COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

# Interpretabilidad Global de Modelos de Machine Learning para la Gestión de Riesgos Financieros: Un Enfoque Agnóstico al Modelo

Autor: Inés Moreno Rabanal
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2024

# Resumen

La inteligencia artificial (IA) ha revolucionado la toma de decisiones y la productividad en el ámbito empresarial, impulsando la adopción masiva de tecnologías de *Machine Learning* (ML). Estas tecnologías facilitan la toma de decisiones basadas en datos, mejorando notablemente la eficiencia y la personalización de servicios. Sin embargo, a pesar de su alta precisión, los modelos de ML son complejos y a menudo resultan en sistemas de "caja negra"que son difíciles de comprender y pueden parecer opacos. La interpretabilidad de estos modelos es fundamental para validar decisiones, fomentar la confianza, identificar sesgos y corregir errores. La transparencia es necesaria para permitir a los interesados evaluar y optimizar las metodologías empleadas. Además, legislaciones recientes, como el Acta de IA de la Unión Europea, destacan la importancia de desarrollar modelos de ML que sean comprensibles y éticos, especialmente en sectores críticos como la evaluación crediticia. Adherirse a estas normativas fomenta una implementación responsable de la IA, reduce prácticas desleales y aumenta la responsabilidad general.

A raíz de la creciente relevancia del riesgo financiero en este campo, este trabajo expone un caso de uso centrado en la aplicación de cinco técnicas de interpretabilidad al análisis de riesgo crediticio: Importancia de la Característica por Permutación (PFI), Gráficos de Dependencia Parcial (PDP), Efectos Locales Acumulados (ALE), Modelos subrogados, y Prototipos y Críticas. Utilizando estas técnicas en un escenario de predicción de clasificación binaria con un modelo de caja negra XGBoost, se determina si un riesgo es bueno o malo. El análisis revela que las variables más determinantes en la evaluación del riesgo crediticio son monto de crédito, su duración y la edad del solicitante. El monto del crédito influye de manera no lineal en el riesgo: inicialmente mejora las probabilidades de aprobación hasta alcanzar un umbral, después del cual el riesgo se incrementa. Los préstamos de mayor duración también conllevan un riesgo elevado, reflejando la incertidumbre sobre la capacidad de pago a largo plazo del prestatario. Además, se observa una discriminación negativa hacia los solicitantes más jóvenes. Otros factores, como el empleo y la información sobre cuentas de ahorro, aunque pertinentes, indican que el modelo podría estar sesgado debido a una distribución desigual entre las distintas categorías de estas variables en el conjunto de datos.

**Palabras clave**: Interpretabilidad global, Machine Learning, Riesgo de crédito, PFI, PDP, ALE, Modelos Subrogados, Prototipos y críticas.

# Abstract

Artificial intelligence (AI) has had a significant impact on decision-making and productivity, resulting in widespread adoption of machine learning (ML) technologies in business. These technologies enable data-driven decision-making, which improves service efficiency and customization. Despite ML models' high accuracy, their complexity often leads to "black box"models that are difficult to understand and opaque.

The interpretability of these models is critical for justifying decisions, building trust, identifying biases, and correcting errors. This transparency is necessary for stakeholders to evaluate and improve methodologies.Furthermore, legal frameworks such as the recently enacted EU AI Act emphasize the importance of transparency, focusing on the development of understandable and ethical machine learning models, particularly in critical areas such as credit scoring. Compliance with such regulations promotes responsible AI deployment, reduces unfair practices, and increases accountability.

Recognizing the growing importance of financial risk in this field, and taking these considerations into account, it was decided in this study to develop a use case. This use case focuses on the application of five interpretability techniques to credit risk: Permutation feature importance (PFI), Partial dependence plots (PDP), Accumulated local effects (ALE), Global surrogate models, and Prototypes and Criticisms. These techniques were used in a binary classification prediction scenario with an XGBoost black box model to determine whether a risk is good or bad. According to the analysis, the most important variables in determining credit risk are the amount of credit, its duration, and the applicant's age. The amount of credit has a nonlinear effect on risk: it initially increases the chances of approval until a certain threshold is reached, at which point the risk increases. Longer-term loans are also riskier because the borrower's long-term payment capacity is uncertain. Additionally, younger applicants face negative discrimination. Other factors, such as employment and information about savings accounts, while relevant, suggest that the model may be biased due to an unequal distribution of these variables across the dataset.

**Keywords**:Global interpretability, Machine Learning, Credit risk, PFI, PDP, ALE, Surrogate models, Prototypes and Criticisms.

# Índice general

**1. Introducción** **1**
1.1. Motivación . . . . . 1
1.2. Objetivos . . . . . 5
1.2.1. Objetivo General . . . . . 5
1.2.2. Objetivos Específicos . . . . . 5
1.3. Estructura del documento . . . . . 6

**2. Interpretabilidad Global en Aplicaciones Financieras: Una revisión de la literatura** **7**

**3. Metodología de Análisis de Datos** **14**
3.1. Importancia de las características por permutación . . . . . 14
3.2. Gráficos de dependencia parcial . . . . . 16
3.3. Efectos Locales Acumulados . . . . . 19
3.4. Interacción entre variables . . . . . 21
3.5. Modelos Subrogados . . . . . 23
3.6. Prototipos y críticas . . . . . 24

**4. Caso de Uso: Aplicación en la Gestión de Riesgos Financieros** **27**
4.1. Descripción y Pre-procesamiento de los datos . . . . . 29
4.2. Entrenamiento del modelo de caja negra . . . . . 32
4.3. Análisis de interpretabilidad global . . . . . 34
4.3.1. Importancia de las características por permutación . . . . . 34
4.3.2. Gráficos de dependencia parcial y Efectos Locales Acumulados . . 36
4.3.3. Modelos subrogados . . . . . 41
4.3.4. Prototipos y críticas . . . . . 45
4.3.5. Resumen de los resultados . . . . . 46

**5. Conclusiones** **49**

**Bibliografía** **55**

# Índice de figuras

1.1. Niveles de interpretabilidad. Elaboración propia. . . . . 4

3.1. Ejemplo de PFI para la variable altura. Elaboración propia: adaptado de (Agarwal, 2022) . . . . . 15
3.2. Representación gráfica de la construcción de un PDP para la variable número de partidos ganados. Elaboración propia: adaptado de (Kuo, 2022). . . . . 17
3.3. Representación gráfica de dos PDP. Elaboración propia. . . . . 17
3.4. Comparación de la salida entre PDP y ALE para una variable. Elaboración propia: adaptado de (Molnar, 2019) . . . . . 20
3.5. Desarrollo de un modelo subrogado. Elaboración propia: adaptado de (Hall, Ambati, y Phan, 2017) . . . . . 24
3.6. Representación de prototipos y críticas en un conjunto de datos. Elaboración propia. Adapatado de: (Molnar, 2019) . . . . . 25

4.1. Secuencia de tareas. Elaboración propia. . . . . 27
4.2. Correlación entre variables. Elaboración propia. . . . . 31
4.3. Funcionamiento de XGBoost. Elaboración propia. Adaptado de: (*XGBoost Working*, 2023) . . . . . 33
4.4. PFI obtenido con el modelo de clasificación XGBoost para el conjunto de prueba y de entrenamiento. Elaboración propia. . . . . 34
4.5. PDP generado para las variables más importantes acorde a PFI. Elaboración propia . . . . . 37
4.6. Gráfico ALE generado para las variables más importantes acorde a PFI. Elaboración propia. . . . . 39
4.7. Importancia de las variables del modelo subrogado según el criterio gini. Elaboración propia. . . . . 42
4.8. Modelo subrogado generado para aproximar el modelo XGBoost. Elaboración propia. . . . . 44
4.9. Prototipos generados para las observaciones objeto de estudio. Elaboración propia. . . . . 45

# Índice de tablas

2.1. Resumen de las aplicaciones de las técnicas de interpretabilidad global en el ámbito de riesgos financieros . . . . . 12

4.1. Explicación de variables. Elaboración propia. . . . . 30

# Acrónimos

| | |
|---|---|
| *ALE* | Accumulated Local Effects |
| *FI* | Feature Interaction |
| *IA* | Inteligencia Artifical |
| *ICE* | Curvas de Expectativa Condicional Individual |
| *MDA* | Mean decrease accuracy |
| *ML* | Machine Learning |
| *MMD* | Máxima Discrepancia de Media |
| *PCA* | Análisis de componentes principales |
| *PDP* | Partial Dependence Pots |
| *PFI* | Permutation Feature Importance |
| *RGPD* | Reglamento General de Protección de Datos |
| *XAI* | Inteligencia Artifical Explicable |

# Capítulo 1

# Introducción

## 1.1. Motivación

La creciente expansión de la inteligencia artificial (IA) ha generado un impacto sin precedentes en la toma de decisiones y la productividad económica a nivel mundial. Según un estudio de McKinsey, la adopción de la IA en las empresas ha experimentado un impresionante crecimiento, pasando del 20 % al 50 % entre 2017 y 2022 (Company, 2023). Diversas industrias y gobiernos han reconocido su inmenso valor, anticipando que para 2027, la productividad derivada de la IA se establecerá como un indicador económico fundamental (Gartner, 2023).

Esta transformación liderada por la IA se ha visto potenciada en gran medida por los avances en el campo del *Machine Learning* , una de sus ramas más dinámicas. Este enfoque posibilita la creación de modelos predictivos capaces de identificar patrones, realizar pronósticos y tomar decisiones basadas en datos históricos disponibles. El aprendizaje automático ha permitido a las organizaciones extraer valiosos conocimientos de grandes volúmenes de datos, optimizando procesos y tomando decisiones más informadas. En el entorno empresarial, el ML se ha convertido en una herramienta clave para la innovación y la competitividad, impulsando la personalización de servicios, la eficiencia operativa y la toma de decisiones estratégicas. Su aplicación va desde el análisis predictivo en la cadena de suministro hasta la detección avanzada de fraudes en el sector financiero, demostrando su capacidad para adaptarse y agregar valor en diversos sectores (Allen, Gan, y Zheng, 2023).

Sin embargo, a medida que estos modelos de ML se vuelven más complejos y efectivos en sus predicciones, también se tornan menos transparentes en cuanto a los factores que influyen en sus pronósticos. Estos modelos, comúnmente conocidos como modelos de “caja negra” debido a la opacidad de sus algoritmos, son extremadamente difíciles de comprender para un observador externo. La falta de interpretabilidad en estos modelos plantea cuestiones significativas en la toma de decisiones críticas, como en situaciones judiciales donde individuos pueden ser condenados injustamente debido a sesgos en modelos de caja negra (Alonso

y Carbo, 2020). Dado que el funcionamiento interno de estos modelos es incomprensible, identificar y corregir estos sesgos es una tarea desafiante. Esto resalta la apremiante necesidad de desarrollar técnicas de interpretabilidad en el ámbito de la IA y el ML. Estas técnicas permiten arrojar luz sobre el funcionamiento de estos modelos opacos, proporcionando una visión clara de cómo se toman las decisiones, y demostrando su capacidad para adaptarse y agregar valor en diversos sectores.

En este contexto, el término *interpretabilidad* adquiere una relevancia particular. Se refiere a la capacidad de justificar el razonamiento detrás de las decisiones tomadas por los modelos de ML (Khan et al., 2022). Aunque existen diferentes definiciones de interpretabilidad en la literatura, en esencia, se trata del grado en que un ser humano puede entender y seguir el proceso lógico que lleva a una decisión específica (Molnar, 2019). Esta comprensión es fundamental para garantizar que las decisiones automatizadas sean transparentes, confiables y justas, especialmente en aplicaciones críticas donde los errores o sesgos pueden tener consecuencias significativas.

La interpretabilidad de los modelos de ML conlleva múltiples ventajas significativas. En primer lugar, habilita una depuración efectiva o *debugging*, lo que implica comprender la forma en la que el modelo toma decisiones a un nivel más profundo. Esto permite la identificación de posibles errores en el funcionamiento del modelo, brindando la oportunidad de validar su desempeño de manera más robusta (Hassija et al., 2023). Además, la interpretabilidad potencia la confianza que las personas tienen en el modelo, ya que los profesionales pueden brindar asistencia a aquellos que se ven afectados negativamente por las predicciones del modelo. Por ejemplo, una persona a la que se le deniega un crédito puede comprender las razones detrás de esta decisión, como la insuficiencia de ingresos o pagos previos incumplidos, lo que le permite tomar medidas para mejorar su situación y aumentar sus posibilidades de éxito en el futuro (Alonso y Carbo, 2020).

