los valores SHAP proporcionan la única medida aditiva de importancia de características que se adhiere a estas tres propiedades y utiliza expectativas condicionales para definir entradas simplificadas. La definición de los valores SHAP está alineada con los valores de Shapley, al tiempo que permite conectar con LIME. La ventaja de SHAP es que tiene una base teórica sólida en la teoría de juegos para llegar a una explicación.

# 3. Objetivos y metodología de trabajo

A continuación, se exponen los objetivos y metodología que se van a utilizar para la elaboración del trabajo.

## 3.1. Objetivo general

El objetivo general del trabajo es realizar una comparativa de distintos modelos de ML interpretables aplicados al dominio financiero en la tarea de la predicción de riesgo crediticio. En la evaluación de dichos modelos se tendrá en cuenta no solo su precisión sino también la interpretabilidad. Es deseable que los modelos tengan una precisión superior al 70% y además que permitan justificar la decisión específica, dando información de las razones por las que se realizaron las predicciones, a partir de los datos de entrada. Para evaluar los modelos por su interpretabilidad se utilizarán métricas basadas en las explicaciones. Es deseable que los modelos mantengan una alta puntuación en dichas métricas de interpretabilidad.

## 3.2. Objetivos específicos

Los objetivos específicos para este trabajo son los siguientes:

- Revisar el estado del arte en interpretabilidad respecto a tres ejes de clasificación, partes interesadas y evaluación.
- Revisar los trabajos más destacados para la aplicación de la interpretabilidad en el dominio financiero.
- Explorar modelos intrínsecamente interpretables y técnicas XAI.
- Analizar el dataset para la predicción del riesgo crediticio, identificando las características más destacadas de los datos y realizando un tratamiento de estos, en caso de ser necesario.
- Implementar los modelos a comparar y aplicar las técnicas de interpretabilidad a aquellos modelos de caja negra no interpretable.
- Evaluar los modelos atendiendo a la precisión predictiva y a la interpretabilidad de estos.
- Comparar los resultados obtenidos por los diferentes modelos entrenados y analizarlos para obtener las conclusiones.

## 3.3. Metodología del trabajo

Para alcanzar los objetivos fijados se utilizará una metodología de trabajo basada en las siguientes fases:

- Fase 1. Realizar un amplio estudio de la literatura existente sobre la interpretabilidad. Se realizará una búsqueda bibliográfica de los distintos enfoques y clasificaciones existentes actualmente para aplicación de interpretabilidad en los modelos de ML.

- Fase 2. Revisión y selección de los modelos a utilizar en la comparativa. Se utilizarán modelos intrínsecamente interpretables y se compararán con modelos de caja negra a los que se aplicarán técnicas de XAI. Las interpretaciones serán locales y permitirán justificar los motivos por los que se llegó a una decisión, a partir de los datos iniciales.

- Fase 3. Análisis del dataset para la predicción del riesgo crediticio. Se ha seleccionado el dataset utilizado para el desafío público planteado por Fair Isaac Corporation (FICO) (FICO community, 2018) para la evaluación del riesgo crediticio. Se analizarán las características que componen el conjunto de datos y los valores asociados, realizándose el tratamiento necesario antes de la utilización en la comparativa. Se realizará la división de los datos entre entrenamiento y test.

- Fase 4. Implementación de los modelos. En esta fase se desarrollarán los distintos modelos elegidos para la comparativa y se ajustarán para alcanzar la máxima precisión. Para los modelos de caja negra se aplicarán técnicas de XAI que permita una explicación local de los modelos.

- Fase 5. Ejecución de modelos y estudio de resultados. Tras entrenar los modelos se ejecutarán sobre el conjunto de test y se comparará la precisión obtenida por cada modelo. Para la evaluación de la interpretabilidad de los modelos se elegirán un conjunto reducido de instancias sobre las que se analizarán las explicaciones dadas por cada uno. Se compararán las unidades básicas que componen las explicaciones analizando la complejidad, número y relaciones entre ellas.

# 4. Planteamiento de la comparativa

A continuación, se presentan las características del conjunto de datos utilizado para la comparativa, y se explica el tratamiento realizado sobre dicho conjunto previo al entrenamiento de los modelos. También se presenta la arquitectura utilizada para cada uno de los modelos y el procedimiento para obtener los parámetros óptimos en cada uno de ellos. Se muestra un ejemplo de interpretación para una instancia determinada en cada uno de los modelos y sus técnicas asociadas, de forma que permita explicar la predicción realizada para una instancia. Se plantean las métricas que se utilizarán para la comparativa de los modelos para medir la precisión de las explicaciones y finalmente se indican los criterios que se tendrán en cuenta para la comparativa de los distintos modelos.

## 4.1. Conjunto de datos

Se ha seleccionado el dataset Home Equity Line of Credit (HELOC), proporcionado por Fair Isaac Corporation (FICO) en 2018 para Explainable Machine Learning Challenge (xML Challenge) (FICO community, 2018) sobre los desafíos y oportunidades para la inteligencia artificial en servicios financieros. El xML Challenge corresponde a una colaboración entre Google, FICO y académicos en Berkeley, Oxford, Imperial, UC Irvine y MIT, con el objetivo promover investigaciones en el área de explicabilidad algorítmica.

El dataset HELOC es un conjunto de datos anónimos sobre las solicitudes de línea de crédito con garantía hipotecaria realizadas por propietarios reales. Una línea de crédito HELOC ofrecida por los bancos estadounidenses se basa en un porcentaje del valor líquido de la vivienda, la diferencia entre el valor de mercado actual de la vivienda y el saldo pendiente de todos los gravámenes, como las hipotecas. En el conjunto de datos HELOC, los clientes solicitaron una línea de crédito en el rango de 5.000 a 150.000 dólares estadounidenses.

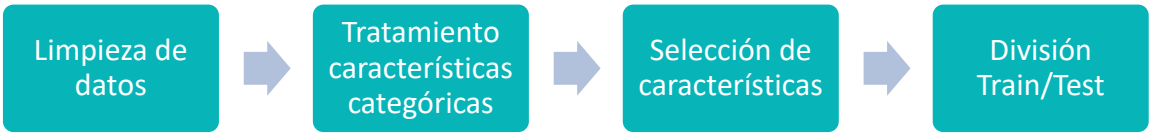
Las variables predictoras son todas cuantitativas o categóricas y provienen de datos anónimos de agencias de crédito. La variable dependiente a predecir es una variable binaria llamada RiskPerformance, cuyo valor *Bad* indica que el solicitante tuvo un retraso de 90 días o más durante un periodo de 2 años desde que abrió la cuenta de crédito. Un valor *Good* indica que ha realizado los pagos con menos de 90 días de atraso. El modelo entrenado en este conjunto de datos puede predecir si un prestatario reembolsará su cuenta HELOC dentro de 2 años. Esta información se puede utilizar posteriormente para decidir si el prestatario puede acceder a una línea de crédito y por cuánto crédito. En la Tabla 1 se muestra la relación de características y su significado obtenida de la especificación del dataset.

Tabla 1. Descripción de las características del dataset HELOC

Fuente: Elaboración propia a partir de la descripción del dataset HELOC (FICO community, 2018). Se muestra cada característica del dataset y su descripción

| Característica | Descripción |
|---|---|
| RiskPerformance | Indicador de pago los últimos 24 meses. |
| ExternalRiskEstimate | Indicador consolidado de marcadores de riesgo. |
| MSinceOldestTradeOpen | Número de meses que han transcurrido desde la primera operación. |
| MSinceMostRecentTradeOpen | Número de meses que han transcurrido desde la última operación abierta. |
| AverageMInFile | Meses promedio sin operaciones. |
| NumSatisfactoryTrades | Número de operaciones satisfactorias. |
| NumTrades60Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 60 días. |
| NumTrades90Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 90 días. |
| NumTotalTrades | Número total de operaciones. |
| NumTradesOpeninLast12M | Número de operaciones abiertas en los últimos 12 meses. |
| PercentTradesNeverDelq | Porcentaje de operaciones que no estaban en mora. |
| MSinceMostRecentDelq | Número de meses que han transcurrido desde la última operación morosa. |
| MaxDelq2PublicRecLast12M | Período de morosidad más largo de los últimos 12 meses. |
| MaxDelqEver | Período de morosidad más largo. |
| PercentInstallTrades | Porcentaje de operaciones a plazos. |
| NetFractionInstallBurden | Fracción neta de la carga de los plazos. Saldo de la cuota dividido por el monto original del préstamo. |
| NumInstallTradesWBalance | Número de operaciones a plazos con saldo. |
| MSinceMostRecentInqexcl7days | Meses desde la última consulta (excepto los últimos 7 días). |
| NumInqLast6M | Número de consultas en los últimos 6 meses. |
| NumInqLast6Mexcl7days | Número de consultas en los últimos 6 meses (excluidos los últimos 7 días). |
| NetFractionRevolvingBurden | Saldo renovable dividido por límite de crédito. |
| NumRevolvingTradesWBalance | Número de operaciones renovable con saldo. |
| NumBank2NatlTradesWHighUtilization | Número de operaciones con un alto índice de utilización (índice de utilización: el monto del saldo de una tarjeta de crédito en comparación con el límite de crédito). |
| PercentTradesWBalance | Porcentaje de operaciones con saldo. |

En la Figura 7 se muestran las etapas seguidas en el procesamiento de los datos, previo a ser utilizados por los modelos, las cuales se detallan a continuación.



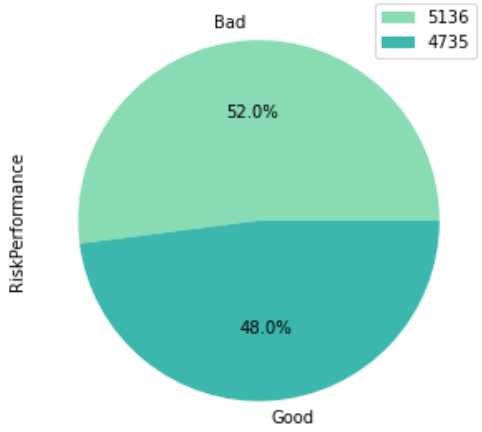

Fuente: Elaboración propia. Se muestran las etapas en las que se ha dividido el procesamiento de los datos.

Figura 7. Etapas en el procesamiento de los datos

El dataset consta inicialmente de información de 10.459 clientes que solicitaron una línea de crédito HELOC. Teniendo en cuenta los valores especiales indicados en las especificaciones del dataset se han identificado 588 instancias las cuales tienen valor -9, ‘Sin registro en la oficina o sin investigación’, en todas las variables predictoras, por lo que carecen de valor para la toma de decisiones. Por ese motivo se eliminan del conjunto original, quedando un total de 9.871 instancias.

Tras revisar el resto de las características con valores especiales: -7, 'Condición no satisfecha (por ejemplo, no hay consultas, no hay morosidad)' y -8, 'No hay operaciones o investigaciones utilizables/válidas', se llega a la conclusión que no es posible sustituir estos valores por otros sin alterar el sentido de las muestras. Sustituir estos valores especiales por el valor medio implicaría que se está considerando como valor aceptable incluso cuando no son válidos. Otra opción sería sustituirlos por 0 aunque, como se explica a continuación, no se consideró como una opción viable. Tres de las columnas con estos valores corresponden al número de meses desde que se cumple cierta condición, en estos casos, al asignar un valor 0 se estaría generando información errónea, para el resto de las características ya existen valores a 0, por lo que se estarían asignando el mismo valor a muestras válidas y muestras con valores especiales. Por ese motivo se decide mantener el resto de los valores especiales, y se tendrán en cuenta a la hora de evaluar los modelos.

En la Figura 8 se puede apreciar la distribución de los datos y el reparto en cada una de las clases. No se aprecia que el conjunto de datos esté desbalanceado entre los posibles valores de la variable predictora. Para el tratamiento de los datos las instancias con la etiqueta *Bad* se actualizarán con el valor 1 y las de la etiqueta *Good* con el valor 0, ya que nuestro objetivo sería identificar aquellos prestatarios con menor probabilidad de reembolsar el crédito en el periodo de 2 años.




Fuente: Elaboración propia de la distribución obtenida tras la limpieza del dataset para cada clase objetivo.

Figura 8. Distribución del conjunto de datos por clase

Existen dos variables categóricas correspondientes al periodo de morosidad más largo y el más largo de los últimos 12 meses. Cada una de ellas tiene asociada una serie de códigos de categorías. El sentido de las categorías de ambas variables está relacionado, pero difieren en los códigos asignados en cada una de ellas. Se hace un tratamiento de los datos para unificar

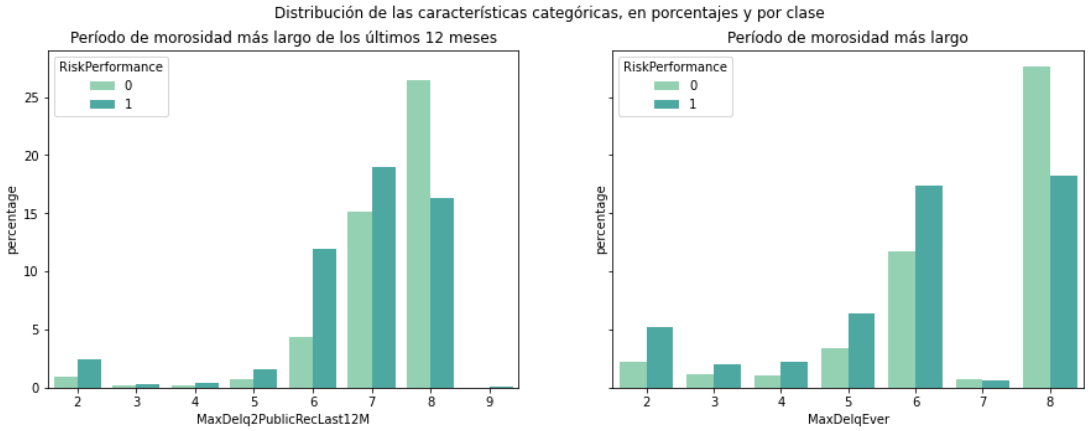
los códigos en ambas características en función de su significado, siendo los valores resultantes los que se muestran en la Tabla 2.

Tabla 2. Categorías unificadas para las variables categóricas

Fuente: Elaboración propia a partir de la descripción del dataset HELOC (FICO community, 2018). Se muestra el código de cada categoría y su significado. Los códigos serán comunes a ambas variables categóricas.

| código | significado |
|---|---|
| 1 | No hay tal valor |
| 2 | comentario negativo |
| 3 | 120+ días de morosidad |
| 4 | 90 días de retraso |
| 5 | 60 días de retraso |
| 6 | 30 días de retraso |
| 7 | morosidad desconocida |
| 8 | actual y nunca moroso |
| 9 | todos los demás |

Según se aprecia en la Figura 9, para ambas variables, la categoría con mayor porcentaje de casos corresponde con los clientes sin morosidad, categoría 8, superando el 25% para la etiqueta *Good* y el 15% a la etiqueta *Bad*, lo cual está alineado con que la posibilidad de reembolso del crédito es mayor para clientes sin morosidad. Para las categorías del 2 al 6 en las que, o bien es un comentario negativo o existe morosidad, el mayor porcentaje sería para la etiqueta *Bad*, lo cual está alineado con el sentido de estos códigos, donde mayor morosidad implicaría menor posibilidad de reembolsar el crédito.




Fuente: Elaboración propia. Se muestra los porcentajes de cada categoría para las etiquetas *Bad* (1) y *Good* (0) en cada característica categórica del periodo de morosidad más largo.

Figura 9. Distribución de Características Categóricas

Para el correcto tratamiento de los valores categóricos por los modelos se realiza una conversión hot encoding a variables dummy. Se sustituye cada variable categórica por las respectivas variables con valor 0 o 1. La variable categórica del periodo de morosidad más

largo se sustituye por 7 variables dummy y la variable categórica del periodo más largo de los últimos 12 meses se sustituye por 8 variables dummy, pasando de 23 variables predictoras a 36.

Se calculan las correlaciones entre las distintas variables predictoras y la variable dependiente, identificándose que la variable con mayor correlación, en este caso negativa, es ExternalRiskEstimate. Esta variable es un indicador consolidado de marcadores de riesgo. Las siguientes variables más relacionadas con la variable dependiente, en este caso con un sentido positivo, son las correspondientes al saldo renovable dividido por el límite de crédito y porcentaje de operaciones con saldo, respectivamente.

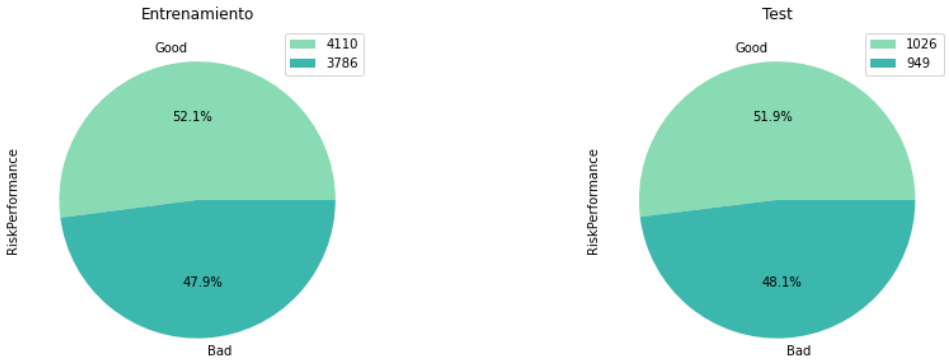
Se calcula la puntuación de la importancia de las características, con el criterio de información mutua, y se seleccionan aquellas que son más representativas en relación con la variable predictora, definiendo para ello el umbral en 0.005. Se considera que la relación de dependencia entre la variable predictora y las características con puntuación inferior a ese umbral no es significativa para la toma de decisiones. Se reduce el número de variables predictoras de entrada de 36 a 24. En la Tabla 3 se muestran las características seleccionadas.