Otra ventaja crucial de la interpretabilidad es su rol en la prevención y mitigación de sesgos en los modelos. Los modelos de "caja negra", caracterizados por su opacidad, pueden originar problemas graves relacionados con la salud, la libertad, los sesgos raciales y la seguridad. Por ello, es fundamental que estos modelos se desarrollen con un enfoque justo y equitativo, evitando que variables sensibles como el nivel socioeconómico, raza, religión o género generen discriminación. La Comisión Europea enfatiza en este aspecto, señalando la importancia de garantizar que los sistemas de IA no generen resultados injustamente sesgados y respeten a los grupos potencialmente vulnerables, incluyendo trabajadores, mujeres, personas con discapacidad, minorías étnicas, niños y consumidores (Commission, 2019). Además, la interpretabilidad no solo es fundamental desde un punto de vista ético y moral, sino que también está siendo cada vez más regulada. Un análisis de McKinsey sobre registros legislativos en 127 países revela que el número de leyes aprobadas que mencionan el término "inteligencia artificial" aumentó de 1 en 2016 a 37 en 2022 (Company, 2023), lo que refleja un creciente interés en la regulación y supervisión de la IA a nivel mundial. Esta

tendencia legislativa subraya la necesidad de interpretación y transparencia en los sistemas automatizados para abordar los desafíos éticos y prácticos asociados con la expansión de la IA.

En este contexto normativo, el 13 de marzo de 2024, el parlamento de la Unión Europea aprobó la directiva de Inteligencia Artificial (*AI Act*). Esta regulación obliga a las empresas que implementan modelos de ML como los utilizados en el *credit scoring* a implementar medidas de interpretabilidad y transparencia de manera urgente. El incumplimiento de estas disposiciones no es una opción viable, ya que conlleva sanciones económicas severas que pueden ascender hasta el 3 % de la facturación global anual de la compañía. La directiva destaca la necesidad de que las entidades financieras aseguren que sus prácticas de *scoring* sean justas, equitativas y comprensibles para los usuarios, promoviendo así un entorno financiero más ético y responsable. Las empresas deben actuar con prontitud para adaptar sus sistemas a este nuevo marco legal y evitar las repercusiones que conllevaría su incumplimiento (*Ley de Inteligencia Artificial de la Unión Europea*, 2024).

Paralelamente, en los últimos años ha emergido un creciente interés en la explicabilidad de los modelos de ML, dando origen al desarrollo del campo conocido como *Inteligencia Artificial Explicable* (XAI, por sus siglas en inglés, *Explainable Artificial Intelligence*). Dentro de la XAI, se distinguen tres niveles de interpretabilidad en los modelos de ML. El esquema presentado en la Figura 1.1 se ofrece como una herramienta para esclarecer de manera efectiva estos niveles. El primero es la interpretabilidad previa a la modelización, que consiste en analizar y sintetizar información a partir de datos disponibles antes de construir el modelo. El segundo nivel incluye modelos intrínsecamente interpretables (*ante-hoc*), como los modelos subrogados o los árboles de decisión. Estos modelos son comprensibles en su funcionamiento paso a paso sin la necesidad de métodos adicionales de explicación y, por su naturaleza, resultan más transparentes.

Finalmente, se identifican los modelos de "caja negra" que pueden ser explicados mediante el enfoque *post-hoc*. Este tipo de estrategias, aplicadas tras la construcción del modelo, se dividen en dos categorías principales:

- **Enfoque específico:** Esta categoría incorpora una aproximación diseñada exclusivamente para un tipo específico de modelo. Una solución a medida es proporcionada, donde se consideran el *input*, *output* y el funcionamiento interno del modelo. No obstante, debido a su naturaleza específica, estos métodos están restringidos a un tipo particular de "caja negra" y no se generalizan a otros modelos.

- **Enfoque agnóstico:** Este enfoque, que será el principal objeto de estudio en este trabajo de fin de grado, es caracterizado por su aplicabilidad a cualquier tipo de modelo, sin depender de las particularidades de un modelo específico. Se basa en técnicas que utilizan la ingeniería inversa para obtener una aproximación del modelo original, observando solamente los *inputs* y *outputs* del modelo de "caja negra" (Khan et al., 2022).

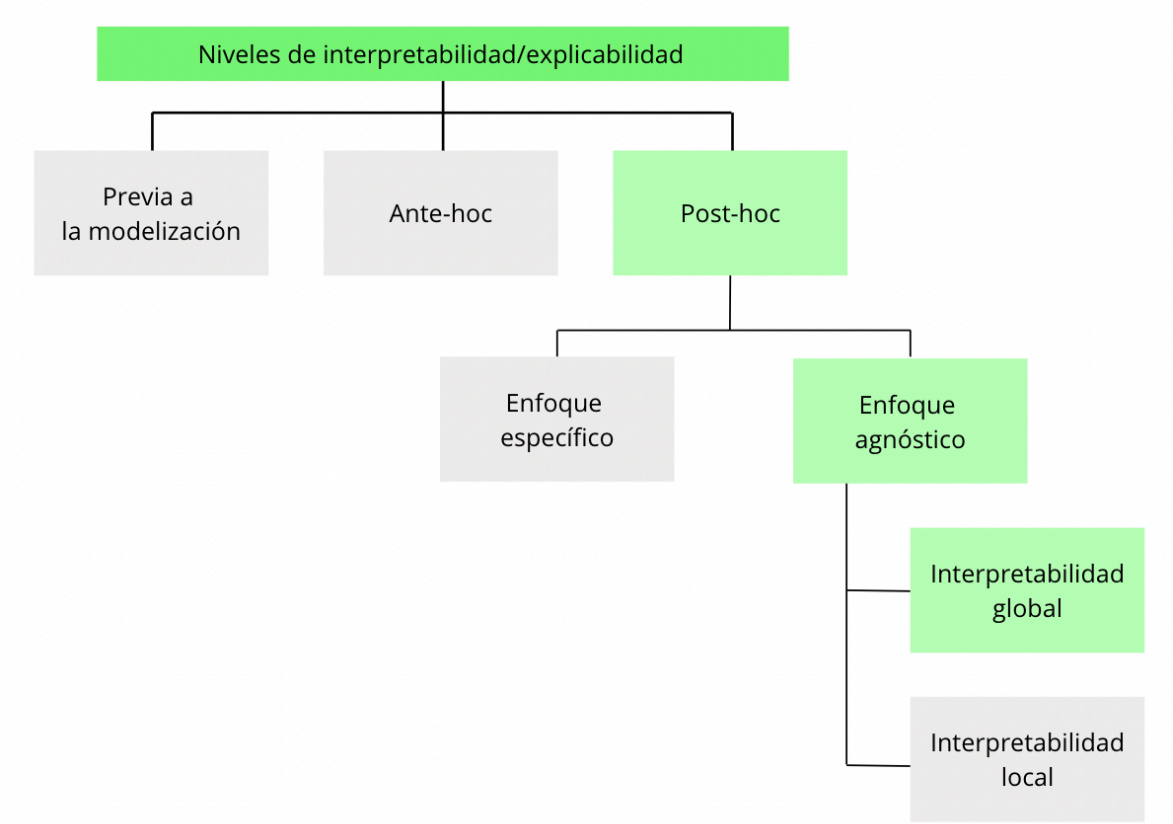



Figura 1.1: Niveles de interpretabilidad. Elaboración propia.

El enfoque agnóstico, basado en principios aplicables de manera universal, se caracteriza por su notable flexibilidad. A diferencia del enfoque específico, permite su uso en cualquier tipo de modelo de ML, como *random forests*, redes neuronales o cualquier tipo de *ensemble*. Esta versatilidad también se refleja en la explicación y representación de los modelos. Las explicaciones pueden adaptarse en su forma y contenido, ajustándose al contexto particular y a las necesidades específicas del usuario.

Dentro del enfoque agnóstico, se distinguen dos tipos de interpretabilidad: **global** y **local**. La interpretabilidad global, que será el foco de este trabajo de fin de grado, busca una comprensión completa del modelo, permitiendo descifrar su lógica y el razonamiento detrás de una variedad de resultados posibles. Esta capacidad es invaluable para comprender el comportamiento general del modelo, ofreciendo información relevante sobre su funcionamiento promedio. Esta interpretación global resulta especialmente útil para reguladores y directivos, como en el ámbito de la calificación crediticia, ya que les permite entender el modelo en su conjunto. Sin embargo, se debe reconocer que alcanzar una completa interpretabilidad global puede ser un reto práctico, dada la complejidad y opacidad de muchos modelos de ML (Demajo, Vella, y Dingli, 2020).

En contraste, la interpretabilidad local no busca explicar el comportamiento global del modelo, sino que se enfoca en ofrecer interpretaciones para decisiones individuales. Este

enfoque permite comprender el impacto de cada variable en la predicción realizada para una observación específica. Por ejemplo, en el caso de clientes que solicitan créditos, este tipo de explicaciones son preferibles, ya que les proporciona detalles específicos sobre las razones de la negativa de su crédito y el razonamiento del modelo para su caso particular. Además, la interpretabilidad local brinda la posibilidad de simular observaciones cercanas a la actual, permitiendo analizar el impacto de posibles variaciones en las variables de entrada sobre la predicción, en diversos escenarios.

Esta exploración inicial en el área de la interpretabilidad de modelos de ML establece el fundamento sobre el cual se construirá el presente trabajo de fin de grado. Este estudio pretende investigar y profundizar en el uso de técnicas de interpretabilidad global dentro de modelos de ML, especialmente enfocado en su aplicación en el sector financiero. La interpretabilidad es esencial en este ámbito, no solo por su valor técnico en la mejora de modelos predictivos, sino también por su relevancia crítica en la toma de decisiones financieras, donde la transparencia y la comprensión clara de los procesos son imperativos (Fritz-Morgenthal, Hein, y Papenbrock, 2022). Dado el papel cada vez más influyente que la IA juega en el sector financiero, principalmente en la gestión de riesgos, este estudio busca aportar una perspectiva valiosa sobre la forma en la que la interpretabilidad puede ser empleada para garantizar decisiones financieras más justas, responsables y comprensibles para la sociedad actual.

## 1.2. Objetivos

### 1.2.1. Objetivo General

El objetivo principal de este trabajo de fin de grado es llevar a cabo una evaluación de las técnicas de interpretabilidad global, que son agnósticas al modelo, dentro del contexto de modelos de machine learning. Esta evaluación se enfoca en su aplicación específica en el ámbito de la gestión de riesgos financieros, con el propósito de mejorar la comprensión y la transparencia de las decisiones tomadas por estos modelos.

### 1.2.2. Objetivos Específicos

- Contextualizar el concepto de interpretabilidad global en el ámbito del aprendizaje automático y destacar su importancia para la toma de decisiones efectiva.
- Analizar y explicar las técnicas de interpretablidad global y agnósticas al modelo para estrategias de aprendizaje automático, destacando sus ventajas y desventajas.
- Desarrollar y entrenar un modelo de aprendizaje automático que sea intrínsecamente no interpretable, con aplicación en el ámbito de la gestión de riesgos financieros.

- Evaluar las técnicas de interpretabilidad global más prevalentes, en el marco de un escenario específico de gestión de riesgos, con el objetivo de proporcionar un análisis detallado de la relevancia de las variables involucradas en la toma de decisiones..

## 1.3. Estructura del documento

Este documento se organiza en cinco capítulos estructurados para facilitar una comprensión integral del tema. El primer capítulo introduce el contexto y la motivación detrás de este estudio, presentando conceptos fundamentales, el objetivo general y los objetivos específicos de la investigación. En el capítulo 2, se lleva a cabo una revisión de la bibliografía, donde se exploran soluciones propuestas para potenciar la interpretabilidad global en la gestión de riesgos financieros. Este capítulo abarca los avances más relevantes en este campo y realiza un análisis detallado de investigaciones significativas. En el capítulo 3 se efectua un análisis y comparación conceptual de diversas técnicas de interpretabilidad global que son agnósticas a los modelos de ML, con especial énfasis en las metodologías disponibles en la actualidad.

El capítulo 4 se centra en un caso de uso específico en la gestión de riesgos financieros, donde se describe el preprocesamiento de datos, el entrenamiento de un modelo de aprendizaje automático no interpretable y la aplicación de técnicas de interpretabilidad global. Este capítulo incluye un análisis detallado y la discusión de los resultados obtenidos. Finalmente, el capítulo 5 presenta las conclusiones del estudio, resumiendo los resultados más relevantes, discutiendo implicaciones y proponiendo futuras áreas de investigación en la interpretabilidad de modelos de ML.

# Capítulo 2

# Interpretabilidad Global en Aplicaciones Financieras: Una revisión de la literatura

En el ámbito de la gestión de riesgos financieros, la adopción de modelos de ML ha transformado la capacidad de las empresas para predecir eventos y tomar decisiones críticas. Sin embargo, la complejidad inherente y la naturaleza opaca de nuevos modelos plantean desafíos significativos en términos de transparencia y confianza, especialmente cuando las decisiones tienen profundas implicaciones económicas y sociales. Este desafío destaca la creciente necesidad de técnicas de interpretabilidad global, que buscan entender el funcionamiento interno de los modelos de ML y proporcionar explicaciones comprensibles de sus decisiones. Por esta razón, diversos investigadores han reconocido la importancia de la interpretabilidad en este contexto y han propuesto una variedad de estrategias de interpretabilidad global. Estos métodos no solo buscan mejorar la transparencia y la confiabilidad de los modelos de ML en la gestión de riesgos financieros, sino también fortalecer la confianza entre los usuarios y las instituciones, garantizando que las decisiones basadas en IA sean justas, éticas y comprendidas de manera integral.

En este contexto, la técnica de la Importancia de la Característica de Permutación (PFI, por sus siglas en inglés, *Permutation Feature Importance*) representa un primer enfoque para determinar la relevancia de las variables en la precisión predictiva de un modelo. Esta estrategia fue introducida inicialmente por Breiman en 1999 en el contexto de su investigación sobre *Random Forests* (Breiman, 2001), y posteriormente fue extendida por Fisher, quien desarrolló una versión agnóstica al modelo que permite evaluar la importancia individual de cada variable. Esta metodología, que altera secuencialmente los valores de una variable para observar el impacto en el rendimiento del modelo, se ha convertido en una herramienta esencial en el campo, permitiendo a los investigadores identificar y priorizar las variables que influyen significativamente en las predicciones (Fisher, Rudin, y Dominici, 2019).

Dentro del campo de la gestión de riesgos, la aplicación de la técnica PFI ha sido explorada en un estudio realizado con datos procedentes de la Oficina de Crédito de Polonia. Esta

investigación incluye informes de crédito de bancos y entidades de préstamo a nivel nacional, y utiliza el método PFI para identificar variables clave en la determinación de la capacidad de una empresa para evitar el impago. El objetivo se extiende más allá de la predicción, buscando orientar a los clientes sobre cómo mejorar su perfil crediticio para reducir el riesgo de ser considerados potenciales morosos por el modelo. En este sentido, se advierte que retrasos en los pagos de más de 30 días pueden disminuir considerablemente las posibilidades de conseguir financiación bancaria en el futuro (Kłosok, Chlebus, et al., 2020). Este enfoque práctico de PFI, centrado en proveer recomendaciones concretas para mejorar la gestión de riesgos financieros, ha sido replicado en varios estudios que resaltan su eficacia en el ámbito de las puntuaciones crediticias (Biecek et al., 2021), (Alonso y Carbó, 2022), (Salter, 2023).

Incorporando el análisis de PFI en el contexto de la gestión de riesgos financieros, es importante reconocer que, aunque PFI puede ser empleado como un buen punto de partida, presenta ciertas limitaciones. Tal como (Toloşi y Lengauer, 2011) han señalado, la efectividad de PFI está fuertemente influenciada por la correlación entre variables, lo que puede afectar la precisión al identificar la contribución de los predictores en la salida del modelo. Este método estima la importancia de las variables pero no permite identificar la naturaleza de la relación entre las entradas y las salidas. De esta manera, para complementar el análisis proporcionado por PFI y obtener una comprensión más detallada del efecto de las variables sobre las predicciones del modelo, se pueden emplear técnicas adicionales como los Gráficos de Dependencia Parcial (PDP, por sus siglas en inglés, *Partial Dependence Plots*) y los Gráficos de Efectos Locales Acumulados (ALE, por sus siglas en inglés, *Accumulated Local Effects*).

Estas estrategias permiten visualizar el impacto individual de cada variable sobre las predicciones del modelo, proporcionando una representación gráfica más detallada de la relación entre las entradas y las salidas. Los PDP, introducidos por (Friedman, 2001), ilustran las variaciones de las predicciones promedio al modificar una o dos variables, manteniendo constantes las demás. A pesar de que estos gráficos son intuitivos y fáciles de interpretar, incurren en limitaciones significativas al asumir la independencia de las variables. Tal suposición es problemática, particularmente bajo condiciones de multicolinealidad, donde la interrelación entre variables puede sesgar la interpretación. Frente a la multicolinealidad, los gráficos ALE se presentan como una solución, evitando los problemas asociados con PDP. A diferencia de estos, los gráficos ALE se enfocan en las variaciones de las predicciones, calculando el promedio de las diferencias dentro de una ventana limitada de valores, lo cual los hace inmunes a los efectos distorsionadores de los outliers, proporcionando así una herramienta de interpretación más confiable.

El fundamento de los gráficos ALE radica en la descomposición funcional, que consiste en dividir una función de predicción en la suma de los efectos individuales de las características. Este enfoque facilita la visualización y comprensión del aporte específico de cada característica a las predicciones del modelo. Varias metodologías aplican este principio,

incluyendo los gráficos ALE, el análisis ANOVA (Hooker, 2004), y modelos de regresión estadística como los Modelos Lineales Generalizados (GLM) o los Modelos Aditivos Generalizados (GAM) (Caruana et al., 2015). Estos métodos permiten una interpretación detallada de la influencia de cada variable de entrada sobre las predicciones del modelo, enriqueciendo así la comprensión de su contribución en diversos contextos y aplicaciones.

Considerando el ámbito específico de la gestión de riesgos financieros, (Cascarino, Moscatelli, y Parlapiano, 2022) aplicaron el método ALE después de identificar las variables más influyentes, utilizando también métodos de interpretación locales como los valores de Shapley para complementar su análisis. Los investigadores centraron su estudio en la explicabilidad de un modelo de ML diseñado para predecir la probabilidad de incumplimiento de empresas no financieras italianas. De este análisis, se concluyó que los ratios financieros, como la liquidez, tienen una gran relevancia en la predicción de incumplimientos, un resultado coherente con la intuición financiera. Este enfoque metodológico no es único, otros investigadores también han combinado el uso de PDP y ALE para entender la dinámica entre las variables y su impacto en las predicciones (Tejero, 2022). Estos estudios demuestran que los gráficos PDP, a pesar de sus limitaciones, siguen siendo prevalentes en la gestión de riesgos financieros. Asimismo, los gráficos ALE ofrecen información valiosa, enriqueciendo el análisis al destacar la relevancia de los indicadores financieros dentro de modelos complejos. además de que los gráficos ALE proporcionan información adicional al evaluar la importancia de los indicadores financieros en modelos complejos (Yi, Liang, Xie, y Li, 2023), (Hashimoto, Miura, y Yoshizaki, 2023), (Biecek et al., 2021).

Tras explorar técnicas que destacan la importancia de cada variable y estiman su efecto individual en la predicción final, surge la necesidad de profundizar en el estudio de las relaciones específicas entre las entradas y la salida del modelo. La interacción entre características (FI, por sus siglas en inglés, *Feature Interaction*) se presenta como una nueva técnica que caracteriza las conexiones complejas entre las variables dentro de un modelo. En este enfoque, es importante destacar que la influencia de una variable sobre el resultado final no es un fenómeno aislado, sino que se encuentra estrechamente vinculada a la interacción con otras variables. Por ejemplo, en la predicción de una variable binaria como el "impago", factores como la edad, los ingresos y la existencia de una pareja se entrelazan en la determinación del riesgo. De hecho, la predicción de impago puede diferir notablemente incluso entre dos personas con ingresos idénticos, dado que la presencia de una pareja podría disminuir la probabilidad de impago al distribuir entre ambos la carga financiera. Este tipo de estrategias ofrece una visión más completa del comportamiento del modelo, resaltando la importancia de analizar el efecto de las combinaciones de variables sobre las predicciones finales.

Ante la complejidad de entender las interacciones entre variables en modelos predictivos, se hace necesario adoptar nuevas estrategias que permitan una comprensión más detallada de estas dinámicas. (Friedman y Popescu, 2008) introdujeron el estadístico $H$, una metodología que cuantifica la interacción entre variables y su influencia en la variabilidad de las predic-

ciones del modelo. En el área de estudio, este enfoque se aplicó en una investigación sobre pólizas, complementándose con el método ALE para proporcionar una interpretación más profunda de las interacciones. En este estudio se reveló la importancia de la interacción entre las características del producto y las primas anuales en la predicción de riesgos y cancelaciones.(Bermúdez, Anaya, y Belles-Sampera, 2023). Además, se han desarrollado otras técnicas para explorar la dinámica entre variables, tales como las Redes de Interacción Variable propuestas por (Hooker, 2004) y la interacción de características basada en dependencia parcial planteada por (Greenwell, Boehmke, y McCarthy, 2018).

En el esfuerzo por mejorar la interpretabilidad global en modelos de ML, los modelos subrogados han surgido como una solución efectiva. Estos constituyen modelos interpretables que buscan replicar el comportamiento de modelos complejos y opacos, denominados comúnmente como "caja negra". Al utilizar tanto los datos de entrada como de salida del modelo original, se emplea un modelo más sencillo e interpretable para obtener las predicciones. Entre los ejemplos destacados de modelos subrogados se encuentran los árboles de decisión, los cuales facilitan la visualización y comprensión del proceso mediante el cual el modelo original llega a sus conclusiones (Molnar, 2019). Este enfoque no solo permite una aproximación práctica para entender la lógica interna de los algoritmos complejos, sino que también hace accesible la interpretación de sus decisiones a un público más amplio, manteniendo una fidelidad razonable con el modelo de caja negra que caracteriza.

Dentro del ámbito financiero, la aplicabilidad de los modelos subrogados ha sido validada a través de diversas investigaciones. Por ejemplo, (Güntay, Bozan, Tığrak, Durdu, y Özkahya, 2022) desarrollaron un estudio que empleó datos de un banco turco para la predicción de la puntuación de crédito. En este caso, se optó por un árbol de decisión como modelo subrogado debido a su rendimiento competitivo en comparación con el algoritmo de caja negra XBoost, destacando su capacidad para aproximar fielmente las predicciones con una interpretación mucho más accesible. Esta elección destaca la preferencia recurrente por los árboles de decisión en diversos estudios como modelos subrogados, valorados especialmente por su claridad y sencillez interpretativa (Ayad, Hegazy, y Dahroug, 2023).

En este contexto, es importante resaltar que la elección de los datos para el entrenamiento impacta directamente en la calidad y la interpretabilidad de los resultados del modelo. Para afrontar este desafío, se han desarrollado diversas estrategias, entre las cuales destaca una metodología basada en el análisis de prototipos y críticas. Dicho enfoque consiste en seleccionar un conjunto de datos representativo, denominado prototipo, y compararlo con un conjunto menos representativo de una muestra más distante, conocido como crítica. Tal estrategia brinda información valiosa sobre el efecto de distintos segmentos de datos en el comportamiento del modelo, facilitando la comprensión de las dinámicas internas. Entre los métodos para llevar a cabo este muestreo, el MMD-critic se destaca por su prevalencia (Kim, Khanna, y Koyejo, 2016), aunque existen alternativas como Protodash, que amplía este enfoque (Gurumoorthy, Dhurandhar, Cecchi, y Aggarwal, 2019), y k-medoids, un método de muestreo

más simple que data de 1987 (Rdusseeun y Kaufman, 1987). Estas técnicas de análisis de datos complementan las estrategias previamente discutidas, como los modelos subrogados y las técnicas de interacción de características, al ofrecer una interpretación más completa de los modelos.