Tabla 3. Características seleccionadas

Fuente: Elaboración propia a partir de la descripción del dataset HELOC (FICO community, 2018). Se muestran las características seleccionadas y su descripción.

| | Características Seleccionadas | Descripción |
|---|---|---|
| **0** | ExternalRiskEstimate | Indicador consolidado de marcadores de riesgo. |
| **1** | MSinceOldestTradeOpen | Número de meses que han transcurrido desde la primera operación. |
| **2** | AverageMInFile | Meses promedio sin operaciones. |
| **3** | NumSatisfactoryTrades | Número de operaciones satisfactorias. |
| **4** | NumTrades60Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 60 días. |
| **5** | NumTrades90Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 90 días. |
| **6** | PercentTradesNeverDelq | Porcentaje de operaciones que no estaban en mora. |
| **7** | MSinceMostRecentDelq | Número de meses que han transcurrido desde la última operación morosa. |
| **8** | NumTradesOpeninLast12M | Número de operaciones abiertas en los últimos 12 meses. |
| **9** | PercentInstallTrades | Porcentaje de operaciones a plazos. |
| **10** | MSinceMostRecentInqexcl7days | Meses desde la última consulta (excepto los últimos 7 días). |
| **11** | NumInqLast6M | Número de consultas en los últimos 6 meses. |
| **12** | NumInqLast6Mexcl7days | Número de consultas en los últimos 6 meses (excluidos los últimos 7 días). |
| **13** | NetFractionRevolvingBurden | Saldo renovable dividido por límite de crédito. |
| **14** | NetFractionInstallBurden | Fracción neta de la carga de los plazos. Saldo de la cuota dividido por el monto original del préstamo. |
| **15** | NumRevolvingTradesWBalance | Número de operaciones renovable con saldo. |
| **16** | NumBank2NatlTradesWHighUtilization | Número de operaciones con un alto índice de utilización. |
| **17** | PercentTradesWBalance | Porcentaje de operaciones con saldo. |
| **18** | MaxDelq2PublicRecLast12M_6 | Período de morosidad más largo de los últimos 12 meses (30 días de retraso). |
| **19** | MaxDelq2PublicRecLast12M_8 | Período de morosidad más largo de los últimos 12 meses (Actual y nunca moroso). |
| **20** | MaxDelq2PublicRecLast12M_9 | Período de morosidad más largo de los últimos 12 meses (Todos los demás). |
| **21** | MaxDelqEver_2 | Período de morosidad más largo (Comentario negativo). |
| **22** | MaxDelqEver_5 | Período de morosidad más largo (60 días de retraso). |
| **23** | MaxDelqEver_8 | Período de morosidad más largo (Actual y nunca moroso). |

Finalmente se realiza una división de los datos en los conjuntos de entrenamiento y test, con una distribución del 80% - 20%. En la Figura 10 se muestra la distribución de los datos para cada conjunto.




Fuente: Elaboración propia. Distribución de los datos una vez realizada la división entre datos de train y test.

Figura 10. Distribución de los conjuntos de datos Train/Test

## 4.2. Arquitecturas de los modelos

La elección de los modelos ha estado condicionada al objetivo de comparar aquellos que son intrínsecamente interpretables con los de caja negra. Para tal fin se han elegido los modelos de Regresión Logística y Árboles de decisión, como intrínsecamente interpretables y Red Neuronal y el clasificador XGBoost como modelos de caja negra.


Fuente: Elaboración propia. Se muestra las etapas en las que se ha dividido la tarea de entrenamiento y validación de los modelos.

Figura 11. Etapas entrenamiento de Modelos

En la Figura 11 se muestran las etapas en el entrenamiento y validación de los distintos modelos. Para cada uno de los modelos se entrena un modelo inicial básico, a partir del cual se han ido ajustando los parámetros. Para el ajuste se ha utilizado la técnica de GridSearchCV de sklearn que permite configurar una serie de parámetros, propios de cada modelo, y hacer una búsqueda exhaustiva de valores para esos parámetros. La técnica genera candidatos a partir de una cuadrícula con los valores proporcionados para cada parámetro y realiza una búsqueda con validación cruzada sobre dicha cuadrícula para encontrar la combinación que aumente la precisión del modelo. En todos los casos se utiliza el atributo *cv* para la activación

del uso de validación cruzada 10-fold, en lugar de la técnica Leave-One-Out Cross-Validation, correspondiente a la validación cruzada dejando un elemento fuera.

Los parámetros considerados para configurar en Regresión Logística con esta técnica son:

- *penality*: correspondiente a la norma utilizada en la penalización, con posibles valores a elegir entre L1 y L2. L1 permite algunos valores de coeficientes $\beta$ a 0, una forma de forzar la selección o no de características. L2 estima $\beta$ pequeños, que sirve para controlar el sobreajuste.
- *C*: inversa de la fuerza de regularización, para la reducción del sobreajuste. Cuanto más pequeño es el valor, mayor es la regularización.
- *solver*: algoritmo a utilizar en el problema de optimización. Como opciones para este parámetro se consideran liblinear, saga, lbfgs.

Los parámetros considerados para configurar en Árboles de Decisión son:

- *criterion*: mide la calidad de la división. Se selecciona como valor gini para la impureza de Gini.
- *max_depth*: profundidad máxima del árbol.
- *max_leaf_nodes*: para la mejor construcción del árbol utilizando el mejor nodo que reduzca impurezas.
- *min_samples_split*: número mínimo de muestras necesarias para dividir un nodo interno.

Para utilizar la búsqueda exhaustiva con los modelos de Redes Neuronales con Keras es necesario crear una función que cree y devuelva el modelo secuencial. Esta función se utilizará como parámetro del constructor de la clase KerasClassifier y el modelo construido será el que se utilice con GridSearchCV.

El modelo secuencial se compone de tres capas densas de 256, 128 y 64 neuronas respectivamente, con activación relu y dropout de 0.2. Y una capa densa final con dos posibles valores correspondientes a la probabilidad de pertenecer a cada una de las dos clases, dada por la activación softmax y la función de pérdida asociada binary_crossentropy.

Los parámetros considerados para configurar en Red Neuronal son:

- *epochs*: número de pases completos a través del conjunto de datos de entrenamiento
- *batch_size*: valor por el que se divide la muestra y sobre el cual se actualiza el gradiente.

- *optimizer*: algoritmo o método utilizado para minimizar la función de pérdida. Los posibles valores utilizados son Adam, Nadam, RMSprop

Los parámetros considerados para configurar en XGBoost son:

- *min_child_weight*: bloquea las interacciones de características potenciales para evitar el sobreajuste. Si el nodo hoja tiene una suma mínima de peso de instancia (calculada por una derivada parcial de segundo orden) inferior al valor de este parámetro, la división del árbol se detiene.
- *gamma*: para el control de la regularización y evitar el sobreajuste. Cuanto mayor sea el valor, mayor será la regularización por lo tanto el sobreajuste
- *learning_rate*: reducción del tamaño del paso utilizada en la actualización para evitar el sobreajuste, con valor entre 0 y 1.
- *subsample*: proporción de submuestras de los datos de entrenamiento antes de cultivar árboles para evitar sobreajuste.
- *colsample_bytree*: proporción de submuestra de columnas al construir cada árbol.
- *max_depth*: profundidad máxima de un árbol.
- *n_estimators*: número máximo de iteraciones o cantidad de árboles a crear.

En la Tabla 4 se muestran los parámetros asociados a cada uno de los modelos obtenidos tras la búsqueda exhaustiva con GridSearchCV.

Tabla 4. Parámetros de configuración por modelo

Fuente: Elaboración propia. Parámetros utilizados en la configuración de cada modelo y los valores finales.

| **Regresión Logística** | |
|---|---|
| penality | l2 |
| C | 1 |
| solver | liblinear |

| **Árboles de Decisión** | |
|---|---|
| criterion | gini |
| max_depth | 5 |
| max_leaf_nodes | 21 |
| min_samples_split | 2 |

| **Red Neuronal** | |
|---|---|
| epochs | 100 |
| batch_size | 128 |
| optimizer | RMSprop |

| **XGBoost** | |
|---|---|
| min_child_weight | 10 |
| gamma | 5 |
| learning_rate | 0.05 |
| subsample | 1.0 |
| colsample_bytree | 0.8 |
| max_depth | 5 |
| n_estimators | 150 |

Una vez ajustados los parámetros para cada uno de los modelos y entrenados, se realiza la validación de estos utilizando para ello el conjunto de datos de test, obteniéndose la matriz de confusión y el valor de área bajo la curva ROC.

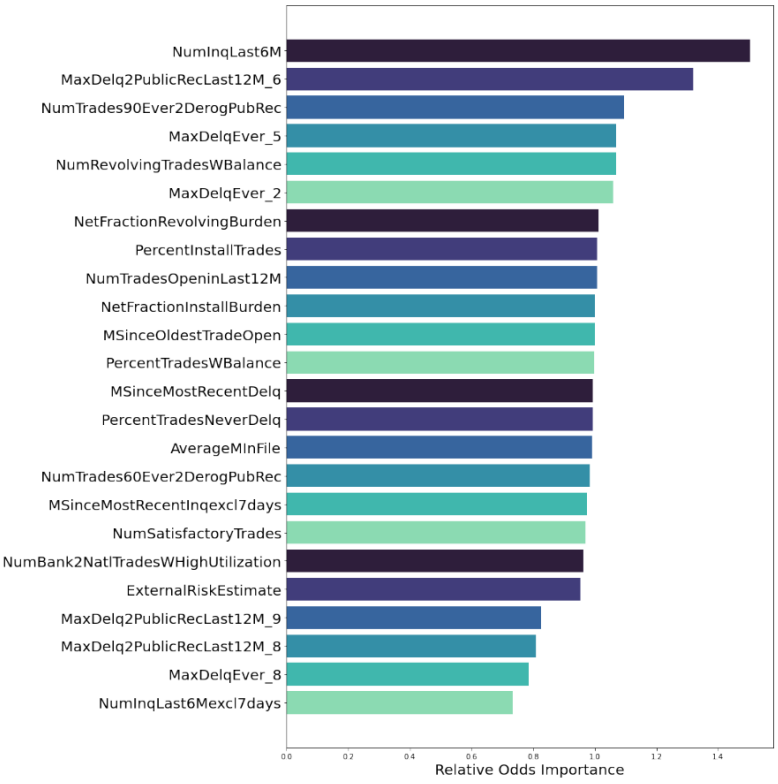
## 4.3. Interpretación de instancias

En la comparativa se han elegido dos modelos intrínsecamente explicables y dos modelos de caja negra a los que se les aplica las técnicas de interpretabilidad seleccionadas, LIME y SHAP. A partir de ello se obtiene la explicabilidad a nivel de instancia en los siguientes casos: Regresión Logística, Árbol de Decisión, Red Neuronal con LIME, XGBoost con LIME, Red Neuronal con SHAP y XGBoost con SHAP. En cada uno de los casos se analiza la explicación local proporcionada.

### 4.3.1. Regresión logística

Como se indicó en el apartado 2.4.1. en relación con la interpretación de Regresión Logística, un cambio de una unidad en una característica cambia la razón entre las posibilidades por un factor de $\exp(\beta_j)$, cuando todas las otras características permanecen constantes.

En la Figura 12 se muestra los coeficientes $\beta_j$ asociados a cada una de las características y su correspondiente valor de odd ($\exp(\beta_j)$).




| features | coef | odds |
|---|---|---|
| NumInqLast6M | 0.407176 | 1.502568 |
| MaxDelq2PublicRecLast12M_6 | 0.276195 | 1.318105 |
| NumTrades90Ever2DerogPubRec | 0.089424 | 1.093544 |
| MaxDelqEver_5 | 0.067424 | 1.069749 |
| NumRevolvingTradesWBalance | 0.066618 | 1.068887 |
| MaxDelqEver_2 | 0.057419 | 1.059100 |
| NetFractionRevolvingBurden | 0.012933 | 1.013017 |
| PercentInstallTrades | 0.007861 | 1.007891 |
| NumTradesOpeninLast12M | 0.007685 | 1.007715 |
| NetFractionInstallBurden | -0.000145 | 0.999856 |
| MSinceOldestTradeOpen | -0.000855 | 0.999145 |
| PercentTradesWBalance | -0.001572 | 0.998429 |
| MSinceMostRecentDelq | -0.005953 | 0.994065 |
| PercentTradesNeverDelq | -0.007430 | 0.992598 |
| AverageMInFile | -0.008066 | 0.991966 |
| NumTrades60Ever2DerogPubRec | -0.016817 | 0.983324 |
| MSinceMostRecentInqexcl7days | -0.025856 | 0.974475 |
| NumSatisfactoryTrades | -0.031428 | 0.969061 |
| NumBank2NatlTradesWHighUtilization | -0.039371 | 0.961394 |
| ExternalRiskEstimate | -0.048740 | 0.952429 |
| MaxDelq2PublicRecLast12M_9 | -0.190457 | 0.826581 |
| MaxDelq2PublicRecLast12M_8 | -0.212509 | 0.808553 |
| MaxDelqEver_8 | -0.242648 | 0.784547 |
| NumInqLast6Mexcl7days | -0.309937 | 0.733493 |

Fuente: Elaboración propia. A la derecha se muestra la tabla con la información del coeficiente $\beta_j$ y su odd $\exp(\beta_j)$ asociado para cada característica y a la izquierda se muestra gráficamente la importancia de cada característica en función de su odd.

Figura 12. Regresión logística, coeficientes y odds

Se puede apreciar que las características con más impacto en la probabilidad de pertenecer a la clase *Bad* serían: El número de consultas en los últimos 6 meses, morosidad con 30 días de retraso de los últimos 12 meses, el número de operaciones que están atrasadas en más de 90 días y los periodos de morosidad más largos de 60 días. Todas estas características indicarían morosidad por parte del cliente, por lo que sería correcta la tendencia a catalogarlo como *Bad*.

En contraposición, entre las características que mayor impacto tendrían en el caso de catalogarse como *Good* estarían los periodos de morosidad más largo y el mayor de los últimos 12 meses con un valor de sin morosidad lo cual también es coherente con dicha tendencia.

Al trasladar esta información a una instancia concreta, para la cual se quisiera calcular la probabilidad de que la variable predictora fuera 1 ($P(y = 1)$), se partiría de los valores de cada característica, sus coeficientes y el producto de estos, según se muestra en la Tabla 5 para una instancia. Aplicando la ecuación (1), se obtiene la probabilidad para una muestra. En este caso un aumento en el número de consultas en los últimos 6 meses cambia (aumenta) las probabilidades de no reembolso del crédito en un factor de 1.502568, cuando todas las demás características siguen siendo las mismas. Un cambio en la morosidad con 30 días de retraso de los últimos 12 meses cambiaría la probabilidad de no reembolso en un factor de 1.318105, siempre considerando que el resto de las características mantienen su valor.

Tabla 5. Regresión Logística, valores coeficientes para una instancia

Fuente: Elaboración propia. Para la instancia de ejemplo y cada característica se muestra su valor, el coeficiente de regresión, el valor de la ratio odd y la multiplicación del valor de la instancia por el coeficiente.

| | features | Xi | coefi | oddsi | Xi_x_coefi |
|---|---|---|---|---|---|
| **0** | biasCoefficient | 1.0 | 4.834994 | 0.000000 | 4.834994 |
| **12** | NumInqLast6M | 9.0 | 0.407176 | 1.502568 | 3.664582 |
| **19** | MaxDelq2PublicRecLast12M_6 | 1.0 | 0.276195 | 1.318105 | 0.276195 |
| **16** | NumRevolvingTradesWBalance | 2.0 | 0.066618 | 1.068887 | 0.133236 |
| **14** | NetFractionRevolvingBurden | 14.0 | 0.012933 | 1.013017 | 0.181062 |
| **10** | PercentInstallTrades | 43.0 | 0.007861 | 1.007891 | 0.338002 |
| **9** | NumTradesOpeninLast12M | 9.0 | 0.007685 | 1.007715 | 0.069166 |
| **15** | NetFractionInstallBurden | 67.0 | -0.000145 | 0.999856 | -0.009682 |
| **2** | MSinceOldestTradeOpen | 148.0 | -0.000855 | 0.999145 | -0.126564 |
| **18** | PercentTradesWBalance | 67.0 | -0.001572 | 0.998429 | -0.105344 |
| **8** | MSinceMostRecentDelq | 6.0 | -0.005953 | 0.994065 | -0.035717 |
| **7** | PercentTradesNeverDelq | 93.0 | -0.007430 | 0.992598 | -0.690959 |
| **3** | AverageMInFile | 47.0 | -0.008066 | 0.991966 | -0.379103 |
| **4** | NumSatisfactoryTrades | 24.0 | -0.031428 | 0.969061 | -0.754266 |
| **1** | ExternalRiskEstimate | 73.0 | -0.048740 | 0.952429 | -3.558036 |
| **13** | NumInqLast6Mexcl7days | 9.0 | -0.309937 | 0.733493 | -2.789436 |

### 4.3.2. Árboles de decisión

Una de las características propias de los Árboles de Decisión es la visualización de dicho árbol. En la Figura 14 se muestra toda la estructura del árbol de decisión creada por el modelo. Para una instancia determinada, partiendo del nodo raíz se puede recorrer el árbol hasta llegar a un nodo hoja, el cual indica la clase en la que se ha clasificado dicha instancia.