En el marco específico del presente trabajo de grado, el método Protodash ha sido implementado en una investigación enfocada en entender el proceso de toma de decisiones de un modelo de Línea de Crédito con Garantía Hipotecaria. Este enfoque se caracteriza por la selección de perfiles de usuarios que presentan similitudes con el solicitante de un préstamo dentro del conjunto de datos de entrenamiento, asignando pesos que reflejan su grado de semejanza. Estos pesos son de gran utilidad para los oficiales de crédito, ya que facilitan la comprensión de las bases sobre las cuales se aprueba o rechaza una solicitud. Al realizar comparaciones con perfiles similares, Protodash proporciona claridad sobre los elementos determinantes en las decisiones del modelo, potenciando la transparencia y el entendimiento en los procedimientos de préstamo (Karpagam, Varma, Samrddhi, et al., 2022). La aplicación de prototipos ha trascendido a otros campos, como la detección de fraudes, demostrando su versatilidad y contribución a la explicabilidad en diversos contextos analíticos (Cirqueira, Nedbal, Helfert, y Bezbradica, 2020).

En el campo de la Interpretabilidad Global de algoritmos de ML, aplicada a la gestión de riesgos financieros, se han destacado siete métodos como los más significativos según la literatura revisada. En la tabla 2.1, se resumen las aplicaciones de estas técnicas de interpretabilidad global, incluyendo la estrategia empleada, el propósito y los resultados de cada estudio académico. Los trabajos analizados comparten el objetivo de investigar y aplicar métodos de interpretabilidad global específicamente en la gestión de riesgos financieros, enfocándose en incrementar la transparencia de los modelos predictivos de crédito. Dichas técnicas son de gran relevancia en la interpretabilidad de modelos complejos, aportando explicaciones comprensibles acerca de las decisiones crediticias, destacando variables clave y señalando posibles sesgos.

En el próximo capítulo, se ofrece una descripción detallada del funcionamiento de cada una de estas técnicas de interpretabilidad global. Además, se explorara sus ventajas y desventajas, teniendo en cuenta su utilidad en el ámbito financiero. Esto posibilita una evaluación más completa de estas técnicas y facilita la elección adecuada de la herramientas interpretables más idóneas para afrontar los retos particulares en la gestión del riesgo financiero.

Tabla 2.1: Resumen de las aplicaciones de las técnicas de interpretabilidad global en el ámbito de riesgos financieros

| Referencia | Estrategia de Interpretabilidad | Objetivo | Resultados |
|---|---|---|---|
| (Kłosok et al., 2020) | PFI, PDP y ALE | Explorar métodos que incrementen la transparencia y confianza en las predicciones de modelos de caja negra en el scoring crediticio. | La tasa de cuotas y el historial de residencia, tienen una influencia constante en las predicciones de predicción de incumplimiento crediticio, mientras que el saldo de la cuenta y la duración del préstamo, tienen un impacto más significativo. |
| (Alonso y Carbó, 2022) | PFI | Analizar la precisión de técnicas de interpretabilidad en modelos de ML mediante la generación y evaluación de conjuntos de datos sintéticos en finanzas. | La precisión de PFI es alta, especialmente para el algoritmo XGBoost, aunque SHAP ofrece mejores resultados. Se destaca que estos resultados podrían variar según el conjunto de datos y el modelo utilizado. |
| (Salter, 2023) | PFI | Evaluar el rendimiento y la explicabilidad de métodos modernos de ML y de *Deep Learning*(XGBoost, Long Short-Term Memory (LSTM) y Regresión Lineal (LR)) en comparación con enfoques tradicionales para el scoring de crédito. | XGBoost tuvo el mejor rendimiento, seguido por LR y LSTM. Aunque LSTM mostró buen rendimiento, su falta de explicabilidad lo hizo menos adecuado para aplicaciones reguladas. En entornos donde la transparencia es fundamental, la simplicidad y explicabilidad de LR pueden ser más valiosas. PFI determinó las características de comportamiento de pago mensual como las más importantes. |
| (Biecek et al., 2021) | PFI y PDP | Realizar una comparación de diversos modelos crediticios para evaluar el riesgo crediticio, e interpretar los resultados obtenidos | Los modelos basados en árboles sobresalen en la predicción de incumplimientos. Las variables que tienen un impacto negativo significativo en el scoring crediticio incluyen el pago de cuotas con más de treinta días de retraso, un uso excesivo de créditos "revolving" y la falta de desarrollo de un historial crediticio. |
| (Cascarino et al., 2022) | ALE | Realizar un estudio de explicabilidad de un modelo de predicción de incumplimiento corporativo basado en *random forest*. | Durante períodos de deterioro financiero, las variables con relaciones más complejas, como la liquidez y el endeudamiento de las empresas, cobran mayor relevancia en la predicción de incumplimientos. |
| (Tejero, 2022) | ALE y PDP | Evaluar la efectividad de técnicas de interpretación en una red neuronal que estima la probabilidad de impago. | Se destaca la influencia de la variable edad en el modelo, así como el principal y el principal restante de la hipoteca. La fuerte dependencia entre estas variables afecta la suposición de independencia subyacente en los métodos de interpretación, lo que dificulta su comprensión. |
| (Yi et al., 2023) | PDP y PFI | Predecir el riesgo financiero en una cadena de suministro utilizando un modelo de caja negra e interpretarlo. | El tiempo promedio de pago afecta la probabilidad de alto riesgo, disminuyendo marginalmente. Además, la oferta de financiación, cambios en los términos de pago y cargos por permanencia en la lista de proveedores influyen en los riesgos financieros y la estabilidad de la cadena de suministro. |
| (Hashimoto et al., 2023) | PDP | Estudiar e interpretar un modelo de credit scoring de probabilidad de impago con ML. | Los ingresos totales, la rotación de activos totales y el Índice de Cobertura de Intereses (ICR) tienen un impacto significativo en la calidad crediticia de las empresas. |
| (Güntay et al., 2022) | Subrogado | Aproximar el modelo de caja negra XGBoost con un modelo inherentemente explicable, aplicado al scoring crediticio. | El rendimiento del árbol de decisión se aproxima lo suficiente al rendimiento del modelo XGBoost, lo que sugiere la posibilidad de desarrollar un modelo de puntuación crediticia explicativo de alto rendimiento con una mínima disminución en la precisión del mismo. |

| Referencia | Estrategia de Interpretabilidad | Objetivo | Resultados |
|---|---|---|---|
| (Karpagam et al., 2022) | Prototipos | Interpretar un modelo predictivo de caja negra de solicitud de préstamo con el framework creado por IBM llamado AIX 360. | El método Protodash identifica un perfil de cliente altamente representativo, lo que sugiere que las personas con perfiles similares tienen mayores probabilidades de que se les apruebe el préstamo. Además, señala que una solicitud puede ser rechazada si hay un bajo número de transacciones satisfactorias o si la deuda del solicitante es excesiva. |

# Capítulo 3

# Metodología de Análisis de Datos

En el capítulo 2, se han introducido las distintas técnicas de interpretabilidad global aplicables a los modelos de ML. En este capítulo, se proporciona una descripción de los fundamentos conceptuales de cada técnica, se analiza sus ventajas y limitaciones, y se destacan las restricciones particulares inherentes a cada método.

## 3.1. Importancia de las características por permutación

La Importancia de las Características por Permutación (PFI, por sus siglas en inglés) permite determinar la importancia de diversas variables dentro de las predicciones de un modelo. La idea central de este enfoque es que la alteración o permutación de una variable no relevante no debería afectar significativamente el rendimiento del modelo. PFI evalúa este principio mediante la comparación del error del modelo antes y después de permutar cada variable. Un incremento notable en el error tras la permutación indica la importancia de dicha variable, demostrando que su modificación impacta de manera significativa las predicciones del modelo. De este modo, PFI proporciona una medida cuantitativa de la contribución de cada variable a la precisión del modelo, resaltando las variables que son fundamentales para las predicciones.

El proceso inicia con la evaluación del error inicial del modelo, proporcionando un indicador de su eficacia basado en las variables originales y su configuración inicial. Posteriormente, se realiza la alteración de las entradas mediante la permutación de las variables, lo que implica el intercambio aleatorio de los valores de cada variable dentro de su respectiva dimensión (Fisher et al., 2019). En la Figura 3.1, se presenta un ejemplo de aplicación de la PFI a la variable de altura, registrada en metros para ciertos individuos. La PFI implica el reasignamiento aleatorio de estas alturas entre los individuos del conjunto de datos. Una vez permutadas las variables, el rendimiento del modelo se recalcula con esta nueva disposición de datos, generando un nuevo error. Teniendo en cuenta estos resultados, la PFI se calcula dividiendo el error obtenido tras la permutación de las variables entre el error inicial, cuando

las variables se mantenían sin cambios. Este cociente ofrece un indicativo del efecto de la alteración de las variables en la capacidad predictiva del modelo. Un valor elevado destaca la importancia de determinadas variables en la precisión de las predicciones efectuadas por el modelo original.

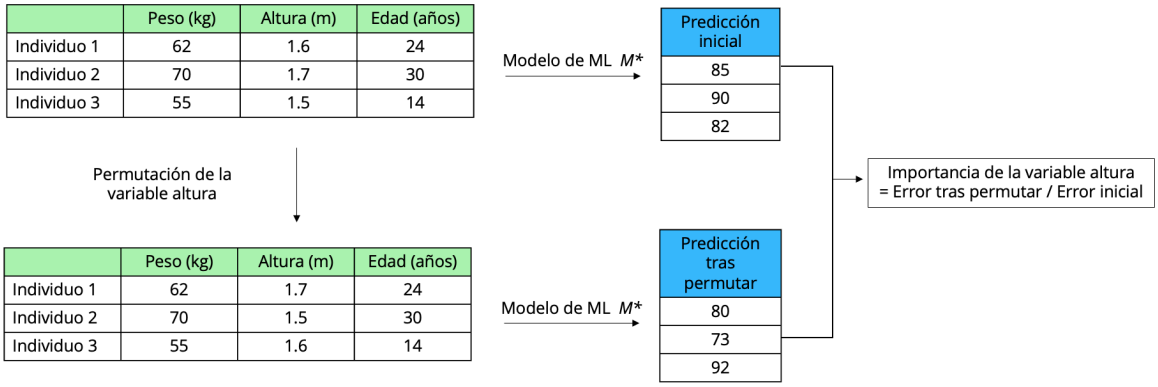



Figura 3.1: Ejemplo de PFI para la variable altura. Elaboración propia: adaptado de (Agarwal, 2022)

.

Para el cálculo de la PFI, se pueden emplear tanto los datos de entrenamiento como los de prueba. Al evaluar el error para determinar el PFI, se recomienda usar datos de prueba, dado que estos comprenden observaciones no expuestas durante el entrenamiento del modelo. Este enfoque asegura una evaluación más objetiva de la capacidad predictiva del modelo bajo condiciones no sesgadas por el conocimiento previo. Sin embargo, el análisis del PFI utilizando los datos de entrenamiento también ofrece perspectivas valiosas, permitiendo una comprensión más profunda de las variables que ejercen un impacto significativo en la formulación del modelo. Identificar estas variables es fundamental, ya que pueden ser determinantes en el proceso de aprendizaje del modelo, y su análisis contribuye a una mayor comprensión de su comportamiento, facilitando oportunidades para optimizar el rendimiento del modelo (Dankers, 2020). De este modo, ambos enfoques son recomendados para obtener una visión integral del modelo, combinando la validación externa con un análisis detallado del aprendizaje, lo cual enriquece la comprensión y la interpretación de las predicciones generadas.

Este método ofrece varias ventajas significativas. Su claridad y sencillez de interpretación lo hacen una herramienta particularmente accesible para analizar la relevancia de las variables dentro de un modelo. Utilizando el ratio de error, facilita la comparación de la importancia de variables a través de distintas aplicaciones, promoviendo la generalización de los resultados y la evaluación de modelos en variados contextos. Un beneficio adicional es la capacidad de este método para operar sin necesidad de reentrenar el modelo, permutando simplemente la variable de interés en lugar de eliminarla. Esta característica preserva la independencia del modelo frente a la manipulación de variables, contribuyendo a su robustez y capacidad de generalización.