A partir de cada nodo intermedio se generan dos ramas en función de la condición de corte asociada una característica. En la Tabla 6 se muestran aquellas características que intervienen en el árbol en alguna condición de corte y su importancia. La característica ExternalRiskEstimate es la es la más importante, con una amplia diferencia en comparación con el resto de las características. Esto es debido a que el nodo raíz y otros nodos intermedios utilizan esta característica como criterio de división y generación de ramas hijas.

Tabla 6. Árbol de Decisión, características que intervienen e importancia

Fuente: Elaboración propia. Se muestran las características que intervienen en la generación del árbol y la importancia asociada para cada una de ellas, en función de su aportación en el árbol.

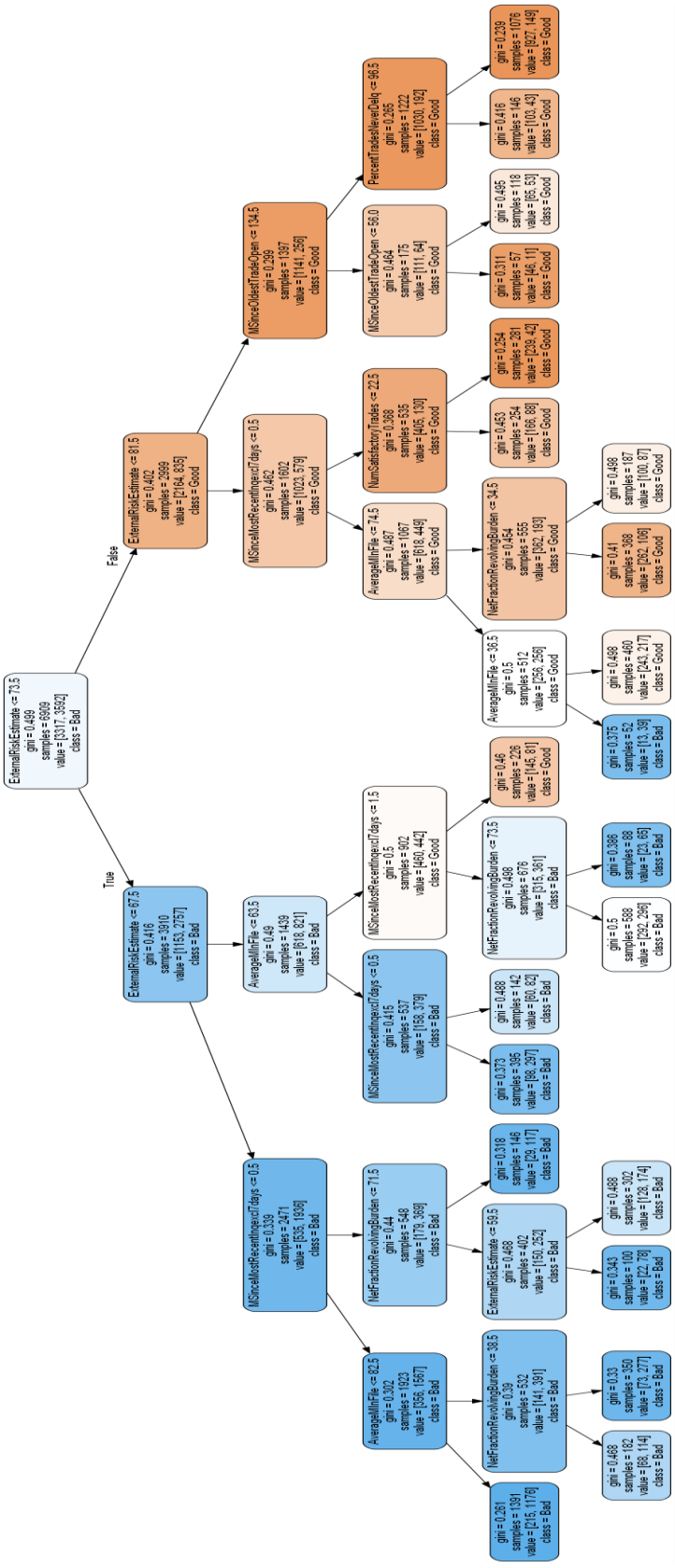
| | feature | importance |
|---|---|---|
| **0** | ExternalRiskEstimate | 0.806 |
| **2** | AverageMInFile | 0.064 |
| **10** | MSinceMostRecentInqexcl7days | 0.060 |
| **13** | NetFractionRevolvingBurden | 0.031 |
| **1** | MSinceOldestTradeOpen | 0.020 |
| **3** | NumSatisfactoryTrades | 0.011 |
| **6** | PercentTradesNeverDelq | 0.007 |

Aplicado a una instancia en concreto se puede recorrer el árbol de decisión para obtener aquellas reglas que aplican a la instancia hasta llegar al nodo hoja. En la Figura 13 se muestra un ejemplo del camino para una instancia, representado como reglas de decisión y en la Figura 15 se puede ver la representación de esas mismas reglas de forma visual en el árbol de decisión.

```
Reglas utlizadas para predecir el ejemplo 428 cuya etiqueta real es 1:
decision node 0 : (X_test[428, ExternalRiskEstimate] = 73.0) <= 73.5)
decision node 1 : (X_test[428, ExternalRiskEstimate] = 73.0) > 67.5)
decision node 4 : (X_test[428, AverageMInFile] = 47.0) <= 63.5)
decision node 7 : (X_test[428, MSinceMostRecentInqexcl7days] = -7.0) <= 0.5)
Clase predicha 1
```

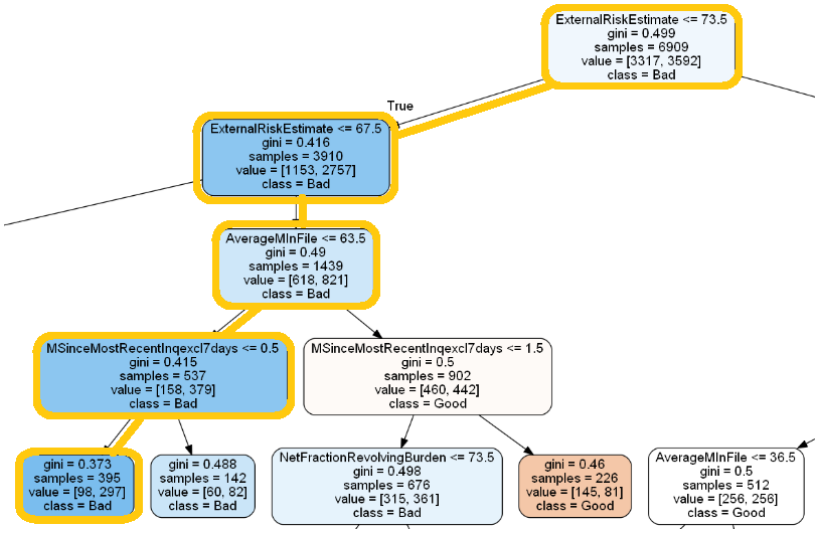
Fuente: Elaboración propia. Para una instancia se muestran las reglas que aplicarían hasta llegar a su nodo hoja.

Figura 13. Árbol de Decisión reglas para una instancia




Fuente: Elaboración propia. Árbol de decisión generado para la clasificación del dataset con 5 niveles de profundidad máxima.

Figura 14. Árbol de Decisión visualización



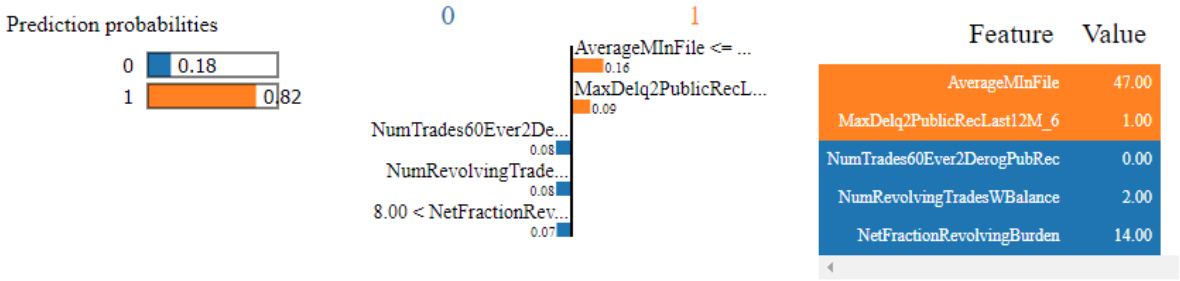

Fuente: Elaboración propia a partir del Árbol de Decisión. Se muestra el camino que aplica para la instancia.

Figura 15. Árbol de Decisión camino seguido por una instancia

### 4.3.3. Red Neuronal + LIME

Para obtener la interpretabilidad del modelo de red Neuronal, en este caso, se ha aplicado la técnica de interpretabilidad LIME, que genera modelos sustitutos locales que se aproximen a las predicciones del modelo de caja negra. Se ha utilizado para ello el método LimeTabularExplainer, cuyo resultado permite explicar la predicción asociada a una instancia, mostrando las características que más influyen en la explicación.

En la Figura 16 se muestra el resultado asociado a una instancia utilizando el modelo de Red Neuronal entrenado y aplicando posteriormente LIME.




Fuente: Elaboración propia a partir de la aplicación de LIME para el modelo de Red Neuronal.

Figura 16. Red Neuronal, aplicación de LIME para una instancia

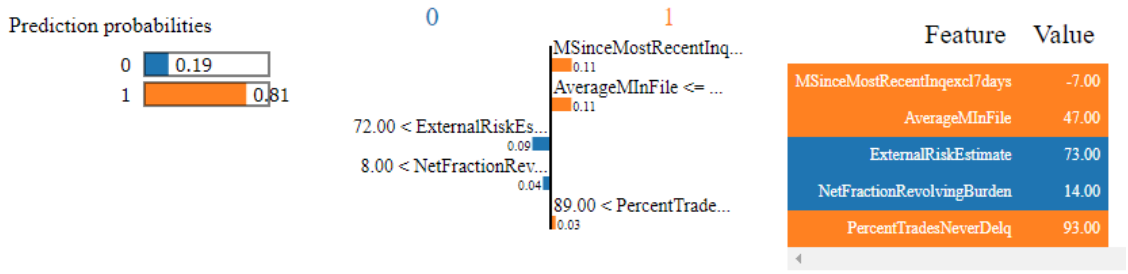
El modelo confía en un 82% que se trata de un cliente que no podría hacer frente al crédito sin demoras durante un período de 2 años, por lo que se etiquetaría como *Bad* ($y = 1$) que coincide con la clasificación original de dicha instancia.

Se marcan en color naranja aquellas características que aumentan la posibilidad de demoras. En color azul se pueden ver aquellas características que hacen decrecer dicha posibilidad y por lo tanto acercan la decisión hacia *Good* ($y = 0$). Las características que más han influido en dicha decisión hacia la etiqueta *Bad*, en color naranja, son el número de meses promedio en archivo, con un valor de 47, y el período de morosidad más largo de los últimos 12 meses, con 30 días de retraso, con un valor de 1 que indica que ha tenido demoras de 30 días en el último año. En color azul se muestran las características del número de operaciones atrasadas en más de 60 días, con un valor de 0 que indica que no existen retrasos, y la información relacionada con las operaciones renovables, tanto el número de operaciones como el saldo renovable dividido por el límite de crédito. Estas tres características disminuyen la posibilidad de demoras.

### 4.3.4. XGBoost + LIME

De la misma forma, tras la generación del modelo XGBoost se aplica la técnica de interpretabilidad LIME, utilizando también el método LimeTabularExplainer. Los valores de LIME obtenidos permiten mostrar cada una de las características más influyentes y en qué forma han influido a la predicción para una instancia.

En la Figura 17 se muestra la misma instancia de ejemplo predicha, en este caso por el modelo XGBoost. En este caso la confianza en la predicción como *Bad* ($y = 1$) es de 81% y se muestran las características más representativas asociadas a dicha decisión.




Fuente: Elaboración propia a partir de la aplicación de LIME para el modelo XGBoost.

Figura 17. XGBoost, aplicación de LIME a una instancia

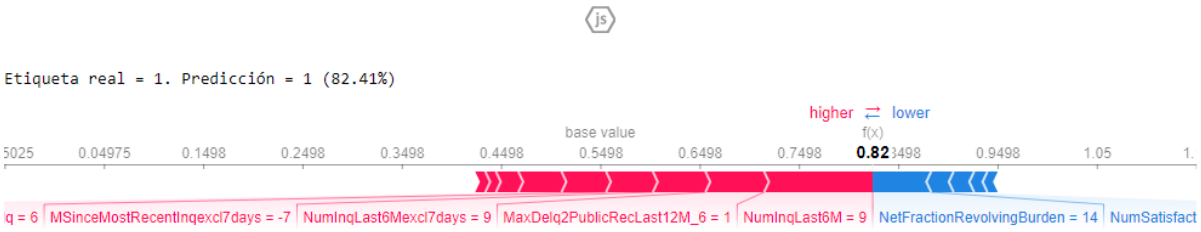
Como se puede apreciar las características más importantes para XGBoost no coinciden completamente con las seleccionadas en el caso de Red Neuronal, aunque aquellas que sí están en ambos modelos coinciden en el sentido de influencia, por ejemplo, en ambos modelos número de meses promedio en archivo aumenta la posibilidad de demora y el saldo renovable dividido por límite de crédito disminuye dicha probabilidad. En el caso de XGBoost utiliza la característica correspondiente al indicador consolidado de marcadores de riesgo,

ExternalRiskEstimate, dicha característica es un indicador obtenido a partir de otros indicadores y no daría información útil al cliente, lo cual se tendrá en cuenta a la hora de evaluar las explicaciones dadas por los distintos métodos.

### 4.3.5. Red Neuronal + SHAP

Para la obtención de la interpretación de una instancia con la técnica de SHAP, a partir del modelo de Red Neuronal, se utiliza el método Kernel de SHAP: KernelExplainer. Este método utiliza una regresión lineal ponderada para calcular la importancia de cada característica.

Para mostrar la información asociada a una instancia determinada, SHAP incorpora el gráfico de fuerza que muestra qué características contribuyen a impulsar la predicción. Tiene como referencia un valor base, que es el promedio del modelo sobre el conjunto de datos que se han indicado al realizar el análisis con SHAP.




Fuente: Elaboración propia a partir de la aplicación de SHAP para el modelo de Red Neuronal.

Figura 18. Red Neuronal, aplicación de SHAP para una instancia

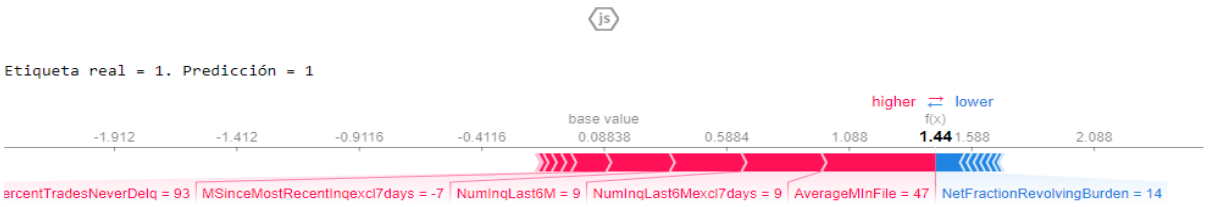
En la Figura 18 se muestra la gráfica de fuerza para una instancia, donde se puede identificar el valor base, en este caso es 0.5498, las características que influyen positivamente en la predicción, mostradas en color rojo y las que influyen negativamente, mostradas en color azul. Como se puede apreciar, la influencia de las características es mayoritariamente positiva, con un mayor color rojo, de ahí el desplazamiento a la derecha del valor de la instancia, 0.82, frente al valor base.

Las características que más influyen positivamente son el número de consultas en los últimos 6 meses, con un valor de 9, seguida de la característica del período de morosidad más largo de los últimos 12 meses, con un retraso de 30 días. En el sentido negativo, pero sin superar las anteriores está la característica de saldo renovable dividido por límite de crédito, con un valor de 14 y número de operaciones satisfactorias. En el gráfico de fuerza se puede apreciar la importancia de cada una de las características, y se ve de forma visual aquellas con más influencia.

### 4.3.6. XGBoost + SHAP

La interpretación con SHAP para el modelo de caja negra XGBoost se ha realizado a través del método Tree de SHAP, TreeExplainer. Este es un método rápido y exacto para estimación de los valores de SHAP para modelos de árboles y conjuntos de árboles.

En la Figura 19 se muestra la gráfica de fuerza asociada para una instancia a partir del modelo XGBoost. El valor de la instancia, 1.44, está desplazado a la derecha del valor base de 0.08838. Al igual que para el modelo de Red Neuronal, la influencia de las características para esta instancia es positiva, acercando la decisión con mayor probabilidad a la categoría *Bad*. Las características positivas corresponden a número de meses en promedio sin operaciones, con un valor de 47, seguido por el número de consultas en los últimos 6 meses (excluidos los últimos 7 días), y el número de consultas en los últimos 6 meses, ambas con un valor de 9. Estas dos últimas características están relacionadas y ambas aparecen influyendo en la decisión en el mismo sentido, lo cual muestra coherencia al resultado de la instancia. La característica que más influye de forma negativa, aunque sin poder contrarrestar las anteriores es el saldo renovable dividido por límite de crédito, con un valor de 14.




Fuente: Elaboración propia a partir de la aplicación de SHAP para el modelo XGBoost.