Sin embargo, es importante reconocer que la PFI mide la importancia de las variables basándose en el rendimiento del modelo, lo cual no ofrece una indicación directa sobre la relevancia intrínseca de las variables. Además, el procedimiento de permutación introduce un factor de aleatoriedad, que podría resultar en una variabilidad incrementada de los resultados. La repetición del proceso de permutación para obtener un promedio también puede elevar el coste computacional de manera significativa. Por lo tanto, es necesario buscar un equilibrio entre la precisión del análisis y la eficiencia en el uso de recursos computacionales. Finalmente, es importante destacar que la permutación podría resultar en estimaciones de importancia de variables poco realistas en casos donde las variables no son independientes o presentan una alta correlación. La correlación entre dos variables implica que un cambio en una afecta a la otra; por lo tanto, al permutar una variable sin alterar la otra correlacionada, se podría subestimar su importancia real. Esto se debe a que no se captura el impacto combinado de estas variables en el resultado del modelo, lo cual es un requerimiento para una evaluación precisa de su relevancia (Toloşi y Lengauer, 2011).

## 3.2. Gráficos de dependencia parcial

Los Gráficos de Dependencia Parcial (PDP, por sus siglas en inglés) ofrecen una representación visual intuitiva del efecto que ejerce una variable específica sobre la predicción generada por un modelo. Para elaborar un PDP, inicialmente se genera un conjunto de datos sintético en el que únicamente la variable de interés experimenta variaciones, mientras que el resto de variables se mantienen constantes, tal y como aparecen en el conjunto de datos original. Posteriormente, se determinan las predicciones del modelo para este conjunto de datos sintético. Este procedimiento se repite, modificando sistemáticamente la variable focal y registrando las predicciones del modelo para cada uno de sus valores (Friedman, 2001).

En la Figura 3.2, se muestra un ejemplo de construcción de PDP. Por ejemplo, al investigar la influencia del número de partidos ganados en la probabilidad de ser finalista en un torneo, se generan conjuntos de datos artificiales incrementando gradualmente el número de partidos ganados desde uno en adelante. Luego, se analiza la respuesta del modelo para cada caso, computando la media de estas respuestas. Los resultados obtenidos de la media, junto con la variable que aumenta gradualmente, se utilizan para trazar el gráfico PDP, lo que permite visualizar el impacto de la variación de la variable número de partidos ganados en la predicción del modelo.En términos generales, el PDP permite comprender la variabilidad en las predicciones del modelo en respuesta a cambios en la variable analizada. Observando la Figura 3.3, se aprecia que un incremento de uno a dos en el número de partidos ganados correlaciona con un aumento del 0.1 en la probabilidad de avanzar a la final. Por otro lado, un segundo PDP que ilustra el número de lesiones revela una tendencia inversa: a medida que el número de lesiones se incrementa, se evidencia una disminución en la probabilidad de

ser finalista, reflejando el impacto negativo de las lesiones en el rendimiento del equipo.

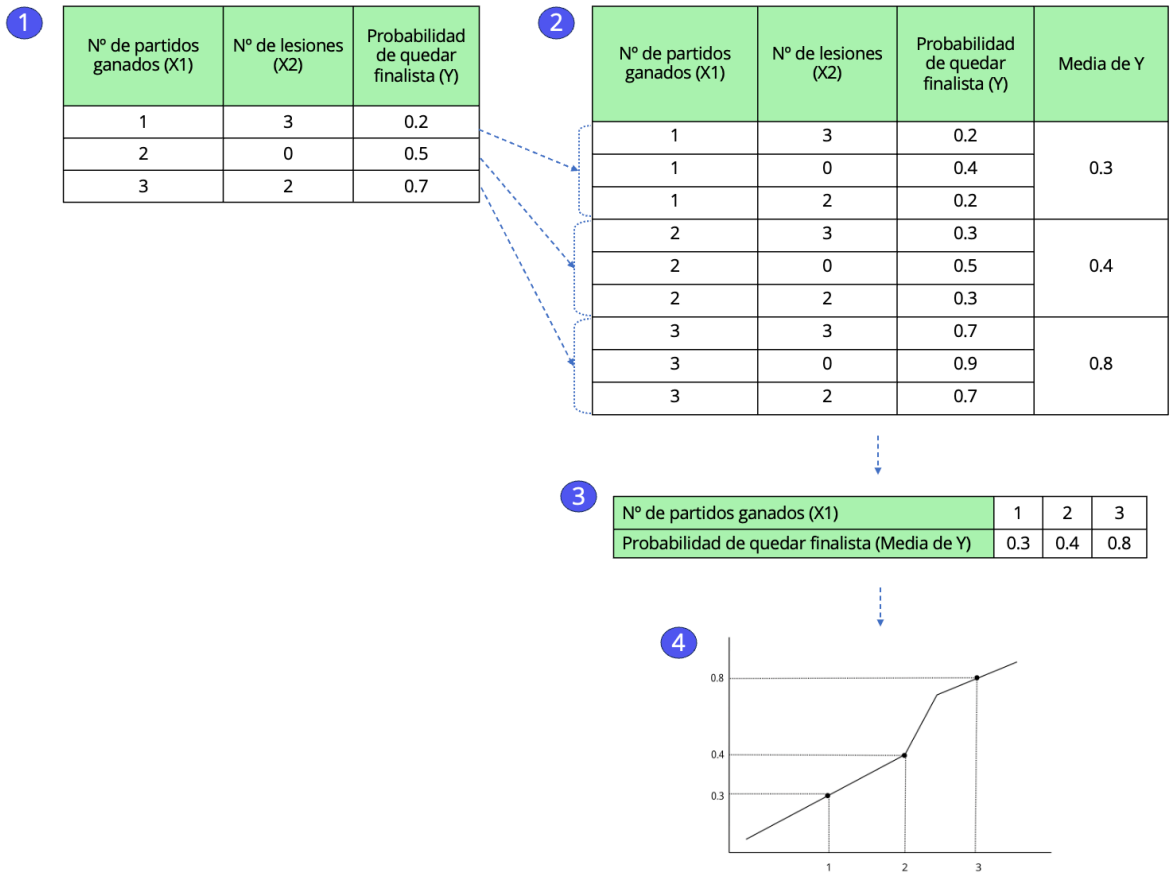



Figura 3.2: Representación gráfica de la construcción de un PDP para la variable número de partidos ganados. Elaboración propia: adaptado de (Kuo, 2022).

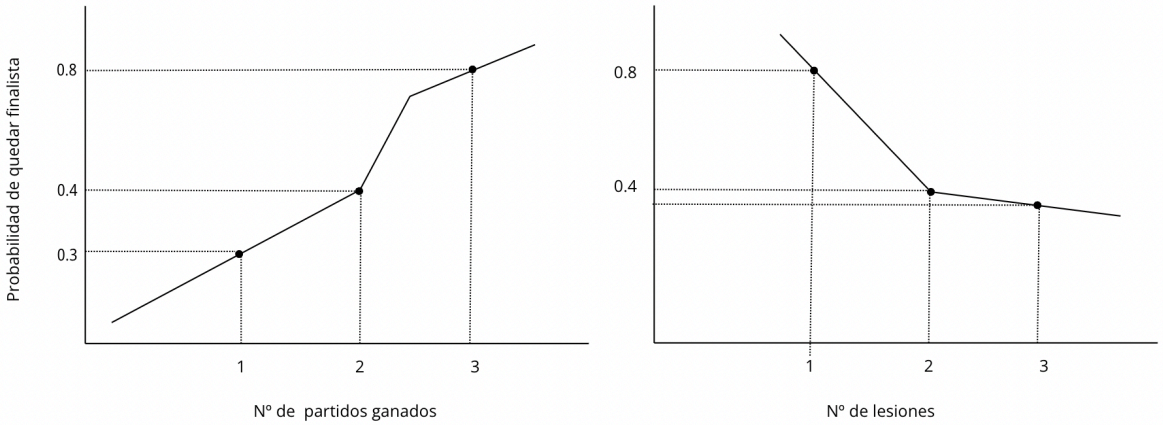



Figura 3.3: Representación gráfica de dos PDP. Elaboración propia.

La visualización resultante de un PDP detalla las variaciones en la salida del modelo a raíz de la modificación de los valores de una variable específica, mientras se mantienen in-

variables las demás. Esta representación gráfica caracteriza la naturaleza del vínculo entre la entrada y la salida del modelo para la variable en cuestión. A través de esta visualización, se puede discernir la complejidad de dicha relación, identificando si es lineal, cuadrática o incluso de mayor complejidad. Un PDP que muestra una línea plana sugiere la ausencia de relación, indicando así que la variable analizada carece de importancia (Greenwell et al., 2018). No obstante, es imprescindible abordar la interpretación de estos gráficos con prudencia, dado que parten de la premisa de independencia entre las variables, una suposición que puede no siempre corresponderse con la realidad.

A pesar de sus beneficios, los PDP presentan ciertas limitaciones. El cálculo de PDP para cada observación requiere considerar todas las combinaciones posibles de variables, un procedimiento que puede ser extenso y demandante desde el punto de vista de los recursos computacionales. La complejidad se incrementa con el aumento en la cantidad de variables y puntos de datos, lo cual puede hacer inviable su uso en conjuntos de datos de gran tamaño o complejidad. Por esta razón, se suele recurrir primero al método de permutación de variables para identificar aquellas de mayor importancia, aplicando posteriormente el PDP solo a estas variables seleccionadas. Otra desventaja significativa de los PDP es que asumen la independencia entre las variables al mantenerlas constantes y modificar sólo una de ellas, lo cual puede ser poco realista en la práctica. Frecuentemente, las variables predictoras están interrelacionadas, y esta correlación puede influir notoriamente en la precisión de las interpretaciones derivadas de los PDP. Si estas interdependencias se ignoran, los PDP podrían ofrecer interpretaciones que no solo sean sesgadas sino también inexactas, lo que restringe su aplicabilidad en ciertos contextos (Hastie, Tibshirani, y Friedman, 2009).

Además, los PDP tienen la tendencia a crear combinaciones de datos que pueden no corresponder a escenarios realistas, lo que puede resultar en una interpretación errónea de la influencia de las variables en la predicción del modelo. Por ejemplo, sería incoherente dentro de un modelo predictivo de competiciones deportivas encontrar un caso donde un equipo que ha ganado cinco partidos se encuentre en la última posición de la liga. Tales anomalías introducen distorsiones en la interpretación de los datos que podrían comprometer la fiabilidad de los PDP en la toma de decisiones informadas.

Por último, los PDP pueden ocultar efectos heterogéneos al presentar solamente los efectos marginales promedio.Si una característica presenta asociaciones positivas en ciertos segmentos del conjunto de datos y negativas en otros, la representación gráfica de la Dependencia Parcial podría indicar una línea plana. Esto se debe a que los efectos opuestos se anulan entre sí en el análisis promedio, lo que podría inducir a la conclusión errónea de que la característica es no influyente, cuando, de hecho, existen efectos compensatorios. Los efectos heterogéneos pueden ser identificados mediante el trazado de Curvas de Expectativa Condicional Individual (ICE, por sus siglas en inglés), que permiten observar cómo la característica afecta las predicciones en cada caso individual, destacando las variaciones del impacto de una característica en diferentes subconjuntos de datos (Goldstein, Kapelner, Bleich, y Pitkin, 2015). Las

ICE, que se orientan a la interpretabilidad local, aunque no son el objeto de estudio de este trabajo, proporcionan una perspectiva valiosa para detectar estos efectos heterogéneos.

## 3.3. Efectos Locales Acumulados

Aunque los gráficos PDP proporcionan una interpretación visual del efecto de una o dos variables en la predicción de un modelo, se enfrentan a notables limitaciones. Estas incluyen la producción de puntos de datos no representativos que no coinciden con la distribución auténtica de los datos y la premisa de independencia entre variables, que puede no ser realista. Dado que la influencia de una característica determinada, como $x$, en el resultado del modelo depende de los valores de las otras variables, no es viable asignar una única predicción para cada valor de $x$. Esta situación evidencia la imposibilidad de adoptar la premisa de independencia entre variables que los PDP presuponen. Por tanto, se hace necesario adoptar un enfoque que permita definir y calcular precisamente la predicción esperada para cada valor específico de $x$, considerando la interrelación estadística existente entre las variables.