Figura 19. XGBoost, aplicación de SHAP a una instancia

## 4.4. Planteamiento de comparativa de interpretabilidad

Como se ha mostrado en el apartado 4.3, para cada una de las instancias se puede obtener su explicación en función de la aportación de cada característica a la predicción.

Para realizar la comparativa de los distintos modelos planteados, se consideran los factores relevantes identificados en relación con las características utilizadas para la toma de decisiones. Estos factores relevantes son:

- Cuántas unidades básicas contiene la explicación.
- Si las unidades básicas se corresponden con características en bruto o derivadas.

- Cuál es la estructura de dichas unidades, si se mantiene una estructura jerárquica entre ellas.
- Interrelaciones existentes entre las unidades básicas.

Dado un subconjunto de instancias del conjunto de test, se obtiene la aportación de las características en la explicación de su predicción y se analizan distintas métricas relacionadas con los factores relevantes. A continuación, se indican cuáles serán dichas métricas.

**1. Número de unidades básicas en la explicación**

Para cada una de las instancias se obtiene la métrica de número de unidades básicas por explicación, considerando aquellas características que han aportado en la explicación de la predicción de una instancia.

El criterio tomado para identificar si la característica ha aportado en cada uno de los modelos ha sido:

- Regresión Logística. Se considera que una característica $j$ ha aportado si el valor $\beta_j x_j$ es distinto de 0.
- Árbol de Decisión. Se considera que una característica $j$ ha aportado si forma parte del conjunto de características incluido en el camino hasta su nodo hoja.
- LIME: el tratamiento sería igual tanto para el modelo de Red Neuronal como el modelo XGBoost. A partir de los datos proporcionados por LIME para cada instancia y cada característica se considerará que una característica $j$ ha aportado si el valor LIME para esa característica multiplicado por el valor de la característica es distinto de 0.
- SHAP: el tratamiento es igual para ambos modelos de caja negra. A partir de los datos proporcionados por SHAP para cada instancia y cada característica se considerará que una característica $j$ ha aportado si el valor SHAP es distinto de 0.

**2. Características en bruto o derivadas**

Por la información facilitada en las especificaciones del dataset se conoce que la característica ExternalRiskEstimate es en sí una característica derivada, dado que es un indicador consolidado de marcadores de riesgo. Dicha característica, al ser un dato consolidado, no aporta información al cliente para la explicación de su predicción, por lo que se considera que una explicación que utilice esta característica no facilita su interpretación y por tanto es un punto de penalización a considerar en las explicaciones.

No se considera ninguna otra característica derivada. Aunque se ha realizado el tratamiento de dos de las características categóricas, no se consideran como características derivadas. En concreto, al haber desglosado dichas características en los valores determinados se pueden identificar con mayor facilidad cuál es el sentido de dicha característica.

Para cada instancia se obtiene la métrica de unidades derivadas. El valor de la métrica para la instancia será 1 si la característica derivada ha aportado a la explicación de la predicción, teniendo en cuenta el criterio de aportación indicado anteriormente, o será 0 en otro caso.

**3. Estructura de dichas unidades**

No se han identificado relaciones de jerarquía entre las características, por lo que no se considerará ninguna medida correspondiente a la relación jerárquica.

Asociado a este factor relevante se tendrá en cuenta, para cada instancia, si las categorías que aportan en la explicación corresponden con alguno de los valores especiales. Se considera que los datos faltantes no aportan valor a la instancia y, por lo tanto, las explicaciones basadas en dichos datos deben penalizarse. Se considera la métrica, por instancia, del número de valores especiales utilizados. Su valor será distinto de 0 si una o más características que aportan a la explicación tienen alguno de los valores especiales.

**4. Interrelaciones entre unidades**

A partir de la información proporcionada de las características del dataset se identifican ciertas relaciones entre ellas. En la Tabla 7 se muestran las relaciones identificadas, que formarán cada una un grupo y las características incluidas en cada uno de estos grupos.

Estas características relacionadas se tienen en cuenta si, para cada grupo, dos o más características aparecen en la explicación de una instancia. Si varias características del mismo grupo aportan a la explicación deberían hacerlo en el mismo sentido. En caso de que no sea igual el sentido de la aportación se considera que la explicación podría ser contradictoria.

El criterio tomado, en cada modelo, para identificar el sentido de cada característica del grupo ha sido:

- Regresión Logística. Signo de $\beta_j x_j$, cuando es mayor o menor que 0.
- Árbol de Decisión. Cada característica que aporta corresponde a un nodo en el árbol. Se considera el sentido de la característica en función de la clase del nodo en el árbol. Para obtener la clase del nodo se contabiliza el número de muestras de entrenamiento de cada clase que aplican a ese nodo y se asociará la clase con el mayor número de muestras.
- LIME: Signo del valor LIME multiplicado por el valor de la característica, mayor o menor que 0.
- SHAP: Signo del valor SHAP, cuando es distinto de 0.

Tabla 7.Agrupación de características interrelacionadas

Fuente: Elaboración propia a partir de la descripción del dataset HELOC (FICO community, 2018) agrupando las características por la relación entre ellas.

| **Grupos** | | | |
|---|---|---|---|
| **Cod** | **Relación** | **Nombre de Característica** | **Descripción** |
| **0** | **Antigüedad de las operaciones** | MSinceOldestTradeOpen | Número de meses que han transcurrido desde la primera operación. |
| | | AverageMInFile | Meses promedio sin operaciones. |
| **1** | **Operaciones atrasadas** | NumTrades60Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 60 días. |
| | | NumTrades90Ever2DerogPubRec | Número de operaciones que están atrasadas en más de 90 días. |
| **2** | **Operaciones morosas** | PercentTradesNeverDelq | Porcentaje de operaciones que no estaban en mora. |
| | | MSinceMostRecentDelq | Número de meses que han transcurrido desde la última operación morosa. |
| **3** | **Periodo de morosidad negativo** | MaxDelq2PublicRecLast12M_6 | Período de morosidad más largo de los últimos 12 meses (30 días de retraso). |
| | | MaxDelq2PublicRecLast12M_9 | Período de morosidad más largo de los últimos 12 meses (Todos los demás) |
| | | MaxDelqEver_2 | Período de morosidad más largo (Comentario negativo) |
| | | MaxDelqEver_5 | Período de morosidad más largo (60 días de retraso) |
| **4** | **Periodo de morosidad negativo** | MaxDelq2PublicRecLast12M_8 | Período de morosidad más largo de los últimos 12 meses (Actual y nunca moroso) |
| | | MaxDelqEver_8 | Período de morosidad más largo (Actual y nunca moroso) |
| **5** | **Operaciones a plazos** | PercentInstallTrades | Porcentaje de operaciones a plazos. |
| | | NetFractionInstallBurden | Fracción neta de la carga de los plazos. Saldo de la cuota dividido por el monto original del préstamo. |
| **6** | **Consultas de operaciones** | MSinceMostRecentInqexcl7days | Meses desde la última consulta (excepto los últimos 7 días). |
| | | NumInqLast6M | Número de consultas en los últimos 6 meses. |
| | | NumInqLast6Mexcl7days | Número de consultas en los últimos 6 meses (excluidos los últimos 7 días). |
| **7** | **Operaciones renovables** | NetFractionRevolvingBurden | Saldo renovable dividido por límite de crédito. |
| | | NumRevolvingTradesWBalance | Número de operaciones renovable con saldo. |

Habrá una medición independiente para cada agrupación y se obtendrá una métrica final que sea la media entre los datos de cada agrupación.

Tabla 8. Métricas de Interpretabilidad

Fuente: Elaboración propia. A partir de los factores relevantes se muestran las métricas de interpretabilidad.

| **Factor relevante** | **Métrica** |
|---|---|
| Número de unidades básicas en la explicación | Número de unidades básicas por explicación |
| Características en bruto o derivadas | Unidades derivadas |
| Estructura de unidades básicas | Valores especiales |
| Interrelaciones entre unidades | Agrupación con signo opuesto |

Como resumen, en la Tabla 8 se muestra, para cada factor relevante, las métricas que se aplicarán. Para cada instancia se medirán cada una de las métricas. Para cada modelo se obtendrá la media y la desviación asociada para las tres primeras métricas. En el caso de agrupación con signo opuesto se realizará un cálculo inicial por cada grupo. Se sumará el número de veces que aparecen características del mismo grupo con signo opuesto, se dividirá por el número de veces que aparecen varias características del mismo grupo y se promediará entre el número de muestras totales. Para obtener la métrica final para cada modelo se calculará la media y desviación asociada respecto a los distintos grupos existentes.

## 4.5. Criterios de éxito en precisión e interpretabilidad

Para determinar la elección de los modelos se considera tanto la precisión del modelo como por la precisión de la interpretabilidad. En función de las métricas identificadas se consideran los siguientes criterios para comparar los modelos.

**1. Métrica de precisión**

- *Área bajo la curva ROC (AUROC)*

  Se considera que un modelo es mejor cuanto mayor sea la precisión del modelo. Es deseable que los modelos tengan una precisión AUROC superior al 70%.

**2. Métricas de Interpretabilidad**

- *Número de unidades básicas por explicación*

  Se considera que los modelos con menor número de unidades básicas por explicación son más interpretables. No hay un valor óptimo para dicha métrica, pero para este trabajo, se considera que explicaciones con 5 o menos características son preferidas a aquellas con mayor número de características.

- *Unidades derivadas*

  Se considera que los modelos con menor número de explicaciones asociadas a la característica agrupada son preferibles a aquellos que basan todas las explicaciones en dicha característica. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.

- *Valores especiales*

  Se considera que los modelos cuyas explicaciones se basan en valores especiales no facilitan explicaciones fiables. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.

- *Agrupación con signo opuesto*

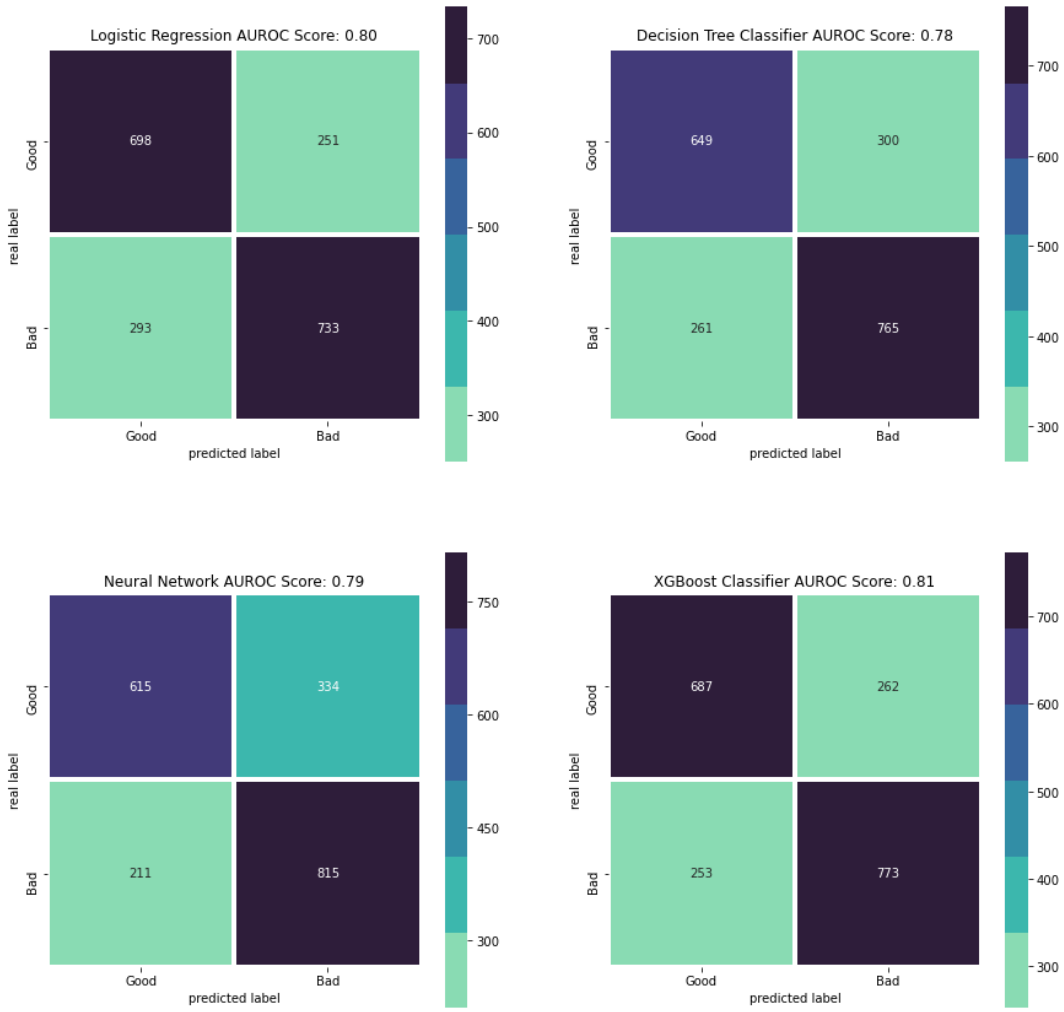
  Se considera que los modelos cuyas explicaciones contienen características del mismo grupo, pero con signos opuestos son contradictorios. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.

# 5. Desarrollo de la comparativa

Como se ha explicado en apartado 4.2, para cada uno de los modelos utilizados en la comparativa se ajustaron los parámetros asociados y se entrenaron los modelos en el conjunto de entrenamiento. Se utilizó el conjunto de test para obtener las medidas de precisión para cada uno de ellos.

Para tener el detalle de los casos predichos como verdaderos positivos ($tp$), verdaderos negativos ($tn$), falsos positivos ($fp$) y falsos negativos ($fn$) se muestran a continuación las matrices de confusión asociada a cada uno de los modelos entrenados.

En la Figura 20 se muestran las matrices de confusión de los modelos de Regresión Logística, Árbol de decisión, Red Neuronal y XGBoost, respectivamente. En la Tabla 9 se muestra la información del valor de la métrica AUROC para los cuatro modelos entrenados.




Fuente: Elaboración propia. Se muestran las matrices de confusión de los modelos para el conjunto de test.

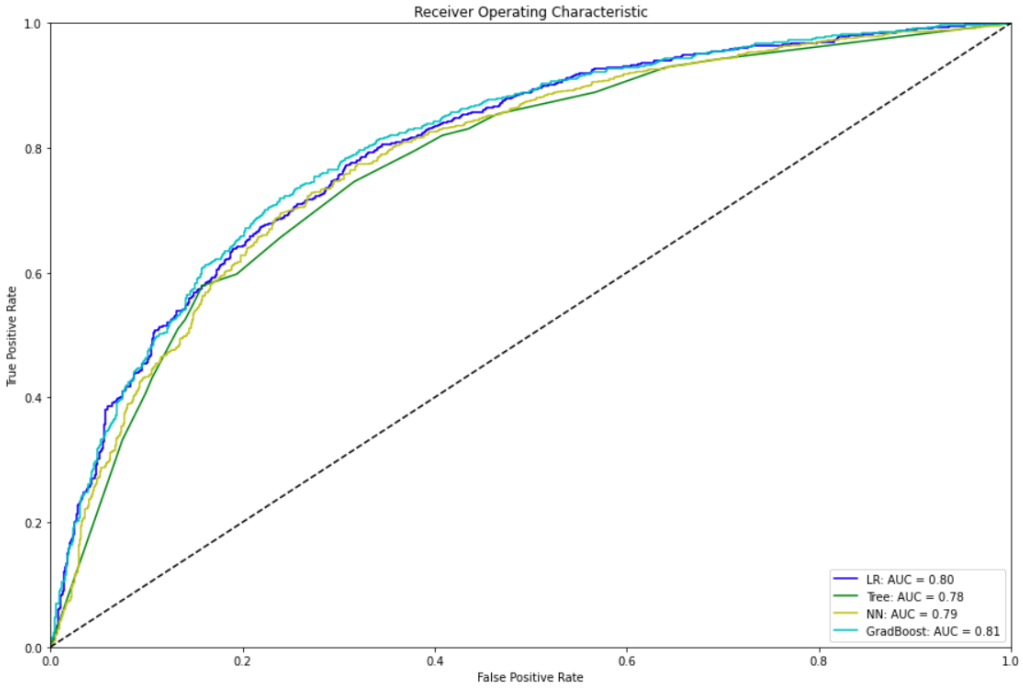
Figura 20. Matriz de confusión de los distintos modelos

Tabla 9. Métrica AUROC de los modelos

Fuente: Elaboración propia. Se muestra la métrica AUROC para cada uno de los modelos generados.

| | AUROC Score |
|---|---|
| **Regresión Logística** | 0.80 |
| **Árbol de Decisión** | 0.78 |
| **Red Neuronal** | 0.79 |
| **XGBoost** | 0.81 |

A partir de los datos en el conjunto de test se puede obtener la gráfica ROC para cada uno de los modelos. En la Figura 21 se muestran las curvas ROC de los cuatro modelos entrenados.




Fuente: Elaboración propia a partir de la ratio de los verdaderos positivos (tp) a medida que se incrementan los valores de los falso positivos (fp).

Figura 21. Curva ROC de los distintos modelos

Para obtener las métricas de interpretabilidad asociadas a las instancias se deben realizar los cálculos de las aportaciones de cada característica a la predicción. Debido al alto coste computacional para las técnicas de LIME y SHAP, no ha sido posible obtener los datos para todas las instancias y se ha reducido la muestra a las mil primeras instancias del conjunto de test.

En el caso de Regresión Logística y Árbol de Regresión la obtención de los datos de cada característica vendrá dados a partir del entrenamiento de los modelos.

En el caso de las Redes Neuronales y el clasificador XGBoost, a los cuales se les aplica las técnicas de LIME y SHAP, será necesario realizar el cálculo de la explicación para cada una de las instancias. Una vez obtenidas dichas explicaciones se podrán utilizar para obtener cada una de las métricas indicadas.

La primera de las métricas a calcular corresponde con el número de características utilizadas en la explicación de una instancia. LIME permite pasar como parámetro configurable el número de características que como máximo puede contener la explicación. Por lo tanto, dado que LIME tiene dicha facilidad, se han restringido la obtención de todas las explicaciones a un número máximo de 5 características. Esta restricción aplica tanto a la explicación de los resultados de la Red Neuronal como XGBoost. De esta forma se ajusta a una buena precisión de los modelos en la métrica de número de unidades básicas de explicación.

A continuación, se muestran los resultados obtenidos para cada una de las métricas de interpretabilidad consideradas. Para simplificar, se utiliza la notación LR, DT, CNN+LIME, XGB+LIME, CNN+SHAP, XGB+SHAP, para identificar a los modelos de Regresión Logística, Árbol de Decisión, Red Neuronal junto con LIME, modelo XGBoost junto con LIME, Red Neuronal aplicado SHAP, modelo XGBoost aplicado SHAP, respectivamente. En la Tabla 10 se muestran la media y varianza recogidos para las cuatro métricas de interpretabilidad utilizadas:

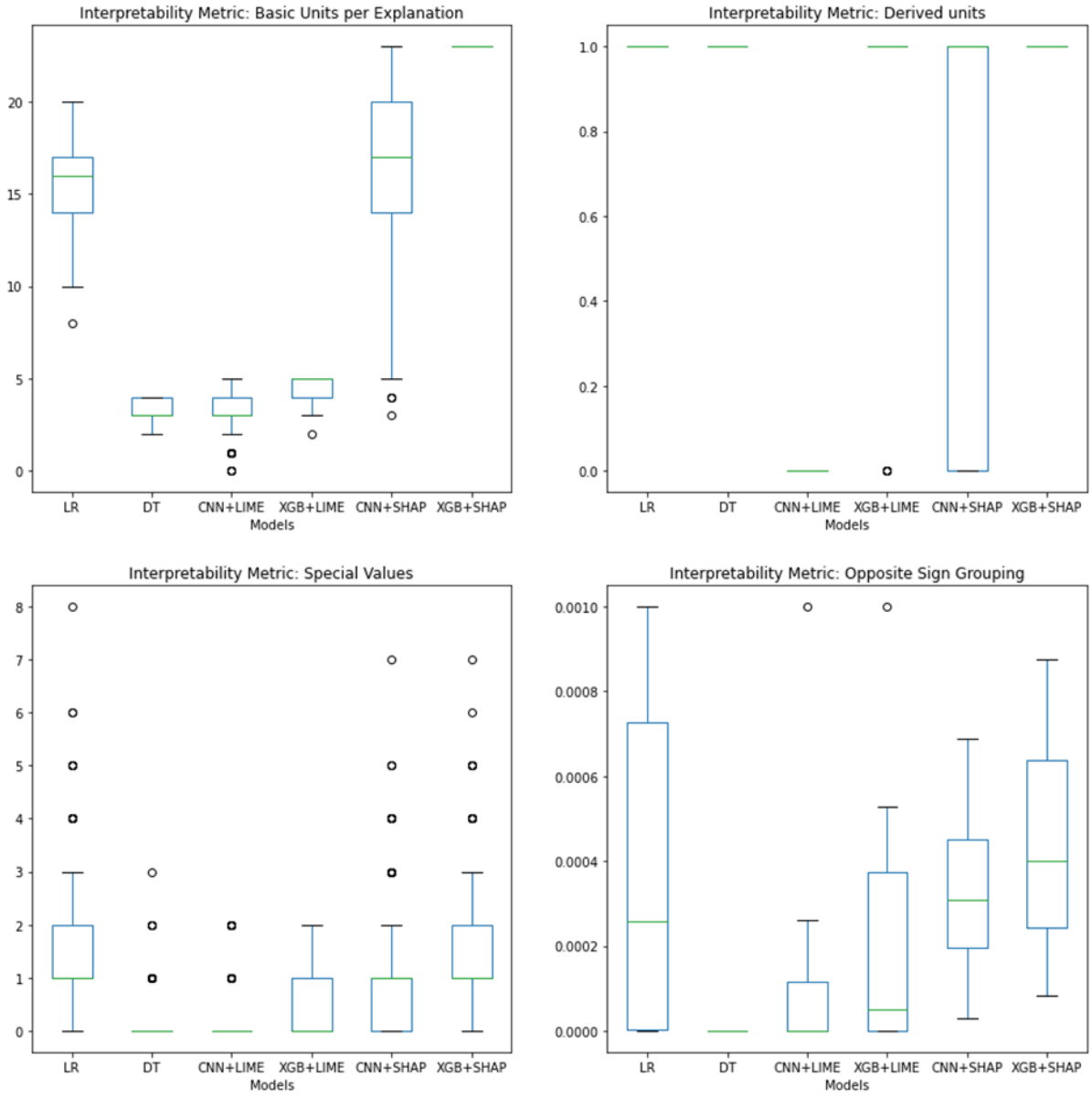
- *Número de unidades básicas por explicación (metric1)*
- *Unidades derivadas (metric2)*
- *Valores especiales (metric3)*
- *Agrupación con signo opuesto (metric4)*

Gráficamente los resultados se muestran en la Figura 22 donde se puede ver, en el diagrama de caja, para cada métrica la distribución de los datos para cada uno de los modelos seleccionados.

Tabla 10. Resultados obtenidos para las métricas de interpretabilidad

Fuente: Elaboración propia. Se muestran los resultados de las métricas de interpretabilidad para cada uno de los modelos.

| | LR | DT | CNN+LIME | XGB+LIME | CNN+SHAP | XGB+SHAP |
|---|---|---|---|---|---|---|
| **Metric1** | 15.61 +/- 1.76 | 3.22 +/- 0.49 | 3.25 +/- 1.02 | 4.47 +/- 0.57 | 16.69 +/- 3.85 | 23.0 +/- 0.0 |
| **Metric2** | 1.0 +/- 0.0 | 1.0 +/- 0.0 | 0.0 +/- 0.0 | 1.0 +/- 0.07 | 0.67 +/- 0.47 | 1.0 +/- 0.0 |
| **Metric3** | 1.24 +/- 1.03 | 0.18 +/- 0.41 | 0.07 +/- 0.3 | 0.27 +/- 0.46 | 0.99 +/- 0.94 | 1.23 +/- 1.02 |
| **Metric4** | 3.94e-04 +/- 0.0 | 0.00e+00 +/- 0.0 | 1.66e-04 +/- 0.0 | 2.44e-04 +/- 0.0 | 3.22e-04 +/- 0.0 | 4.38e-04 +/- 0.0 |




Fuente: Elaboración propia. Diagrama de cajas para mostrar los valores de cada métrica de interpretabilidad.

Figura 22. Gráfica del número de unidades básicas por explicación

El cálculo para la última métrica, correspondiente a la agrupación con signo opuesto, se ha realizado a partir de la información de cada uno de los grupos. La Tabla 11 muestra, para cada modelo y cada grupo, la información del número de veces que aparecen características con signo opuesto, dividido por el número de veces que aparecen varias características del grupo y promediado entre el número de muestras totales. Para obtener la métrica total por modelo se ha calculado la media y la desviación de los distintos grupos.

Tabla 11. Resultados de las agrupaciones con signo opuesto.

Fuente: Elaboración propia. Se muestra para cada modelo el resultado por grupo de características, teniendo en cuenta los signos de las características del mismo grupo. A partir de estos datos se obtiene la métrica de Agrupación con signo opuesto para cada modelo

| | LR | DT | CNN+LIME | XGB+LIME | CNN+SHAP | XGB+SHAP |
|---|---|---|---|---|---|---|
| **Group0** | 0.000022 | 0.0 | 0.000066 | 0.000098 | 0.000262 | 0.000271 |
| **Group1** | 0.001000 | 0.0 | 0.000000 | 0.000000 | 0.000060 | 0.000082 |
| **Group2** | 0.000495 | 0.0 | 0.000000 | 0.000529 | 0.000688 | 0.000692 |
| **Group3** | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000513 | 0.000876 |
| **Group4** | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000029 | 0.000159 |
| **Group5** | 0.000638 | 0.0 | 0.000000 | 0.000000 | 0.000353 | 0.000433 |
| **Group6** | 0.000993 | 0.0 | 0.001000 | 0.001000 | 0.000428 | 0.000618 |
| **Group7** | 0.000002 | 0.0 | 0.000261 | 0.000324 | 0.000240 | 0.000370 |

# 6. Discusión y análisis de resultados

Como se puede ver en la Tabla 9 de la comparativa realizada, la precisión de los modelos varía entre 78% del modelo del Árbol de Decisión hasta 81% en el modelo XGBoost.

Aunque se esperaba que la diferencia fuese más apreciable entre los modelos de caja negra y los modelos intrínsecamente explicables los resultados de precisión muestran una diferencia de tres puntos entre ellos.

En concreto, para el modelo de Red Neuronal, se han planteado varios modelos modificando la arquitectura con el número de capas densas, el número de neuronas en cada capa, los valores de dropout y el resto de los parámetros utilizados en la búsqueda exhaustiva. Ninguno de los modelos planteados ha superado los valores obtenidos para XGBoost. Este modelo es el que mejor precisión tiene, superando al Árbol de Decisión, lo que demuestra que la combinación de distintos clasificadores débiles permite obtener una mejor precisión. Es notable resaltar la precisión obtenida con el modelo de Regresión Lineal con un valor de 80%, superando a la precisión del modelo de Red Neuronal y cercano al obtenido por XGBoost.

Se analizan a continuación los resultados obtenidos por las métricas de interpretabilidad utilizadas en el trabajo.

- *Número de unidades básicas por explicación*

  Como se puede observar en los resultados mostrados en la Tabla 10, los modelos y técnicas con menor número de unidades básicas corresponden a Árbol de Decisión y los modelos de caja negra a los que se le ha aplicado la técnica de LIME. El número medio de atributos utilizado para la interpretación de una instancia es próximo a 3 para el caso de Árbol de Decisión y Red Neuronal con LIME. En el caso de XGBoost al cual se ha aplicado la técnica de LIME el número medio de características utilizadas para la interpretación es cercano a 4, redondeando al número entero más cercano. Este número reducido de características es posible por la configuración realizada para cada uno de ellos. En el Árbol de Decisión se configuró la profundidad máxima del árbol como 5. Del mismo modo se consideró un número máximo de características para la explicación con la técnica de LIME igual a 5. Estas configuraciones permiten reducir el número de características a aquellas con más peso en la decisión. Facilitando una explicación más clara y entendible, centrada en las características más relevantes.

  En contraposición con estos resultados, los obtenidos para el resto de los modelos y técnicas muestran cómo se utilizan alrededor de 16 características de media para la

interpretación de una instancia. En particular, el modelo XGBoost con la técnica de SHAP llega a utilizar 23 de las 24 características, solo una de ellas no cambia el valor predicho en ninguna combinación con otras características y, por lo tanto, según las propiedades de los valores de Shapley su aportación es igual a 0 (dummy). Si bien, como se mostró gráficamente en el caso de SHAP, aunque todas las características aportaran, no todas tenían la misma importancia en la decisión. Visualmente se podían identificar las características más importantes frente a otras con una aportación mucho más reducida. Aun así, todas ellas aportan a la explicación y no hay un punto de corte que pueda ser utilizado de forma común para todos los casos, que permita elegir las más representativas y reducir el número de características utilizadas de forma general. Por lo tanto, las interpretaciones en estos casos pueden llegar a ser menos claras.

- *Unidades derivadas*

  A partir de los datos de la Tabla 10 para la métrica de unidades derivadas se puede comprobar cómo todos los modelos salvo la Red Neuronal junto con LIME utilizan el indicador consolidado de marcadores de riesgo, ExternalRiskEstimate para sus explicaciones. Como se indicó anteriormente esta característica derivada es un marcador cuyo dato puede ser importante para el ajuste de los modelos, pero su inclusión en la interpretación de una instancia no aporta valor al prestatario que pide conocer el motivo de la decisión.

  El caso más claro en el que se puede ver la importancia de esta característica en la interpretación es el del Árbol de Decisión. El primer nodo del árbol es ExternalRiskEstimate, que además interviene en otros nodos internos para la clasificación de las instancias. La característica es la de mayor importancia, con un valor de 0.806 como se muestra en la Tabla 6, frente a la segunda característica en importancia, el número de meses promedio sin operaciones, con un valor de 0.064. Esto implica que la clasificación de los datos está altamente relacionada con el valor de la unidad derivada, aportando menor valor a la interpretabilidad de la decisión para una instancia.

- *Valores especiales*

  Los resultados de la métrica de valores especiales de la Tabla 10 muestran que los modelos de Regresión Lineal y XGBoost, junto con la técnica SHAP, utilizan en la explicación las características con valores especiales, la media en estos casos es de 1.23 para ambos. El modelo de Red Neuronal con la técnica de SHAP tiene un valor menor en esta métrica, pero también cercano a 1. Estos resultados tienen relación con

la primera de las métricas de interpretabilidad analizada, puesto que, como se indicó anteriormente, estos tres casos correspondían a los que mayor número de características utilizaban para la interpretación, existiendo mayor posibilidad de valores especiales entre las características utilizadas.

Para los modelos de Árbol de Decisión, y las cajas negras junto con la técnica de LIME se muestra cómo el valor medio de características con valores especiales utilizadas en la interpretación se reduce, siendo el mejor de los casos la Red Neuronal junto con LIME con un valor medio próximo a cero.

- *Agrupación con signo opuesto*

  De los datos mostrados en la Tabla 10 se obtiene que el mejor resultado para esta métrica es el modelo de Árbol de Decisión, cuyo valor es 0 en todos los grupos, esto es debido a que las características incluidas en el árbol de decisión no comparten grupo y por tanto no se presenta la opción de tener signos opuestos.

  La Red Neuronal junto con la técnica de LIME también tiene un comportamiento parecido, ya que en cinco de los ocho grupos su métrica es igual a 0. Esto tiene relación con el hecho de que el número de características utilizadas para la explicación se redujeron a cinco o menos.

  En contraposición el mayor valor de la métrica es para el modelo XGBoost y la técnica de SHAP, lo cual podría ser debido a que en este caso se utiliza una media de 23 características para la explicación, por lo que es más probable que, en ciertos casos, puedan darse diferencias de signos dentro de un grupo.

En función de las métricas anteriormente analizadas se podría concluir que, si bien el modelo de Red Neuronal no era el modelo con el más alto porcentaje de predicción frente al resto, al aplicar el método de LIME y teniendo en cuenta el resto de las métricas de interpretabilidad, mantiene una alta puntuación en cada una de las métricas, mejorando al resto de modelos y técnicas, permitiendo una compensación entre precisión e interpretabilidad.

# 7. Conclusiones y trabajo futuro

A continuación, se muestran las conclusiones obtenidas a partir del trabajo realizado, y se indican diversas líneas de trabajo futuro relacionadas con el mismo.

## 7.1. Conclusiones

A lo largo de este trabajo se ha mostrado la importancia que ha adquirido la interpretación de los modelos de ML, debido a que cada vez está más ampliamente extendido su uso y son más las personas impactadas por decisiones tomadas por estos modelos. En especial en los dominios de alto riesgo, como el financiero, donde el coste de realizar una predicción incorrecta es alto y se deben adaptar a los reglamentos actuales como GDPR.

Como se presentó en el apartado 2, los estudios sobre técnicas de XAI que se pueden aplicar en función del alcance de la interpretación, complejidad de los modelos o desideratas de los interesados, está ampliamente estudiado en la literatura, sin embargo, no hay un estudio tan exhaustivo ni un consenso sobre cómo medir dichas técnicas para evaluar la calidad de los métodos de explicación.

En este trabajo se ha planteado el problema de la elección del mejor modelo de ML para la predicción de riesgo crediticio, abordando dicho problema no solo desde la perspectiva de la precisión del modelo, sino también desde la interpretabilidad. Para ello, se han utilizado métricas de precisión altamente consensuadas para el análisis de los modelos y se han aportado métricas, basadas en las características, que han permitido la evaluación de la interpretabilidad.

Para realizar la comparativa se llevó a cabo un análisis del estado del arte en interpretabilidad y de las aplicaciones realizadas en el dominio financiero. Este análisis permitió identificar los ejes más representativos para tener en cuenta en la elección de los modelos y técnicas: taxonomía, partes interesadas y evaluación. Se seleccionó los modelos en función de su complejidad, utilizándose modelos intrínsecamente interpretables como Regresión Logística y Árbol de Decisión y modelos con mayor complejidad, llamados de caja negra, como son las Redes Neuronales y modelos que aplican técnicas de boosting como es XGBoost. Estos dos últimos modelos, debido a su naturaleza, no son interpretables, por lo que fue necesario utilizar técnicas de XAI. Entre los distintos enfoques atendiendo a las partes interesadas en la interpretabilidad, este trabajo se centra en los prestatarios interesados en conocer la justificación de la decisión. Este eje determinó el alcance local de las explicaciones, que

permita la justificación para la decisión de una instancia específica. Atendiendo al alcance local de las explicaciones se eligieron las técnicas agnósticas al modelo de LIME y SHAP para la interpretación de los modelos de caja negra. Estas técnicas de atribución de características aditivas han sido ampliamente utilizadas en la literatura, además SHAP tiene una base sólida basada en la teoría de juegos. Aun existiendo otras técnicas de interpretación local, como Anchors, LORE o explicaciones contrafácticas, ha sido necesario acotar el alcance del trabajo para garantizar la viabilidad de realización de este, siendo la utilización de estas técnicas una línea de ampliación futura.