El método de Efectos Locales Acumulados (ALE, por sus siglas en inglés) surge como una solución a estos desafíos (Apley y Zhu, 2020). Dicho método se enfoca en una variable $x$ de interés y calcula su influencia como la suma acumulativa de las diferencias promedio en las predicciones. Estos cálculos se obtienen de alteraciones mínimas de $x$ dentro de un rango estrecho alrededor de cada uno de sus valores, mientras se mantienen las demás variables prácticamente invariables. Esta característica distingue a ALE de los gráficos PDP: ALE promedia las predicciones dentro de una ventana limitada de valores, en contraposición a PDP, que promedia sobre todos los posibles valores, incluyendo aquellos que pueden ser irreales y sintéticos. Así, ALE destaca los efectos de cambios marginales en $x$, preservando la constancia de otros factores. Al acumular estos efectos locales, se facilita la comprensión del impacto global de las variaciones en $x$. Además, para permitir comparaciones efectivas entre los efectos de diversas variables, se suele restar el efecto promedio para que se represente como cero en el gráfico. En este contexto, la Figura 3.4 ilustra claramente la discrepancia entre las predicciones generadas por ALE y PDP, resaltando el sesgo inherente en esta última estrategia. Mientras ALE refleja con precisión una tendencia lineal, PDP muestra una variación que no corresponde a una regresión lineal pura, evidenciando su tendencia a estar sesgado por la inclusión de combinaciones de datos no plausibles. Esta distorsión es un ejemplo del sesgo que puede introducir PDP en comparación con ALE, que calcula el efecto promedio en un rango específico y estrecho, ofreciendo así una representación más confiable del efecto real de la variable en estudio.

El método ALE se destaca por sus ventajas comparativas respecto a otros métodos, como los PDP. Esta estrategia muestra una robustez superior ante la correlación entre variables, proporcionando resultados más objetivos y confiables. En contraste con los PDP, ALE evita el

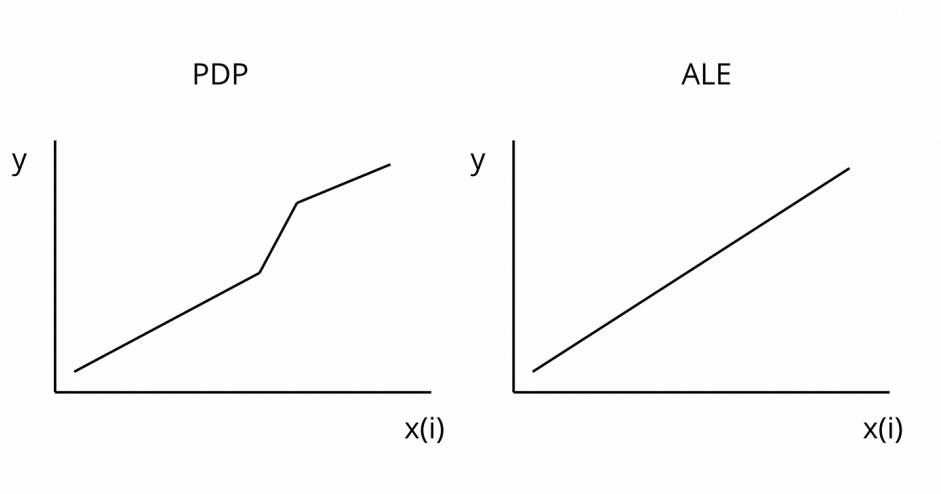



Figura 3.4: Comparación de la salida entre PDP y ALE para una variable. Elaboración propia: adaptado de (Molnar, 2019)

sesgo y no presupone la independencia entre variables. Es también más eficiente en términos computacionales, ya que se enfoca en calcular el cambio en la predicción promedio para intervalos reducidos alrededor de cada valor de la variable, en lugar de abarcar todo el rango de valores posibles, como es el caso con los PDP. Además, la interpretación de los datos obtenidos mediante ALE es intuitiva y directa, gracias al hecho de centrar la variable en cero, lo que equivale a que cada punto en la curva destaque el efecto incremental con respecto a la predicción promedio (Hassija, Chamola, y Mahapatra, 2024).

A pesar de sus beneficios, los ALE presentan ciertas restricciones. Por su naturaleza de ser datos acumulados, los efectos reflejados en las predicciones del modelo se calculan dentro de intervalos específicos, los cuales son luego acumulados para generar una representación gráfica suavizada y de fácil comprensión. ALE, por su diseño, resalta los efectos de cambios incrementales en la variable $x$, asumiendo la constancia del resto de factores. No obstante, en escenarios donde existe una fuerte correlación entre factores, la variación en una característica puede estar intrínsecamente ligada a variaciones en otra, complicando la interpretación aislada de su efecto. Esta aproximación puede conducir a interpretaciones incorrectas si no se consideran debidamente las interrelaciones. Además, ALE podría no generalizar bien los efectos específicos identificados en cada intervalo a lo largo de todo el rango de la variable (Okoli, 2024).

Asimismo, la implementación de ALE puede presentar complicaciones al establecer la cantidad de intervalos a utilizar. Un exceso de intervalos puede llevar a estimaciones volátiles, mientras que una reducción excesiva puede suponer una pérdida de detalles significativos sobre el comportamiento del modelo. Aunque ALE no provee información sobre los efectos heterogéneos que otras estrategias como las Curvas de Expectativa Condicional Individual (ICE) podrían revelar para los PDP, proporciona una perspectiva distinta sobre la influencia

de las variables. A diferencia de los PDP, que examinan el cambio en las predicciones para cada punto de datos individual, ALE segmenta los datos en intervalos y evalúa cómo varían las predicciones en cada uno de ellos. Este enfoque resulta valioso para obtener una visión general del efecto de las entradas sobre las predicciones a lo largo de diferentes rangos de valores de los predictores.

Como norma general, se considera que ALE supera a PDP en precisión, principalmente porque no presenta sesgo. A pesar de que se ha señalado que ALE puede encontrarse con dificultades ante datos altamente correlacionados, este desafío no es exclusivo de ALE; más bien, es una cuestión común en las técnicas de interpretabilidad cuando se trata de variables con fuerte correlación. En tales casos, resulta más apropiado examinar conjuntamente el efecto de ambas variables. Si el interés radica en discernir los diferentes efectos de una característica sobre varios grupos dentro del conjunto de datos, se recurriría a PDP acompañados de curvas ICE (Bodensteiner, 2020). No obstante, para comprender la influencia general de una variable en las predicciones, los gráficos ALE pueden ofrecer una visión global más representativa.

## 3.4. Interacción entre variables

La interacción entre variables es un aspecto fundamental de la capacidad predictiva de los modelos, puesto que las variaciones en una variable pueden influir en el comportamiento de otras, alterando así las predicciones finales. Por ejemplo, al predecir el precio de una casa, se examinarían variables como el tamaño de la vivienda, el número de habitaciones, la ubicación y la edad del inmueble. Teniendo en cuenta que cada una de estas variables puede afectar el precio de forma independiente, también puede existir una influencia mutua entre ellas que incida conjuntamente en la estimación del valor. Por ejemplo, se puede observar una alta correlación entre el tamaño de la casa y el número de habitaciones. Un aumento en el área de la propiedad suele estar relacionado con un incremento en el número de habitaciones, lo cual podría influir en el precio de una manera distinta que si se evaluase cada variable de forma aislada. La comprensión de esta interacción entre variables resulta fundamental para entender el comportamiento del modelo e identificar los factores que ejercen la mayor influencia en la predicción final. Una variable por sí sola puede no ser representativa, pero en combinación con otras, su relevancia puede potenciarse y convertirse en una variable significativa en el resultado final del modelo.

El estadístico $H$, desarrollado por Friedman y Popescu, se establece como una herramienta cuantitativa eficaz para medir la interacción entre variables en modelos predictivos (Friedman y Popescu, 2008). Este índice evalúa la fracción de la varianza de la función de predicción que es atribuible a la interacción de las variables. Los valores de este estadístico varían entre 0 y 1, donde un valor de 0 denota la inexistencia de interacción y un valor de 1

implica que la totalidad de la varianza de la función de predicción se debe a las interacciones entre las variables. Con el fin de entender los valores asociados al estadístico $H$, se hace necesario comprender la influencia de las interacciones entre variables en la función predictiva del modelo. Dicha función se puede conceptualizar como la agregación de los efectos individuales y de interacción de las variables. Específicamente, se compone de la suma del gráfico PDP de una variable aislada, $x_j$, más el PDP del conjunto restante de variables, que contempla la interacción de la variable $x_j$ con las demás. Esta estructura facilita la representación de todas las interacciones potenciales presentes en la función de predicción. La ecuación 3.1 representa este concepto.

$$\hat{f}(x) = PD(x_j) + PD_{-j}(x_{-j}) \tag{3.1}$$

Donde $PD(x_j)$ representa el gráfico PDP para la variable $x_j$ considerada de forma independiente, y $PD_{\neg j}(x_{\neg j})$ caracteriza el gráfico PDP para todas las variables en interacción, exceptuando $x_j$. Esta función de predicción desempeña un papel central en el cálculo del estadístico $H$, como se muestra en la ecuación 3.2.

$$H_j^2 = \frac{\sum_{i=1}^{n} \left[\hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)})\right]^2}{\sum_{i=1}^{n} \hat{f}^2(x^{(i)})} \tag{3.2}$$

El numerador cuantifica la variación debida a las interacciones, calculando la diferencia entre las predicciones del modelo y la suma de los efectos aislados y combinados de la variable $j$ con otras variables. El denominador, por su parte, representa la varianza total de la función predictiva. Por lo tanto, el estadístico $H_j^2$ se interpreta como la proporción de la varianza de la función predictiva del modelo que se atribuye a las interacciones entre la variable $j$ y las demás variables. El estadístico $H$ se destaca por varias ventajas significativas. En primer lugar, proporciona una interpretación intuitiva lo que facilita la comprensión de la relación entre las variables en el modelo. Además, al ser adimensional, facilita las comparaciones directas entre distintos modelos y variables, una cualidad de gran valor en el análisis y selección de predictores. Otra ventaja es su capacidad para analizar interacciones entre tres o más variables, lo que brinda una visión más completa de la relación entre los predictores y la variable de interés. Finalmente, el estadístico $H$ es capaz de detectar cualquier tipo de interacción presente en los datos, lo que contribuye a una modelización más precisa y detallada.

No obstante, la implementación del estadístico $H$ puede presentar inconvenientes. En primer lugar, su aplicación puede resultar computacionalmente costosa, especialmente al utilizar gráficos de dependencia parcial, ya que estos requieren evaluar todos los puntos y crear todas las posibles combinaciones de variables, como se describe en la sección 3.2. Para abordar este problema, algunas propuestas, como la sugerida por Friedman y Popescu, sugieren tomar una muestra de los datos, aunque esto puede comprometer la estabilidad del estadís-

tico $H$ (Friedman y Popescu, 2008). Además, el estadístico $H$ puede resultar engañoso en determinadas circunstancias, particularmente cuando las variables están altamente correlacionadas. Incluso si el efecto total de dos variables en la salida del modelo es insignificante, lo que conduce a una varianza y un denominador bajos, el numerador puede seguir siendo alto debido a la fuerte interacción entre estas variables debido a su alta correlación. Este escenario podría inducir a conclusiones erróneas sobre la importancia relativa de las variables, ya que el estadístico $H$ podría sugerir una influencia significativa, aunque en realidad el efecto combinado de estas variables en el resultado final sea mínimo (Molnar, 2019). Por último, aunque el estadístico $H$ indica la existencia de interacciones, no distingue el tipo específico de interacción, lo cual podría ser una limitante en contextos donde se requiere un análisis de interacción más detallado.

## 3.5. Modelos Subrogados

Un modelo subrogado sirve como una aproximación interpretable para las predicciones generadas por modelos complejos y opacos, comúnmente denominados modelos de caja negra. Un ejemplo clásico es el uso de un árbol de decisión para comprender las predicciones hechas por una red neuronal. La finalidad de este tipo de modelo es ofrecer una mayor transparencia sobre el proceso de toma de decisiones del modelo original, proporcionando información relevante sobre su funcionamiento interno. Cabe destacar que el propósito del modelo subrogado no es evaluar la calidad intrínseca del modelo de caja negra, sino traducir sus decisiones a una forma más accesible para la comprensión humana (Molnar, 2019).

En la Figura 3.5 se presenta la visualización del desarrollo de un modelo subrogado. Este proceso comienza con la selección de un conjunto de datos, que puede coincidir con el usado para el entrenamiento del modelo de caja negra o bien consistir en nuevos datos que compartan una distribución semejante. Se procede entonces a entrenar el modelo de caja negra con estos datos, obteniendo las predicciones correspondientes. Estas predicciones se utilizan como la variable respuesta '$y$' para el modelo subrogado, como por ejemplo un árbol de decisión como se visualiza en el ejemplo. La calidad de este último se evalúa a través del coeficiente de determinación R cuadrado, el cual cuantifica la proporción de la variabilidad en las predicciones del modelo de caja negra que el modelo subrogado logra explicar. El rango de R cuadrado se sitúa entre 0 y 1, donde un valor próximo a 1 indica que el modelo subrogado replica con alta fidelidad el comportamiento del modelo de caja negra. Un R cuadrado cercano a 1 sugiere que, si el modelo subrogado alcanza una precisión y resultados comparables al modelo de caja negra, sería más conveniente optar por el primero por su mayor interpretabilidad.

Los modelos subrogados se distinguen por ofrecer una amplia variedad de beneficios. Su flexibilidad es particularmente destacable, ya que posibilita la implementación de un amplio

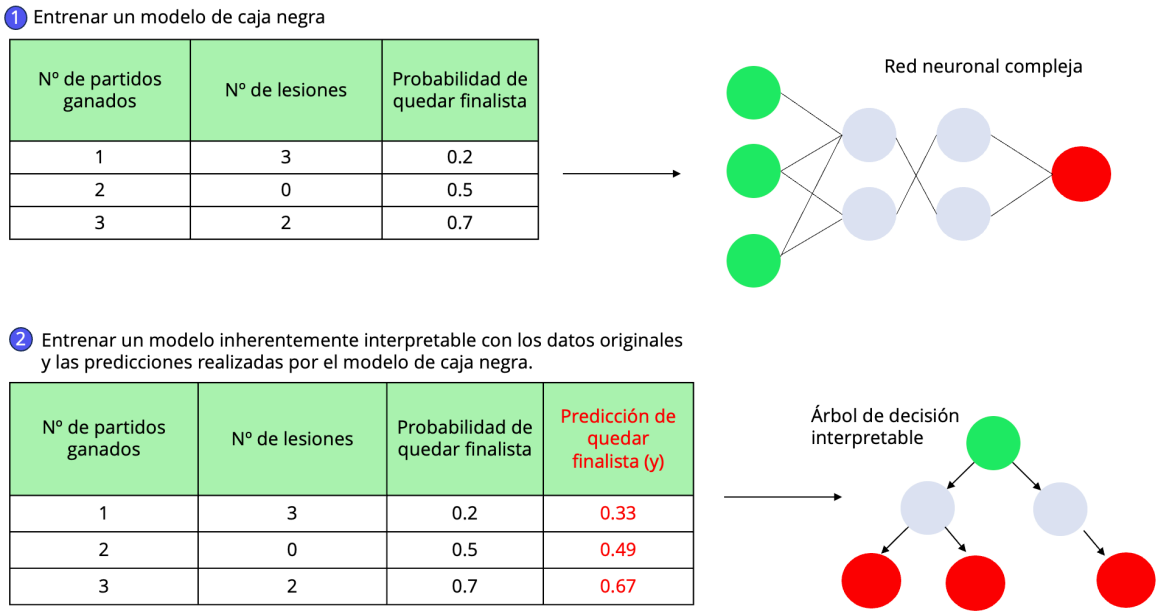



Figura 3.5: Desarrollo de un modelo subrogado. Elaboración propia: adaptado de (Hall et al., 2017)

.

espectro de modelos interpretables con el fin de entender las decisiones de un modelo de caja negra específico. Esta versatilidad es fundamental para elegir la herramienta interpretativa más idónea. Asimismo, la naturaleza intuitiva y fácil de entender inherente a los modelos subrogados los convierte en herramientas especialmente efectivas para comunicar conceptos de ML a personas con menos experiencia en el campo. Esta capacidad de traducir la complejidad del modelo de caja negra en términos más accesibles puede mejorar la colaboración interdisciplinaria y la toma de decisiones informadas (Jones y Chen, 2021). Sin embargo, es importante tener en cuenta las limitaciones asociadas con los modelos subrogados. Las interpretaciones derivadas de estos modelos se centran en el comportamiento general del modelo en sí mismo, no en predicciones específicas. Aunque un modelo subrogado puede proporcionar una buena aproximación del modelo de caja negra, no hay un criterio definitivo que determine cuándo esta aproximación es lo suficientemente precisa . Además, la calidad y precisión del modelo subrogado pueden verse influenciadas por la elección del conjunto de datos utilizado para su entrenamiento. Diferentes conjuntos de datos pueden conducir a resultados variados, lo que destaca la importancia de una evaluación cuidadosa de la idoneidad del modelo subrogado para cada contexto particular .

## 3.6. Prototipos y críticas

La selección de muestras de datos ejerce una influencia significativa durante el entrenamiento de los modelos de ML. Dado que es prácticamente imposible analizar todos los

datos, se recurre a la utilización de prototipos y críticas para representar el conjunto de datos de manera más eficiente. La figura 3.6 ilustra una distribución de datos, en la cual algunos representan los datos prototipos (rombo rojo) y otros las críticas (círculo rojo).

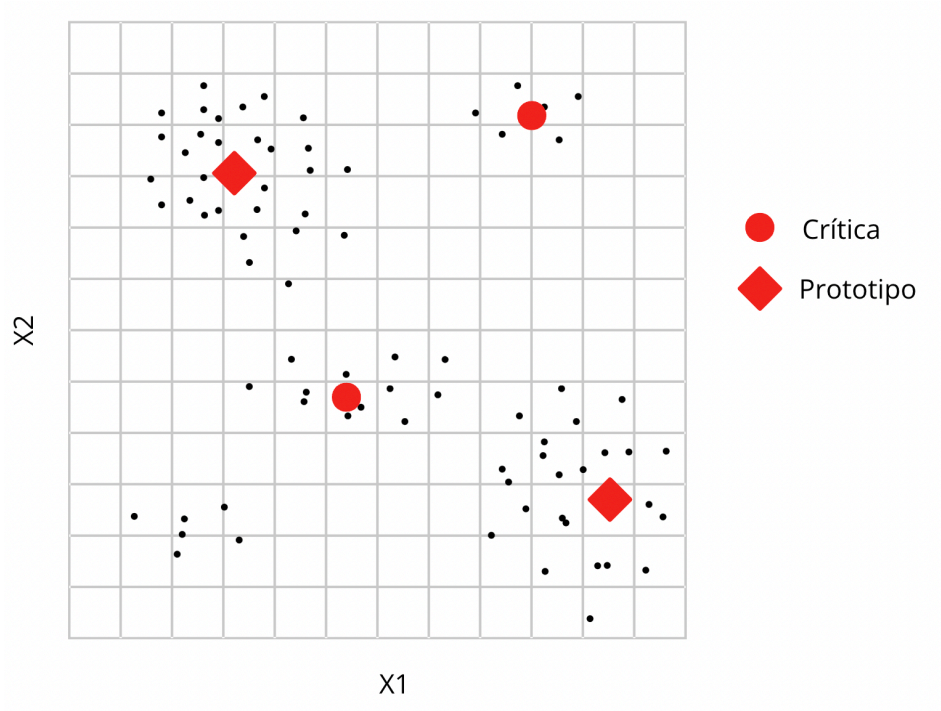



Figura 3.6: Representación de prototipos y críticas en un conjunto de datos. Elaboración propia. Adapatado de: (Molnar, 2019)

.

Los prototipos son instancias que representan de manera confiable la distribución general del conjunto, funcionando como ejemplos predominantes de la población de datos. Por otro lado, las críticas consisten en puntos de datos que no son adecuadamente representados por los prototipos. Estas son muestras que, aunque no son atípicas o *outliers*, se encuentran en regiones menos pobladas del espacio de datos, ofreciendo perspectivas sobre comportamientos poco frecuentes o sobre aspectos del modelo que podrían necesitar mejoras. Este enfoque permite una revisión crítica del modelo, asegurando que se consideren tanto los patrones comunes como las excepciones que podrían revelar vulnerabilidades o áreas de mejora en la predicción del modelo

En conjunto, prototipos y críticas actúan como un sustituto del conjunto de datos completo, ofreciendo una representación condensada pero integral que mantiene las características esenciales del conjunto original. Es fundamental reconocer la interdependencia entre prototipos y críticas; su análisis conjunto enriquece la comprensión del modelo, dado que la selección de las críticas depende del proceso de identificación de los prototipos. Esta relación asegura que la interpretación de los datos y su aplicabilidad práctica se maximicen al considerar ambos elementos en su interacción, permitiendo una visión más integral y deta-

llada del comportamiento del conjunto de datos. El proceso para construir este conjunto de datos sustituto comprende las siguientes etapas:

1. Seleccionar el número de prototipos y críticas.
2. Seleccionar los prototipos utilizando la distancia de MMD (Máxima Discrepancia de Media).
3. Seleccionar las críticas mediante la aplicación de la función "testigo" ("*witness*").

La selección de prototipos se fundamenta en el criterio de MMD, utilizando una función de kernel, simbolizada por $K$, como medida de similitud entre los puntos de datos (Kim et al., 2016). Esta función varía entre 0 y 1, donde el valor 1 señala una coincidencia perfecta entre dos puntos. La tarea de elegir un prototipo implica minimizar el valor de MMD para identificar aquel cuya distribución es más parecida a la del conjunto de datos completo. De forma similar, las críticas se seleccionan mediante la función "testigo", que mide la disparidad entre las estimaciones de densidad en ciertas localizaciones del conjunto de datos, permitiendo así identificar las muestras poco representadas por los prototipos seleccionados.

ProtoDash, una innovación reciente en el criterio MMD, ha sido desarrollada en colaboración por investigadores de IBM y Amazon (Gurumoorthy et al., 2019). Este enfoque, a diferencia de la aplicación estándar de MMD, asigna pesos a cada prototipo, permitiendo que aquellos con pesos menores sean identificados como críticas (Bodria, Giannotti, Guidotti, y et al., 2023). ProtoDash busca prototipos diversos, es decir, ejemplos que ilustren el conjunto de datos desde diferentes perspectivas para proporcionar una imagen más completa del conjunto de datos. Tras la selección del primer prototipo, el más representativo del conjunto, el algoritmo busca un segundo prototipo que, si bien comparte patrones con el primero, introduce también características distintivas. Este proceso iterativo continúa hasta completar la selección de todos los prototipos necesarios, garantizando así una cobertura representativa de la variabilidad dentro del conjunto de datos (Ronaghan, 2020).

La implementación de prototipos y críticas brinda ventajas considerables, facilitando la comprensión de conjuntos de datos complejos. La recopilación de estas muestras específicas permite obtener una percepción detallada de su naturaleza y características. Este proceso permite evaluar la presencia y relevancia de datos dentro de la distribución y verificar su coherencia con el modelo analizado. Además, permite la detección de datos con potencial para reforzar el modelo y, en el caso de identificar muestras redundantes, ofrece la oportunidad de intervenir mediante su eliminación o minimización. Sin embargo, a pesar de sus beneficios, los prototipos y las críticas también tienen limitaciones. La selección de los prototipos puede estar fuertemente influenciada por los hiperparámetros seleccionados en la función kernel. Además, es necesario determinar previamente el número de prototipos a utilizar, lo que puede resultar desafiante, ya que no existe un punto intermedio ideal (Ali et al., 2023).

# Capítulo 4

# Caso de Uso: Aplicación en la Gestión de Riesgos Financieros

Este capítulo se enfoca en la aplicación específica de las técnicas de interpretabilidad presentadas en el Capítulo 3, a un caso de estudio detallado en riesgo crediticio. El objetivo principal es obtener información relevante sobre la importancia y el impacto de diferentes variables en la predicción del riesgo crediticio, facilitando así la toma de decisiones basada en datos y mejorando la transparencia en las evaluaciones de riesgo. Se adoptará un enfoque metodológico estructurado, desglosado en las siguientes etapas, ilustradas en la Figura 4.1:

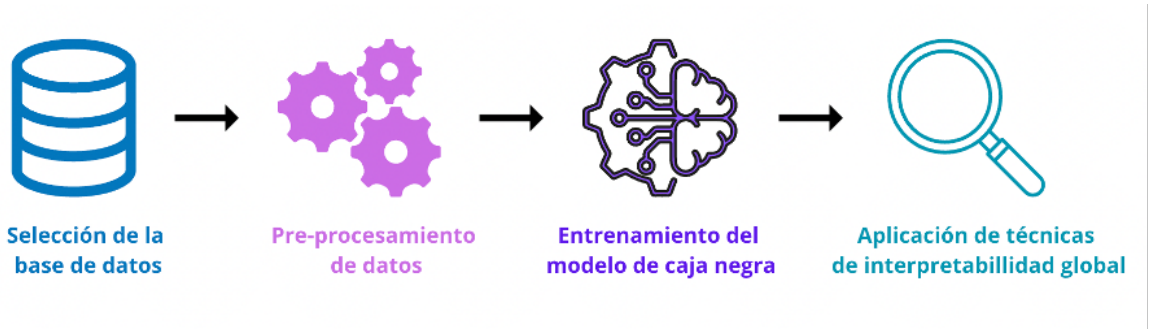



Figura 4.1: Secuencia de tareas. Elaboración propia.