Para la aplicación de los modelos explicables en el ámbito financiero se ha utilizado el dataset HELOC, con el conjunto de datos anónimos de solicitudes de línea de crédito con garantía hipotecaria. Previo a la utilización del dataset por los distintos modelos se realizó un análisis de los datos, que llevó a la limpieza de los datos y tratamiento de las variables categóricas. Se seleccionaron las características más importantes y finalmente, se realizó la división entre conjunto de entrenamiento y test.

Tras el ajuste de los datos, se entrenaron los modelos elegidos para la comparativa. Se analizaron y configuraron los parámetros asociados a cada una de las arquitecturas con el objetivo de obtener el mejor rendimiento. Una vez entrenados los modelos se obtuvieron los datos de precisión. Todos los modelos superaron el 70% en la métrica AUROC, consiguiéndose así el objetivo marcado respecto a la precisión de los modelos. Entre ellos, la menor precisión correspondió al Árbol de Decisión, con un 78%, seguido de la Red Neuronal, con 79%, Regresión Logística con un 80% y finalmente el modelo XGBoost con el mejor porcentaje de 81% de precisión.

Para la obtención de la justificación de una decisión específica, a partir de los datos de una instancia, se aplicaron las técnicas de interpretabilidad a los modelos de caja negra. Esto permitió mostrar la forma en la que se facilitaban las razones que explican la predicción realizada para una instancia, tanto para los modelos intrínsecamente explicables como para los modelos de caja negra con su técnica de interpretabilidad.

Se abordó entonces una de las partes importantes del objetivo general marcado, la evaluación de los modelos por su interpretabilidad y no solo por su precisión. A partir del proceso indicado por (Doshi-Velez & Kim, 2017) para definir y evaluar la interpretabilidad, se definieron los principios generales de la evaluación de la interpretabilidad de los modelos basados en:

- La necesidad de la interpretabilidad. La incompletitud de la formulación del problema, de generar modelos que sean justos en la decisión de conceder un préstamo, hacía

necesaria la interpretabilidad que permitiese evaluar si el modelo no discrimina y sigue las bases éticas.

- El nivel de la evaluación de la interpretación. De los tres niveles planteados por el autor, dos de ellos están basados en humanos, lo cual no permitió abordar dicho nivel de evaluación en este trabajo, centrándose en el nivel basado en la funcionalidad. Este nivel utiliza una definición formal de la interpretabilidad para medir la calidad de la explicación, sin requerir experimentos con humanos.

- Factores relevantes para la interpretación. Correspondientes a los factores para tener en cuenta para la evaluación de la interpretación local de las instancias. Se identificaron dichos factores como el número de unidades básicas que contiene la explicación, si son características en bruto o derivadas, la estructura o jerarquía que mantienen entre ellas y las interrelaciones existentes. Se definió el criterio de aportación de una característica a la explicación y se utilizó para calcular el número de unidades básicas por explicación. Se tomó el indicador consolidado de marcadores de riesgo como única característica derivada. Tras el análisis de los datos y características, no se identificaron relaciones de jerarquía entre las unidades básicas. Sí se consideró relevante las unidades con valores especiales en las interpretaciones. Se identificaron las agrupaciones de las características relacionadas, y se definió el criterio de sentido de aportación.

Siguiendo los pasos indicados por el autor e identificando los factores, se llegó a la definición de las métricas para la evaluación de las interpretaciones, lo que permitió evaluar los modelos por su interpretabilidad. Esta es una aportación significativa frente a otros estudios realizados en la interpretabilidad de los modelos, incluidos el financiero, dado que no se limita a la aplicación de una técnica, sino que permite evaluar cuál sería el mejor modelo interpretable para la tarea en cuestión.

La aplicación de las métricas a las distintas interpretaciones de los modelos ha permitido comprobar que la técnica LIME aplicada al modelo de Red Neuronal obtiene una alta puntuación en todas las métricas de interpretabilidad. Si bien la precisión de este modelo no era la más alta, al considerar ambos criterios de precisión e interpretabilidad, permiten seleccionarlo como el modelo más interpretable de los analizados. Esto confirma que no siempre el modelo más preciso se ajustará mejor a las necesidades de interpretación necesarias para dominios de alto riesgo, como el financiero y que es necesario mantener un equilibrio entre ambos criterios adaptados a las necesidades.

## 7.2. Líneas de trabajo futuro

Como futuras líneas de ampliación de este trabajo se propone analizar otras técnicas de interpretabilidad locales, que puedan ser comparadas con las incluidas en este trabajo y que

permitan analizar el comportamiento de los modelos en las métricas de interpretabilidad definidas.

Adicionalmente se podría ampliar el nivel de evaluación de la interpretación. Como se ha indicado, este trabajo se centró en el nivel basado en la funcionalidad. Para dar más robustez a las métricas, se podría estudiar si los resultados obtenidos en el trabajo se confirman en evaluaciones basadas en aplicaciones y basadas en humanos, permitiendo así afianzar dichas métricas o por el contrario analizar los resultados para mejorar las mismas.

En última instancia se plantea ampliar la búsqueda de nuevos factores relevantes asociados a los modelos, que permita ampliar las métricas de evaluación de interpretabilidad. El objetivo sería obtener métricas comunes, independientes de los modelos y dominios de aplicación, que pudieran adoptarse de forma general, no solo en el dominio financiero y que permitieran la evaluación de la interpretabilidad de los modelos.

# 8. Bibliografía

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3-31. https://doi.org/10.1016/S0167-7012(00)00201-3

Benítez, R., Escudero, G., Kanaan, S., & Maship Rodo, D. (2013). *Inteligencia artificial avanzada*. Editorial UOC, S.L.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine Learning Explainability in Finance: An Application to Default Risk Analysis* (SSRN Scholarly Paper ID 3435104). Social Science Research Network. https://doi.org/10.2139/ssrn.3435104

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, *57*(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An Interpretable Model with Globally Consistent Explanations for Credit Risk. *arXiv:1811.12615 [cs, stat]*. http://arxiv.org/abs/1811.12615

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *2016 IEEE Symposium on Security and Privacy (SP)*, 598-617. https://doi.org/10.1109/SP.2016.42

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. http://arxiv.org/abs/1702.08608

European Commission. Joint Research Centre. (2020). *Robustness and explainability of Artificial Intelligence: From technical to policy solutions.* Publications Office. https://data.europa.eu/doi/10.2760/57493 Último acceso: 22/07/2021

FICO community. (2018). *Explainable Machine Learning Challenge*. https://community.fico.com/s/explainable-machine-learning-challenge Último acceso: 22/07/2021

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *Boletín de exploración de ACM SIGKDD de*, *15*(1), 1-10. https://doi.org/10.1145/2594473.2594475

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80-89. https://doi.org/10.1109/DSAA.2018.00018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 93:1-93:42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [cs, stat]*. http://arxiv.org/abs/1705.07874

Molnar, C. (2021). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. bookdown. https://christophm.github.io/interpretable-ml-book/ Último acceso: 22/07/2021

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986. https://doi.org/10.1016/j.eswa.2020.113986

Munkhdalai, L., Ryu, K. H., Namsrai, O.-E., & Theera-Umpon, N. (2021). A Partially Interpretable Adaptive Softmax Regression for Credit Scoring. *Applied Sciences*, *11*(7), 3227. https://doi.org/10.3390/app11073227

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

Nguyen, A., & Martínez, M. R. (2020). On quantitative aspects of model interpretability. *arXiv:2007.07584 [cs, stat]*. http://arxiv.org/abs/2007.07584

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), Pub. L. No. 32016R0679, 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng Último acceso: 22/07/2021

Reina, J. L. R. (2018). Evaluación de modelos. *Universidad de Sevilla. Razonamiento Asistido por Computador (2018-19)*, *Tema 7*, 40.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*. http://arxiv.org/abs/1602.04938

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv:1811.10154 [cs, stat]*. http://arxiv.org/abs/1811.10154

Shapley, L. S. (2016). 17. A Value for n-Person Games. En *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307-318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *arXiv:1711.00399 [cs]*. http://arxiv.org/abs/1711.00399

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), 593. https://doi.org/10.3390/electronics10050593

# Anexos

## Anexo I. Repositorio de código fuente

El código realizado para desarrollar la comparativa mostrada en este trabajo se ha publicado en el repositorio de GitHub.

El trabajo está estructurado siguiendo los siguientes apartados:

1. Procesamiento de datos
2. Arquitectura de los modelos
3. Interpretabilidad de los modelos
4. Métricas de interpretabilidad

Acceso al repositorio: https://github.com/sortegamart/TFM_ML_Interpretable

## Anexo II. Artículo de investigación

Se incluye el artículo de investigación asociado al trabajo realizado sobre la comparativa de modelos de Machine Learning interpretables para la predicción de riesgo crediticio.

# Comparativa de modelos de Machine Learning interpretables para la predicción de riesgo crediticio

Sonia Ortega Martín

Universidad Internacional de la Rioja, Logroño (España)

22/07/2021

## Resumen

Los sistemas de Inteligencia Artificial son ampliamente utilizados e impactan en la vida diaria de las personas de forma creciente, en gran medida debido al avance en Machine Learning y su alta precisión. Los modelos suelen ser opacos y dificultan el entendimiento de su lógica. Su interpretabilidad se hace más necesaria, sobre todo en dominios de alto riesgo. Este trabajo se ha centrado en el dominio financiero, en la tarea predicción de riesgo crediticio, sobre el conjunto de datos de solicitudes de crédito con garantía hipotecaria (HELOC). Se han planteado distintos modelos y aplicado técnicas para obtener modelos interpretables. Se han definido métricas de interpretabilidad, que permiten la comparativa de modelos atendiendo a criterios de interpretabilidad y no únicamente de precisión. Los resultados obtenidos muestran que la elección de un modelo no solo debe estar basado en la precisión, sino que debe mantener un equilibrio entre precisión e interpretabilidad.

## Palabras Clave

Inteligencia Artificial, Machine Learning, Interpretabilidad, Métricas de Interpretabilidad, Riesgo crediticio.

## I. Introducción

Los avances en Machine Learning (ML) y Deep Learning (DL) han propiciado un aumento en el uso de modelos en distintas áreas y del número de personas afectadas por la implantación de estos sistemas. En ciertos dominios de alto riesgo, donde el coste de la predicción incorrecta es muy alto, se hace esencial la interpretabilidad. En estos dominios, como el financiero, los enfoques de Inteligencia Artificial Explicable (XAI) tienen alto potencial, proporcionando información que permita confiar en el resultado obtenido. Debiendo existir una compensación entre precisión e interpretabilidad.

Como primer paso, para comprender el significado de interpretabilidad, se hace un resumen de las definiciones dadas por algunos autores. Debido al carácter subjetivo del término, no hay una definición estándar y globalmente aceptada para la interpretabilidad, utilizándose a veces indistintamente los términos de interpretable y explicable.

Según [1] a través de la interpretabilidad del sistema se puede explicar su razonamiento de forma comprensible al ser humano y así se podría verificar si es sólido. [2] considera que "La explicabilidad está estrechamente relacionada con el concepto de interpretabilidad: los sistemas interpretables son explicables si sus operaciones pueden ser entendidas por el ser humano." Para [3] las explicaciones deben ser competas e interpretables y [4] considera que el objetivo de la evaluación de la calidad de las explicaciones está en medir hasta qué punto se satisfacen las características de fidelidad e interpretabilidad. Por su parte [5] concluye que interpretabilidad y explicabilidad no se implican mutuamente, asociando a la interpretabilidad la capacidad de identificar la relación entre entradas y salidas del sistema y la interpretabilidad a la lógica interna del sistema de aprendizaje automático. [6] considera la interpretación como la extracción de conocimiento del dominio a través de un modelo que ha aprendido las relaciones contenidas en los datos. El aprendizaje automático interpretable proporciona un conocimiento relevante sobre el problema del dominio en cuestión

Según [7] con la interpretabilidad se puede comprender el mecanismo interno del sistema a la vez que se demuestra si se ajusta a las especificaciones y cumple con las normas éticas. Esto estaría alineado con los cuatro factores que [2] considera para la necesidad de la explicabilidad: justificar los resultados y decisiones, controlar para detectar vulnerabilidades, mejorar continuamente los modelos y descubrir aprendiendo nuevos hechos. La interpretabilidad también es necesaria por razones legales como la aplicación del Reglamento General de Protección de Datos (GDPR) [8], ya que en los casos de decisiones automatizadas se deberá facilitar información significativa de la lógica aplicada y sus consecuencias.

Se considera que la interpretabilidad permite comprender la relación existente entre las entradas y salidas presentándose de forma sencilla y entendible por el humano y así comprobar la equidad de los modelos e identificar y corregir posibles sesgos. Así se fomenta la confianza, una mayor credibilidad en el modelo y el cumplimiento de la legislación, que puede llevar a la elección de los modelos interpretables frente a los que no lo son, sobre todo en aquellos dominios de alto riesgo.

Todo lo expuesto permite tener una idea de la importancia que ha tomado la interpretabilidad en los sistemas de Inteligencia Artificial. Eso ha llevado a un creciente número de publicaciones de investigación sobre interpretación de modelos de ML.

Este trabajo se centra en el dominio financiero y en concreto en la predicción de riesgo crediticio. Se estudiarán y aplicarán técnicas que permitan obtener un modelo interpretable, cuya información sustente el resultado predictivo del mismo. Los modelos obtenidos se compararán en función de la precisión e interpretabilidad. Las métricas para determinar la precisión predictiva están ampliamente estudiadas y consensuadas como el área bajo la curva ROC (AUROC). Sin embargo, en la literatura no se encuentra el mismo consenso sobre las métricas que permitan determinar la interpretabilidad de un modelo. Este trabajo se centra en el proceso expuesto por [1] para definir y evaluar la interpretabilidad. Para ello, se definen los principios generales sobre los que se basa la evaluación de la interpretabilidad, en respuesta a: la necesidad de la interpretabilidad debida a la incompletitud de la formulación del problema, el nivel al que se realiza la evaluación de la interpretación y los factores relevantes para la interpretación.

## II. Estado del Arte

En los últimos años ha habido un aumento en la investigación sobre la interpretabilidad de los modelos, en parte debido a iniciativas como la lanzada por la Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA) de Estados Unidos [9], donde se planteaban tres desafíos de investigación: producir modelos más explicables, diseñar interfaces de explicación y comprender los requisitos psicológicos para explicaciones más efectivas. A medida que estas investigaciones aumentan también lo hacen las clasificaciones de las técnicas utilizadas. En este apartado, se analizan las taxonomías o clasificaciones más representativas, aunque no exhaustivas, de las técnicas para el aprendizaje automático interpretable.

La clasificación de las técnicas de interpretación en aprendizaje automático no debe ser unilateral, como indica [5], ya que existen diferentes puntos de vista y aspectos a la hora de estudiar los métodos de interpretabilidad. Según [6] la elección del enfoque de explicabilidad deberá tener en cuenta características esenciales como la naturaleza del problema, el contexto y la audiencia. La revisión realizada por [2] de las obras más relevantes en la literatura se organiza en cuatro ejes principales: la taxonomía de los métodos de interpretabilidad, la medición y evaluación de las explicaciones, la figura del humano en el bucle y el equilibrio entre explicar y predecir.

En su gran mayoría los estudios realizados sobre explicabilidad se centran en el primero de estos ejes, la búsqueda de métodos, procedimiento y estrategias para la explicación de los modelos y la realización de una clasificación de estos. En comparación con este eje, los otros tres mencionados son menos representativos en la literatura. A continuación, se desarrollan los tres primeros: taxonomía, la figura del humano en el bucle y medición y evaluación.

### *Taxonomía*

Los enfoques más aceptados para la clasificación de los métodos y procedimientos son las realizadas en función del alcance, complejidad y dependencia del modelo.

- *Alcance de las explicaciones*. Pudiendo identificarse como local o global. Un alcance local proporciona información sobre predicciones individuales del modelo, mostrando una porción del comportamiento del sistema. Entre las formas existentes se puede realizar a través del análisis de las características con más peso en la toma de decisión o bien utilizando explicaciones contrafácticas [10]. La propuesta de [11] para la explicación local del modelo es LIME. Por su parte, [12] introdujeron una técnica de enfoque local con sólida base teórica llamada SHapley Additive exPlanation, SHAP. Un alcance global proporciona información general del modelo, con el objetivo de explicar la lógica de este que permita un conocimiento del razonamiento de todos los resultados. Este enfoque se realiza a través de modelos simples e interpretables que aproximan a un modelo más complejo. La información producida por estos modelos puede requerir un conocimiento previo del dominio para su comprensión.

- *Complejidad del modelo utilizado para la predicción*. Se distinguen entre interpretables intrínsecos y modelos de caja negra interpretable. Los primeros, por su naturaleza, son fácilmente comprensibles por los humanos. Se incluiría en este grupo los árboles de decisión, modelos basados en reglas y aproximaciones lineales [11][13][14]. Estos modelos pueden tener un problema de precisión, aun así, estudios como [15] abogan por la utilización de este tipo de modelos, en especial en dominios de alto riesgo. Los modelos de caja negra interpretable parten de un modelo base no interpretable y de alta complejidad y utilizan técnicas que permiten extraer información explicativa del modelo. [16] divide la explicación de caja negra, en función del enfoque de la explicación: explicación del modelo, dando una visión global del comportamiento de la caja negra utilizando para ello un modelo interpretable, explicación del resultado, interpretando la predicción de una instancia, e inspección del modelo como punto intermedio entre las anteriores, para una representación de una propiedad específica del modelo o de sus predicciones.

- *Dependencia del modelo*. Considerando si los métodos utilizados para la interpretación son o no independientes del modelo. Se clasifican en métodos agnósticos o específicos del modelo. En los primeros las explicaciones dependen de la observación de entrada y salida y funcionan para cualquier tipo de modelo. Los segundos tienen en cuenta características propias del modelo que están analizando y no podrían ser utilizados con otro tipo de modelo inicial.

### *Figura del humano en el bucle*

En relación con el eje de la figura del humano y sus intereses sobre la explicación, [17] enfatiza el papel de las partes interesadas y cómo se debe satisfacer sus desideratas de comprensión. Identifica varios grupos de interesados y sus deseos a ser cubiertos: usuarios que utilizan las recomendaciones para toma de decisiones, cuyos deseos para la explicabilidad sería la usabilidad y confianza. Desarrolladores que diseñan, programan y construyen los sistemas artificiales, los principales intereses de este grupo son verificación y rendimiento. Partes afectadas por las decisiones de los sistemas, un grupo cada vez más amplio debido al gran crecimiento de decisiones automatizadas. Cuyo interés en la explicabilidad sería la equidad y la ética de los modelos. Implantadores de un sistema cuyos intereses fundamentales serían que el sistema implantado por ellos sea aceptado y que cumpla con las legislaciones. Reguladores que estipulan las normas legales y éticas, para los cuales la explicabilidad de los modelos es una forma de facilitar la legalidad, ética y robustez de los sistemas.

### *Medición y evaluación*

Otro eje a considerar de la interpretabilidad es el de la evaluación de efectividad de las explicaciones, validando y

comparando para cuantificar la mejora que representan. [6] establece un marco predictivo, descriptivo y relevante (PDR) para seleccionar y evaluar los métodos de interpretación. Entendiendo la predicción predictiva como la capacidad del modelo para aproximar las relaciones de los datos, y la predicción descriptiva como la capacidad de las interpretaciones para explicar lo aprendido por el modelo. El marco considera que la interpretación confiable maximiza la predicción predictiva y descriptiva, además de ser relevante para una audiencia de un dominio particular. Las mejoras en la precisión predictiva son fáciles de medir utilizando métricas de precisión, [18][19], mientras que no hay un protocolo de evaluación estándar para evaluar las mejoras en precisión descriptiva o relevancia. [1] propone un enfoque de evaluación con tres categorías: basada en aplicaciones, basada en humanos y basada en funciones. Los dos primeros requieren experimentos con humanos. La evaluación basada en aplicaciones se realiza con expertos en el dominio, para evaluar la calidad de la explicación en el contexto en cuestión. La evaluación basada en humanos se realiza con humanos legos, para probar la calidad de la explicación respecto a nociones más generales. La evaluación basada en la funcionalidad utiliza una definición formal de interpretabilidad para medir la calidad de la explicación, sin requerir experimentos con humanos.

A partir de este enfoque [20] se centran en la evaluación basada en funciones y consideran que las explicaciones poseen tres aspectos cuantitativos que pueden ser medidos objetivamente: fidelidad, sencillez y amplitud. En un estudio posterior [4] analiza tanto las evaluaciones basadas en experimentos con humanos como las evaluaciones basadas en funciones. Para las primeras no existen criterios consensuados y es difícil la comparación. Para las segundas se basa en las características de las explicaciones: interpretabilidad y fidelidad.

En el dominio financiero [21] desarrolla un marco analítico para abordar el problema de la explicabilidad, mediante el estudio de las entradas y salidas. Identifica cinco tipos de explicaciones significativas y los relaciona con las partes interesadas en la explicación. Compara, para cada tipo de explicación significativa, el comportamiento de los modelos de Regresión Logística y Gradient Tree Boosting (GTB), con la técnica de explicabilidad Quantitative Input Influence (QII) [22]. Para la medición del riesgo de créditos en plataformas peer to peer (P2P) [23] compara modelos en función de su precisión predictiva, empleando posteriormente una técnica de XAI que consigue la explicabilidad. Se selecciona el modelo Gradient Boosting (XGBoost) [24][25] y utiliza valores Shapley [26] para explicar la contribución de cada una de las variables explicativas. En su estudio comparativo para plataformas P2P, [27] elige tres modelos con la mejor precisión predictiva y aplica técnicas XAI, adicionalmente evalúa los modelos en términos de explicabilidad, según el protocolo descrito en [11]. Para la calificación crediticia [28] propone un nuevo modelo, PIA-Soft, que combina modelos de red neuronal y regresión softmax, comparando dicho modelo con otros como Regresión Logística o modelos de caja negra. En un enfoque de utilizar modelos intrínsecamente interpretables [29] presenta un modelo de riesgo aditivo en dos capas y una herramienta de visualización interactiva, demostrando que no siempre son necesarios los modelos de caja negra para la evaluación de riesgo crediticio.

## III. Objetivos y Metodología

El objetivo general del trabajo es realizar una comparativa de distintos modelos de ML interpretables aplicados al dominio financiero en la tarea de la predicción de riesgo crediticio. En la evaluación de dichos modelos se tendrá en cuenta no solo la precisión de los modelos sino también la interpretabilidad. Es deseable que los modelos tengan una precisión superior al 70% y además que permitan justificar la decisión específica, dando información de las razones por las que se realizaron las predicciones, a partir de los datos de entrada. Para evaluar los modelos por su interpretabilidad se utilizarán métricas basadas en las explicaciones. Es deseable que los modelos mantengan una alta puntuación en dichas métricas de interpretabilidad. Para realizarlo se definen los siguientes objetivos específicos:

- Revisar el estado del arte en interpretabilidad respecto a tres ejes de clasificación, partes interesadas y evaluación.
- Revisar los trabajos más destacados para la aplicación de la interpretabilidad en el dominio financiero.
- Explorar los modelos intrínsecamente interpretables y las técnicas XAI.
- Analizar el dataset para la predicción del riesgo crediticio.
- Implementar los modelos a comparar y aplicar las técnicas de interpretabilidad a los modelos de caja negra.
- Evaluar los modelos atendiendo a la precisión predictiva y a la interpretabilidad de estos.
- Comparar los resultados obtenidos por los diferentes modelos y analizarlos para obtener las conclusiones.

Para alcanzar los objetivos fijados se utilizará una metodología de trabajo basada en las siguientes fases:

1. Realizar un amplio estudio de la literatura existente sobre la interpretabilidad.
2. Revisión y selección de los modelos a utilizar en la comparativa.
3. Análisis de dataset para la evaluación del riesgo crediticio.
4. Implementación de los modelos y aplicación de técnica XAI a modelos de caja negra.
5. Ejecución de modelos y estudio de resultados aplicando métricas de precisión e interpretabilidad.

## IV. Contribución

Con el objetivo de realizar la comparativa de distintos modelos de ML interpretables atendiendo no solo a la precisión sino a la interpretabilidad se presentan a continuación el planteamiento realizado para la obtención de los resultados.

### *Conjunto de datos*

Para la comparativa se ha seleccionado el dataset Home Equity Line of Credit (HELOC) [30], que proporciona información sobre las solicitudes de línea de crédito con garantía hipotecaria realizadas por propietarios reales. Las variables predictoras son todas cuantitativas o categóricas y provienen de datos anónimos de agencias de crédito. La variable a predecir es una variable binaria, cuyo valor *Bad* indica que el solicitante tuvo un retraso de 90 días o más durante un periodo de 2 años desde que abrió la cuenta de crédito. Un valor *Good* indica que ha realizado los pagos con menos de 90 días de atraso. El modelo entrenado en este conjunto de datos puede predecir si un prestatario reembolsará su crédito dentro de 2 años. El procesamiento del dataset se dividió en varias etapas: limpieza de datos, tratamiento de características categóricas, selección de características y división en train y test.

A partir de las especificaciones del dataset se identifican 588 instancias con el valor especial -9, 'Sin registro en la oficina o sin investigación', en todas sus características predictoras, por lo que

carecen de valor para la toma de decisiones. Por ese motivo se eliminan del conjunto original. Se analizan las instancias con otros valores especiales, -7, 'Condición no satisfecha (por ejemplo, no hay consultas, no hay morosidad)' y -8, 'No hay operaciones o investigaciones utilizables/válidas', se llega a la conclusión que no es posible sustituir estos valores por otros sin alterar el sentido de las muestras y se decide mantener las instancias sin cambios, teniéndose en cuenta posteriormente en la evaluación de los modelos.

Existen dos variables categóricas correspondientes al periodo de morosidad más largo y el más largo de los últimos 12 meses. Se unifican los códigos en ambas características en función de su significado y se realiza una conversión hot encoding a variables dummy para el correcto tratamiento de los valores categóricos por los modelos.

Se calculan las correlaciones entre las distintas características predictoras y la variable dependiente y se calcula la puntuación asociada a la importancia de las características, seleccionando aquellas que son más representativas, por encima de un umbral. Reduciendo el número de variables predictoras a un total de 24 características.

Finalmente se realiza una división de los datos en los conjuntos de entrenamiento y test, con una distribución del 80% - 20%.

### *Arquitectura de los modelos*

Con el objetivo de comparar modelos intrínsecamente interpretables y modelos de caja negra se seleccionan los modelos de Regresión Logística, Árboles de Decisión del primer tipo y Red Neuronal y XGBoost del segundo.

Para cada modelo se llevan a cabo las etapas de entrenamiento, búsqueda de parámetros, y validación. Partiendo de un modelo inicial básico se utiliza la técnica de GridSearchCV de sklearn, que permite hacer una búsqueda exhaustiva de valores para ciertos parámetros del modelo. Se construyen los distintos modelos con dicha configuración y se realiza la validación en el conjunto de test, obteniéndose para cada uno la matriz de confusión y el valor de área bajo la curva ROC.

### *Comparativa de interpretabilidad*

Para la interpretación de las instancias en cada uno de estos modelos ha sido necesaria la aplicación de técnicas XAI a los modelos de caja negra. Las técnicas elegidas corresponden con LIME y SHAP. Una vez aplicadas se analiza explicabilidad a nivel de instancia para los distintos casos de Regresión Logística, Árbol de Decisión, Red Neuronal con LIME, Red Neuronal con SHAP, XGBoost con LIME y XGBoost con SHAP.

Para realizar la comparativa de interpretabilidad de los distintos modelos interpretables planteados se consideran los factores relevantes identificados en relación con las características utilizadas para la toma de decisiones. Estos factores relevantes son: cuántas unidades básicas contiene la explicación, si las unidades básicas se corresponden con características en bruto o derivadas, cuál es la estructura de dichas unidades, si existe una estructura jerárquica entre ellas y las interrelaciones existentes entre las unidades básicas.

Dado un subconjunto de instancias del conjunto de test, se obtiene la aportación de las características en la explicación de su predicción y se definen métricas relacionadas con los factores relevantes.

- *Número de unidades básicas en la explicación*. Para cada una de las instancias se obtiene la métrica de número de unidades básicas por explicación, considerando aquellas características que han aportado en la explicación de la predicción de una instancia. El criterio tomado para identificar si la característica ha aportado en cada uno de los modelos ha sido:
    - Regresión Logística. Se considera que una característica aporta si la multiplicación del coeficiente por el valor de la característica es distinta de 0.
    - Árbol de Decisión. Se considera que una característica ha aportado si forma parte del conjunto de características incluido en el camino hasta su nodo hoja.
    - LIME: el tratamiento es igual para ambos modelos de caja negra. A partir de los valores de LIME para cada instancia y cada característica se considerará que una característica ha aportado si el valor LIME por el valor de la característica es distinto de 0.
    - SHAP: el tratamiento es igual para ambos modelos de caja negra. A partir de los datos proporcionados por SHAP para cada instancia y cada característica se considerará que una característica ha aportado si el valor SHAP es distinto de 0.
- *Características en bruto o derivadas*. A partir de las especificaciones del dataset se identifica una de las características como dato derivado, ya que es un indicador consolidado de marcadores de riesgo. Este dato no aporta información al prestatario para la explicación de su predicción. Se considera que una explicación que utilice esta característica no facilita su interpretación y es, por tanto, un punto de penalización a considerar en las explicaciones. Para cada instancia se obtiene la métrica de unidades derivadas, cuyo valor será 1 si la característica derivada ha aportado a la explicación de la predicción y 0 en otro caso, teniendo en cuenta el criterio de aportación indicado anteriormente.
- *Estructura de las unidades*. No se ha identificado relaciones de jerarquía entre las características y no se considera ninguna métrica relativa a la jerarquía. Asociado a este factor relevante se tiene en cuenta, para cada instancia, si las categorías que aportan en la explicación corresponden con alguno de los valores especiales. Se considera que los datos faltantes no aportan valor a la instancia y, por lo tanto, las explicaciones basadas en dichos datos deben penalizarse. Se considera la métrica de valores especiales utilizados, su valor será distinto de 0 si una o más características que aportan a la explicación tienen alguno de los valores especiales.
- *Interrelaciones entre unidades*. A partir de la información proporcionada de las características se identifican relaciones entre estas lo que ha permitido agruparlas. Si varias características del mismo grupo aportan a la explicación deberían hacerlo en el mismo sentido. Si el sentido de la aportación no es igual se considera que la explicación podría ser contradictoria. Para cada grupo el valor de la métrica se considera igual a 1 si las características asociadas a ese grupo aportan en sentidos opuestos, será 0 si no aportan o su sentido es el mismo. Se considera una métrica final de agrupación con signo opuesto como media de los datos por cada agrupación.

Para cada modelo se obtiene la media de cada métrica entre todas las instancias consideradas. En el caso de agrupación con signo opuesto se considera el número de veces que aparecen características con signo opuesto dividido entre el número de veces que aparecen varias características del mismo grupo y promediado entre el número de muestras totales.

### *Criterios de éxito en precisión e interpretabilidad.*

Para determinar la precisión de los modelos se considera tanto la precisión del modelo como por la precisión de la interpretabilidad. En función de las métricas identificadas se consideran los siguientes criterios para comparar los modelos.

- *Área bajo la curva ROC (AUROC)*. Se considera que un modelo es mejor cuanto mayor sea la precisión del modelo. Es deseable que los modelos tengan una precisión AUROC superior al 70%.
- *Número de unidades básicas por explicación*. Se considera que los modelos con menor número de unidades básicas por explicación son más interpretables. No hay un valor óptimo para dicha métrica, pero para este trabajo se considera que explicaciones con 5 o menos características son preferidas a aquellas con mayor número de características.
- *Unidades derivadas*. Se considera que los modelos con menor número de explicaciones asociadas a la característica agrupada son preferibles a aquellos que basan todas las explicaciones en dicha característica. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.
- *Valores especiales*. Los modelos cuyas explicaciones se basan en valores especiales no facilitan explicaciones fiables. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.
- *Agrupación con signo opuesto*. Se considera que los modelos cuyas explicaciones contienen características del mismo grupo, pero con signos opuestos son contradictorios. Se consideran mejores modelos si el valor de esta métrica está próximo a cero.

TABLA I

RESULTADO MÉTRICA DE PRECISIÓN

| Modelo | AUROC |
|---|---|
| Regresión Logística | 0.80 |
| Árbol de Decisión | 0.78 |
| Red Neuronal | 0.79 |
| XGBoost | 0.81 |

*Tabla I: Resultados obtenidos para la métrica de precisión AUROC en cada uno de los modelos de la comparativa.*

## V. RESULTADOS

En la Tabla I se muestran los resultados de la métrica de precisión, área bajo la curva ROC, para cada uno de los modelos en el conjunto de test y en la figura 1 se muestra la representación de la curva ROC para todos los modelos.

Para obtener las métricas de interpretabilidad asociadas a las instancias se debe realizar los cálculos de las aportaciones de cada característica en la predicción. Debido al alto coste computacional para las técnicas de LIME y SHAP, se reduce la muestra a las mil primeras instancias del conjunto de test. En el caso de Regresión Logística y Árbol de Regresión la obtención de los datos de cada característica vendrá dados a partir del entrenamiento de los modelos.

La técnica LIME permite configurar el número máximo de características que tendrá una explicación. Para ajustar la presión de los modelos en la métrica de unidades básicas de explicación, se configura este parámetro con un valor igual a 5.

TABLA II

RESULTADO DE MÉTRICAS DE INTERPRETABILIDAD

| Modelo<br>Métrica | LR | DT | CNN+ LIME | XGB+ LIME | CNN+ SHAP | XGB+ SHAP |
|---|---|---|---|---|---|---|
| Unidades básicas por explicación | 15.61 ± 1.76 | 3.22 ± 0.49 | 3.25 ±1.02 | 4.47 ±0.57 | 16.69 ±3.85 | 23.00 ±0.00 |
| Unidades Derivadas | 1.00 ±0.00 | 1.00 ±0.00 | 0.00 ±0.00 | 1.00 ±0.07 | 0.67 ±0.47 | 1.00 ±0.00 |
| Valores especiales | 1.24 ±1.03 | 0.18 ±0.41 | 0.07 ±0.30 | 0.27 ±0.46 | 0.99 ±0.94 | 1.23 ±1.02 |
| Agrupaciones con signo opuesto | 3.94e-4 ±0.00 | 0.00 ±0.00 | 1.66e-4 ±0.00 | 2.44e-4 ±0.00 | 3.22e-4 ±0.00 | 4.38e-4 ±0.00 |

*Resultados obtenidos para las métricas de interpretabilidad en cada uno de los modelos de la comparativa: Regresión Logística (LR), Árbol de Decisión (DT), Red Neuronal con LIME (CNN+LIME), XGBoost con LIME (XGB+LIME), Red Neuronal con SHAP (CNN+SHAP), XGBoost con SHAP (XGB+SHAP).*

En la Tabla II se puede ver el resultado de la aplicación de cada una de las métricas de interpretabilidad a cada modelo interpretable.

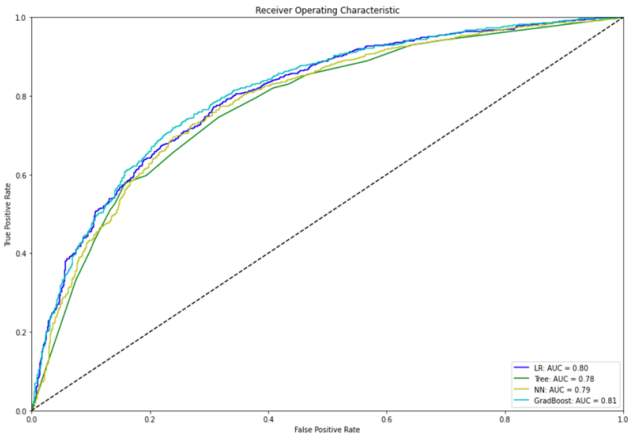



*Figura 1. Curva ROC para los modelos de la comparativa.*

## VI. DISCUSIÓN

Como se puede ver en el desarrollo de la comparativa realizada, la precisión de los modelos varía entre 78% del modelo del Árbol de Decisión hasta 81% en el modelo XGBoost.

Aunque se esperaba que la diferencia fuese más apreciable entre los modelos de caja negra y los modelos intrínsecamente explicables los resultados de precisión muestran una diferencia de tres puntos entre ellos.

En concreto, para el modelo de Red Neuronal, se han planteado varios modelos modificando la arquitectura con el número de capas densas, el número de neuronas en cada capa, los valores de dropout y el resto de los parámetros utilizados en la búsqueda exhaustiva. Ninguno de los modelos planteados ha superado los valores obtenidos para XGBoost. Este modelo es el que mejor precisión tiene, superando al Árbol de Decisión, lo que demuestra que la combinación de distintos clasificadores débiles permite obtener una mejor precisión. Es notable resaltar la precisión

obtenida con el modelo de Regresión Lineal con un valor de 80%, superando a la precisión del modelo de Red Neuronal y cercano al obtenido por XGBoost.

Se analizan a continuación los resultados obtenidos por las métricas de interpretabilidad utilizadas en el trabajo.

- *Número de unidades básicas por explicación*

  Como se puede observar en los resultados mostrados en la Tabla II, los modelos y técnicas con menor número de unidades básicas corresponden a Árbol de Decisión y los modelos de caja negra a los que se le ha aplicado la técnica de LIME.

  El número medio de atributos utilizado para la interpretación de una instancia es próximo a 3 para el caso de Árbol de Decisión y Red Neuronal con LIME. En el caso de XGBoost al cual se ha aplicado la técnica de LIME el número medio de características utilizadas para la interpretación es cercano a 4, redondeando al número entero más cercano. Este número reducido de características es posible por la configuración realizada para cada uno de ellos. En el Árbol de Decisión se configuró la profundidad máxima del árbol como 5. Del mismo modo se consideró un número máximo de características para la explicación con la técnica de LIME igual a 5. Estas configuraciones permiten reducir el número de características a aquellas con más peso en la decisión. Facilitando una explicación más clara y entendible que está centrada en las características más relevantes.

  En contraposición con estos resultados, los obtenidos para el resto de los modelos y técnicas muestran cómo se utilizan alrededor de 16 características de media para la interpretación de una instancia. En particular, el modelo XGBoost con la técnica de SHAP llega a utilizar 23 de las 24 características, solo una de ellas no cambia el valor predicho en ninguna combinación con otras características y, por lo tanto, según las propiedades de los valores de Shapley su aportación es igual a 0 (dummy). Si bien, aunque todas las características aporten, no todas tienen la misma importancia en la decisión. No hay un punto de corte que pueda ser utilizado de forma común para todos los casos, que permita elegir las más representativas y reducir el número de características utilizadas de forma general. Por lo tanto, las interpretaciones en estos casos pueden llegar a ser menos claras.

- *Unidades derivadas*

  A partir de los datos de la Tabla II para la métrica de unidades derivadas se puede comprobar cómo todos los modelos salvo la Red Neuronal junto con LIME utilizan el indicador consolidado de marcadores de riesgo para sus explicaciones. Como se indicó anteriormente, esta característica derivada es un marcador cuyo dato puede ser importante para el ajuste de los modelos, pero su inclusión en la interpretación de una instancia no aporta valor al prestatario que pide conocer el motivo de la decisión.

  El caso más claro en el que se puede ver la importancia de esta característica en la interpretación es el del Árbol de Decisión. El primer nodo del árbol es la característica derivada, que además interviene en otros nodos internos para la clasificación de las instancias por lo que es la característica de mayor importancia. Esto implica que la clasificación de los datos está altamente relacionada con el valor de la unidad derivada, aportando menor valor a la interpretabilidad de la decisión para una instancia.

- *Valores especiales*

  Los resultados de la métrica de valores especiales de la Tabla II muestran que los modelos de Regresión Lineal y XGBoost, junto con la técnica SHAP, utilizan en la explicación las características con valores especiales, la media en estos casos es de 1.23 para ambos. El modelo de Red Neuronal con la técnica de SHAP tiene un valor menor en esta métrica, pero también cercano a 1. Estos resultados tienen relación con la primera de las métricas de interpretabilidad analizada, puesto que estos tres casos correspondían a los que mayor número de características utilizaban para la interpretación, existiendo mayor posibilidad de valores especiales entre las características utilizadas.

  Para los modelos de Árbol de Decisión, y las cajas negras junto con la técnica de LIME se muestra cómo el valor medio de características con valores especiales utilizadas en la interpretación se reduce, siendo el mejor de los casos la Red Neuronal junto con LIME con un valor medio próximo a cero.

- *Agrupación con signo opuesto*

  De los datos mostrados en la Tabla II se obtiene que el mejor resultado para esta métrica es el modelo de Árbol de Decisión, cuyo valor es 0 en todos los grupos, esto es debido a que las características incluidas en el árbol de decisión no comparten grupo y por tanto no se presenta la opción de tener signos opuestos.

  La Red Neuronal junto con la técnica de LIME también tiene un comportamiento parecido, ya que en cinco de los ocho grupos su métrica es igual a 0. Esto tiene relación con el hecho de que el número de características utilizadas para la explicación se redujeron a cinco o menos.

  En contraposición el mayor valor de la métrica es para el modelo XGBoost y la técnica de SHAP, lo cual podría ser debido a que en este caso se utiliza una media de 23 características para la explicación, por lo que es más probable que, en ciertos casos, puedan darse diferencias de signos dentro de un grupo.

En función de las métricas anteriormente analizadas se podría concluir que, si bien el modelo de Red Neuronal no era el modelo con el más alto porcentaje de predicción frente al resto, al aplicar el método de LIME y teniendo en cuenta el resto de las métricas de interpretabilidad, mantiene una alta puntuación en cada una de las métricas, mejorando al resto de modelos y técnicas, permitiendo una compensación entre precisión e interpretabilidad.

## VII. Conclusiones

A lo largo de este trabajo se ha mostrado la importancia que ha adquirido la interpretación de los modelos de ML, debido a que cada vez está más ampliamente extendido su uso y son más las personas impactadas por decisiones tomadas por estos modelos. En especial en los dominios de alto riesgo, como el financiero, donde el coste de realizar una predicción incorrecta es alto y se deben adaptar a los reglamentos actuales como GDPR.

Como se presentó en el estado del arte, los estudios sobre técnicas de XAI que se pueden aplicar en función del alcance de la interpretación, complejidad de los modelos o desideratas de los interesados, está ampliamente estudiado en la literatura, sin embargo, no hay un estudio tan exhaustivo ni un consenso sobre cómo medir dichas técnicas para evaluar la calidad de los métodos de explicación.

En este trabajo se ha planteado el problema de la elección del mejor modelo de ML para la predicción de riesgo crediticio, abordando dicho problema no solo desde la perspectiva de la precisión del modelo, sino también desde la interpretabilidad. Para ello, se han utilizado métricas de precisión altamente consensuadas para el análisis de los modelos y se han aportado métricas, basadas en las características, que han permitido la evaluación de la interpretabilidad.

Para realizar la comparativa se llevó a cabo un análisis del estado del arte en interpretabilidad y de las aplicaciones realizadas en el dominio financiero. Este análisis permitió identificar los ejes más representativos para tener en cuenta en la elección de los modelos y técnicas: taxonomía, partes interesadas y medición y evaluación. Se seleccionó los modelos en función de su complejidad, utilizándose modelos intrínsecamente interpretables como Regresión Logística y Árbol de Decisión y modelos con mayor complejidad, llamados de caja negra, como son las Redes Neuronales y modelos que aplican técnicas de boosting como es XGBoost. Estos dos últimos modelos, debido a su naturaleza, no son interpretables, por lo que fue necesario utilizar técnicas de XAI. Entre los distintos enforques atendiendo a las partes interesadas en la interpretabilidad, este trabajo se centra en los prestatarios interesados en conocer la justificación de la decisión. Este eje determinó al alcance local de las explicaciones que permita la justificación para la decisión de una instancia específica. Atendiendo al alcance local se eligieron las técnicas de LIME y SHAP para la interpretación de los modelos de caja negra. Estas técnicas de atribución de características aditivas han sido ampliamente utilizadas en la literatura, además SHAP tiene una base sólida basada en la teoría de juegos. Aun existiendo otras técnicas de interpretación local, como Anchors, LORE o explicaciones contrafácticas, ha sido necesario acotar el alcance del trabajo para garantizar la viabilidad de realización de este, siendo la utilización de estas técnicas una línea de ampliación futura.

Para la aplicación de los modelos explicables en el ámbito financiero se ha utilizado el dataset HELOC, con el conjunto de datos anónimos de solicitudes de línea de crédito con garantía hipotecaria. Previo a la utilización del dataset por los distintos modelos se realizó un análisis de los datos, que llevó a la limpieza de los datos y tratamiento de las variables categóricas. Se seleccionaron las características más importantes y finalmente, se realizó la división entre conjunto de entrenamiento y test.

Tras el ajuste de los datos, se entrenaron los modelos elegidos para la comparativa. Se analizaron y configuraron los parámetros asociados a cada una de las arquitecturas con el objetivo de obtener el mejor rendimiento. Una vez entrenados los modelos se obtuvieron los datos de precisión. Todos los modelos superaron el 70% en la métrica AUROC, consiguiéndose así el objetivo marcado respecto a la precisión de los modelos.

Para la obtención de la justificación de una decisión específica se aplicaron las técnicas de interpretabilidad a los modelos de caja negra. Esto permitió mostrar la forma en la que se facilitaban las razones que explican la predicción realizada para una instancia, tanto para los modelos intrínsecamente explicables como para los modelos de caja negra con su técnica de interpretabilidad.

Se abordó entonces una de las partes importantes del objetivo general marcado, la evaluación de los modelos por su interpretabilidad y no solo por su precisión. A partir del proceso indicado por [1] para definir y evaluar la interpretabilidad, se definieron los principios generales de la evaluación de la interpretabilidad de los modelos basados en:

- *La necesidad de la interpretabilidad*. La incompletitud de la formulación del problema de generar modelos que sean justos en la decisión de conceder un préstamo hace necesaria la interpretabilidad que permita evaluar si el modelo no discrimina y sigue las bases éticas.
- *El nivel de la evaluación de la interpretación*. De los tres niveles planteados por el autor, dos de ellos estaban basados en humanos, lo cual no permitió abordar dicho nivel en este trabajo, centrándose en el nivel basado en la funcionalidad. Este nivel utiliza una definición formal de la interpretabilidad para medir la calidad de la explicación, sin requerir experimentos con humanos.
- *Factores relevantes para la interpretación*. Correspondientes a los factores para tener en cuenta para la evaluación de la interpretación local de las instancias. Se identificaron dichos factores como el número de unidades básicas que contiene la explicación, si son características en bruto o derivadas, la estructura o jerarquía que mantienen entre ellas y las interrelaciones existentes. Tras el análisis de los datos y características, no se identificaron relaciones de jerarquía entre las unidades básicas. Sí se consideró relevante las unidades con valores especiales en las interpretaciones.

Siguiendo los pasos indicados por el autor e identificando los factores, se llegó a la definición de las métricas para la evaluación de las interpretaciones que han permitido evaluar los modelos por su interpretabilidad. Esta es una aportación significativa frente a otros estudios realizados en la interpretabilidad de los modelos, incluidos el financiero, dado que no se limita a la aplicación de una técnica, sino que permite evaluar cuál sería el mejor modelo interpretable para la tarea en cuestión.

La aplicación de las métricas a las distintas interpretaciones de los modelos ha permitido comprobar que la técnica LIME aplicada al modelo de Red Neuronal obtiene una alta puntuación en todas las métricas de interpretabilidad. Si bien la precisión de este modelo no era la más alta, al considerar ambos criterios de precisión e interpretabilidad, permiten seleccionarlo como el modelo más interpretable de los analizados. Esto confirma que no siempre el modelo más preciso se ajustará mejor a las necesidades de interpretación necesarias para dominios de alto riesgo, como el financiero y que es necesario mantener un equilibrio entre ambos criterios adaptados a las necesidades.

Como futuras líneas de ampliación de este trabajo se propone analizar otras técnicas de interpretabilidad locales, que puedan ser comparadas con las incluidas en este trabajo y que permitan analizar el comportamiento de los modelos en las métricas de interpretabilidad definidas.

Adicionalmente se podría ampliar el nivel de evaluación de la interpretación. Como se ha indicado, este trabajo se centró en el nivel basado en la funcionalidad. Para dar más robustez a las métricas, se podría estudiar si los resultados se confirman en evaluaciones basadas en aplicaciones y basadas en humanos, permitiendo así afianzar dichas métricas o por el contrario analizar los resultados para mejorar las mismas.

En última instancia se plantea ampliar la búsqueda de factores relevantes asociados a los modelos, que permita ampliar las métricas de evaluación de interpretabilidad. El objetivo sería obtener métricas comunes, independientes de los modelos y dominios de aplicación, que pudieran adoptarse de forma general, no solo en el dominio financiero y que permitieran la evaluación de la interpretabilidad de los modelos.

## REFERENCIAS


[1] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [cs, stat]. http://arxiv.org/abs/1702.08608

[2] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80-89. https://doi.org/10.1109/DSAA.2018.00018

[4] Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics, 10(5), 593. https://doi.org/10.3390/electronics10050593

[5] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

[6] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

[7] European Commission. Joint Research Centre. (2020). Robustness and explainability of Artificial Intelligence: From technical to policy solutions. Publications Office. https://data.europa.eu/doi/10.2760/57493 Último acceso: 22/07/2021

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), Pub. L. No. 32016R0679, 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng Último acceso: 22/07/2021

[9] Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

[10] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. arXiv:1711.00399 [cs]. http://arxiv.org/abs/1711.00399

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat]. http://arxiv.org/abs/1602.04938

[12] Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs, stat]. http://arxiv.org/abs/1705.07874

[13] Freitas, A. A. (2014). Comprehensible classification models: A position paper. Boletín de exploración de ACM SIGKDD de, 15(1), 1-10. https://doi.org/10.1145/2594473.2594475

[14] Molnar, C. (2021). Interpretable machine learning. A Guide for Making Black Box Models Explainable. bookdown. https://christophm.github.io/interpretable-ml-book/ Último acceso: 22/07/2021

[15] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154 [cs, stat]. http://arxiv.org/abs/1811.10154

[16] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 93:1-93:42. https://doi.org/10.1145/3236009

[17] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473. https://doi.org/10.1016/j.artint.2021.103473

[18] Reina, J. L. R. (2018). Evaluación de modelos. Universidad de Sevilla. Razonamiento Asistido por Computador (2018-19), Tema 7, 40.

[19] Benítez, R., Escudero, G., Kanaan, S., & Maship Rodo, D. (2013). Inteligencia artificial avanzada. Editorial UOC, S.L.

[20] Nguyen, A., & Martínez, M. R. (2020). On quantitative aspects of model interpretability. arXiv:2007.07584 [cs, stat]. http://arxiv.org/abs/2007.07584

[21] Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine Learning Explainability in Finance: An Application to Default Risk Analysis (SSRN Scholarly Paper ID 3435104). Social Science Research Network. https://doi.org/10.2139/ssrn.3435104

[22] Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. 2016 IEEE Symposium on Security and Privacy (SP), 598-617. https://doi.org/10.1109/SP.2016.42

[23] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable Machine Learning in Credit Risk Management. Computational Economics, 57(1), 203-216. https://doi.org/10.1007/s10614-020-10042-0

[24] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

[25] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[26] Shapley, L. S. (2016). 17. A Value for n-Person Games. En Contributions to the Theory of Games (AM-28), Volume II (pp. 307-318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

[27] Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986. https://doi.org/10.1016/j.eswa.2020.113986

[28] Munkhdalai, L., Ryu, K. H., Namsrai, O.-E., & Theera-Umpon, N. (2021). A Partially Interpretable Adaptive Softmax Regression for Credit Scoring. Applied Sciences, 11(7), 3227. https://doi.org/10.3390/app11073227

[29] Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An Interpretable Model with Globally Consistent Explanations for Credit Risk. arXiv:1811.12615 [cs, stat]. http://arxiv.org/abs/1811.12615

[30] FICO community. (2018). Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge Último acceso: 22/07/2021