1. **Selección de la base de datos:** Se optará por un conjunto de datos que proporcione un compendio de información sobre préstamos, seleccionado cuidadosamente para asegurar su alta calidad. Este conjunto incluirá variables financieras requeridas para una evaluación detallada del riesgo crediticio.

2. **Pre-procesamiento de datos**: Previo a los análisis predictivos, es necesario llevar a cabo tareas de pre-procesamiento en el conjunto de datos seleccionado. Dicho proceso incluye actividades como la limpieza de datos, la eliminación de valores atípicos, la normalización de variables y la codificación de variables categóricas. Este paso es

necesario para garantizar tanto la calidad como la consistencia de los datos, lo cual permitirá obtener resultados más confiables en análisis posteriores.

3. **Entrenamiento del modelo de caja negra:** Se llevará a cabo el entrenamiento de un modelo no interpretable, empleando el conjunto de datos previamente seleccionado. La finalidad de este proceso es que el modelo aprenda a identificar patrones y características significativas en la determinación del riesgo crediticio. Durante este procedimiento, se realizarán ajustes al modelo con el propósito de reducir la diferencia entre sus predicciones y los valores reales asociados al riesgo crediticio reportado en el conjunto de datos. Una vez completado el entrenamiento, el modelo tendrá la capacidad de efectuar predicciones sobre conjuntos de datos nuevos.

4. **Aplicación de técnicas de interpretabilidad global:** Tras finalizar el entrenamiento del modelo, se implementarán las técnicas de interpretabilidad global expuestas en el Capítulo 3. Este paso es importante para entender el funcionamiento y las decisiones subyacentes del modelo, proporcionando así una comprensión más amplia de sus mecanismos internos. A continuación, se especifican las librerías empleadas durante la implementación de cada técnica:

    - Para el enfoque PFI y los gráficos PDP, se utilizó la librería "scikit-learn".
    - Los gráficos ALE se obtuvieron con la librería "PyALE".
    - Para el modelo subrogado, en el cual se aproximó el modelo de caja negra mediante un árbol de decisión, se recurrió nuevamente a la librería "scikit-learn".
    - Los prototipos y críticas fueron implementados utilizando el algoritmo ProtoDash de la librería "AIX360". Este proceso implicó la gestión de versiones específicas de paquetes de Python para evitar conflictos con otras herramientas utilizadas en este proyecto. Para ello, se configuró un entorno virtual dedicado a manejar las dependencias necesarias para la utilización de "AIX360". Las instrucciones detalladas para la instalación de estas dependencias están disponibles en: (Trusted AI, 2019).

Para el desarrollo del código requerido en este estudio, se ha seleccionado Python, debido a su amplia aceptación y aplicación en los campos de la ciencia de datos y el ML. Este lenguaje de programación proporciona un extenso conjunto de librerías y herramientas que facilitan la aplicación de técnicas de interpretabilidad global. El código desarrollado para esta investigación se encuentra accesible en el enlace proporcionado a continuación: (Moreno, 2023)

## 4.1. Descripción y Pre-procesamiento de los datos

La transparencia en la evaluación del riesgo crediticio tiene una importancia crítica para las entidades financieras, así como para el cumplimiento de las normativas y requisitos legales, aspecto que se abordó en el Capítulo 1. En este escenario, resulta fundamental una interpretación clara de los modelos complejos de ML, comúnmente referidos como de caja negra, los cuales pueden presentar desafíos en su comprensión y albergar sesgos que afecten de forma adversa las decisiones crediticias. Las técnicas de interpretabilidad global surgen, por tanto, como herramientas valiosas para comprender la dinámica interna del modelo y de sus relaciones, garantizando así una evaluación justa y transparente del riesgo.

En este contexto, se realizará la implementación y evaluación de estrategias de interpretabilidad global empleando un conjunto de datos de crédito bancario alemán, disponible en Kaggle (Ferreira, 2023). Este conjunto de datos ha sido escogido por su relevancia como referencia en el ámbito de la gestión de riesgos financieros, destacándose por su alta calidad, la representatividad de sus muestras y su disponibilidad como recurso de acceso abierto para descarga. Esta base de datos contiene 1,000 registros y 10 atributos, incluyendo una variable objetivo binaria llamada 'Riesgo'. Cada registro del conjunto representa a un individuo solicitante de un crédito bancario. La variable 'Riesgo' categoriza a los solicitantes en términos de su riesgo crediticio, identificándolos como 'bueno' o 'malo'. Aquí, un calificativo de 'bueno' sugiere una probabilidad reducida de incumplimiento en los pagos del crédito, mientras que la etiqueta 'malo' señala una probabilidad elevada de incumplimiento.

Tras la selección del conjunto de datos para este estudio, se procedió a realizar un análisis descriptivo preliminar. Esta evaluación permitió identificar la presencia de valores faltantes en dos variables significativas: 'Checking account', con un 39,4 % de ausencias, y 'Savings account', con un 18,3 %. Dada la elevada proporción de datos faltantes en 'Checking account', se optó por excluir esta variable del análisis. La decisión se sustenta no solo en la cantidad de información ausente, sino también en la relevancia financiera de la variable 'Savings account', la cual se considera que puede ser más informativa sobre la situación financiera del solicitante que la cuenta corriente. Para manejar los valores ausentes en 'Savings account', se introdujo una nueva categoría denominada 'no_inf', asignada a todas las observaciones que carecen de datos en esta variable, asegurando así la integridad del análisis sin descartar estas observaciones por completo. Las demás variables se mantuvieron sin cambios para su uso posterior en el análisis. La Tabla 4.1 presenta estas variables, detallando su tipo y significado, lo cual constituye una información de gran utilidad para el análisis subsiguiente de interpretabilidad.

Con los registros ya completos, el tratamiento de los datos avanzó a través de varias etapas cruciales para su adecuada preparación y análisis. Inicialmente, se implementó la técnica de codificación *one-hot* en las variables categóricas, tanto nominales como binarias. Este método transforma cada categoría de una variable en una serie de vectores binarios, en los

| Variable | Tipo | Significado |
|---|---|---|
| Age | Cuantitativa discreta | Edad del solicitante en años [19-75]. |
| Sex | Cualitativa binaria | Sexo del solicitante (male, female). |
| Job | Cualitativa ordinal | Experiencia laboral del solicitante (0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled). Representa el nivel de habilidades y residencia del solicitante en su campo laboral. |
| Housing | Cualitativa nominal | Situación de propiedad de vivienda (own, rent, or free) Describe la condición de propiedad de la vivienda del solicitante. |
| Saving accounts | Cuantitativa discreta | Nivel de ahorros del solicitante (little, moderate, quite rich, rich). |
| Credit amount | Cuantitativa discreta | Cantidad solicitada en marcos alemanes [250-18.400]. |
| Duration | Cuantitativa discreta | Duración del préstamo en meses [4-72]. |
| Purpose | Cualitativa nominal | Finalidad del préstamo (car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others) . |
| Risk | Cualitativa binaria | Evaluación del préstamo (good, bad). Indica la evaluación del riesgo crediticio del solicitante, clasificado como bueno o malo. |

Tabla 4.1: Explicación de variables. Elaboración propia.

cuales cada elemento indica la presencia (1) o ausencia (0) de la categoría correspondiente. Tomando como ejemplo la variable 'Housing', que incluye las categorías (*own*, *rent*, *free*), se generaron tres columnas binarias nuevas: 'Housing_own', 'Housing_rent', y 'Housing_-free'. En estas columnas, un valor de 1 señala la presencia de la categoría específica, mientras que 0 indica su ausencia, facilitando así la interpretación y el análisis posterior de los datos en el contexto del modelo. La variable 'Job', por su parte, representa diferentes grados de experiencia laboral, por lo cual puede considerarse como una secuencia ordinal de habilidades y competencias. Para reflejar adecuadamente esta estructura jerárquica, se ha optado por tratar 'Job' como una variable cuantitativa discreta, asignando valores de 0 a 3 que abarcan desde el nivel de experiencia más básico hasta el más avanzado. Así, a los individuos con la experiencia laboral más limitada se les asigna el valor 0, mientras que aquellos con el mayor nivel de experiencia se categorizan con el valor 3. Este método de codificación no solo permite una transcripción efectiva de la experiencia laboral a un formato numérico, sino que también conserva la relación ordinal inherente entre las categorías. Por otro lado,

es relevante señalar que las variables cuantitativas discretas como ‘Age’, ‘Credit amount’ y ‘Duration’ no han necesitado modificación adicional. Estas variables ya se encuentran en formatos numéricos aptos para el análisis, lo que simplifica su integración directa en las etapas asociadas a la predicción y evaluación del modelo.

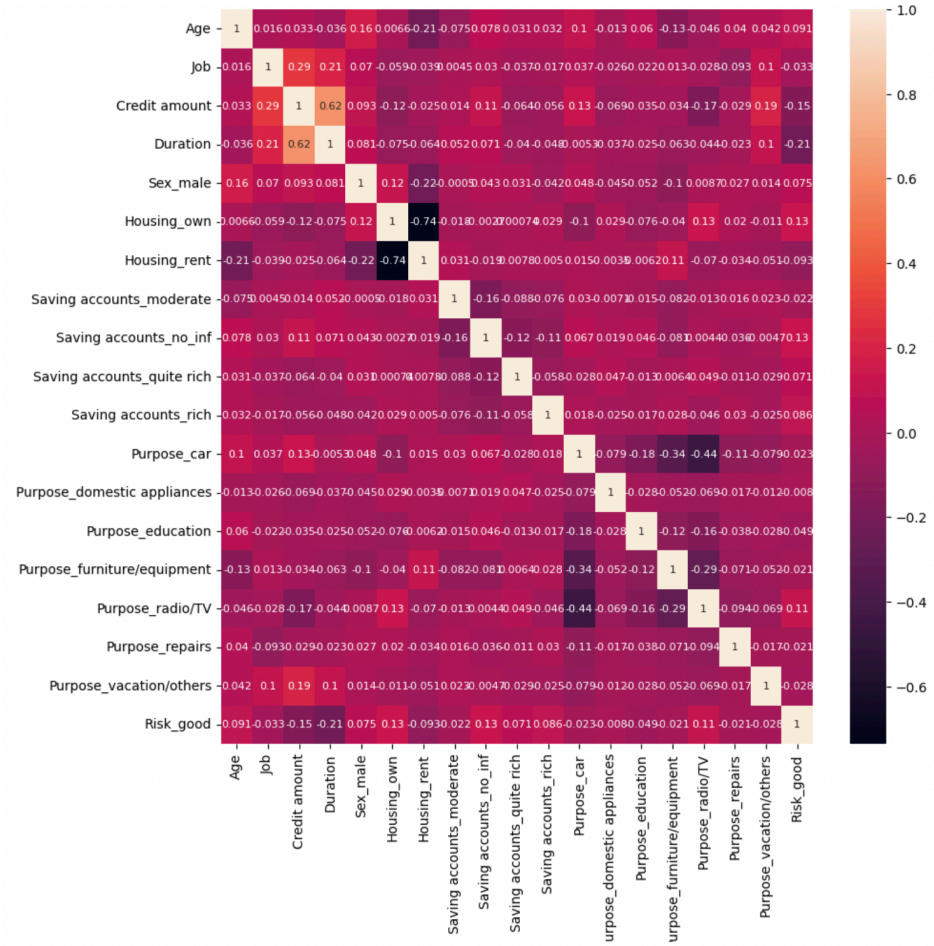



Figura 4.2: Correlación entre variables. Elaboración propia.

Posteriormente, se realizó un análisis de correlación entre variables. En la Figura 4.2 se presenta la matriz de correlaciones entre las variables, mostrando que, si bien las correlaciones no son particularmente elevadas, existe una correlación moderadamente alta de 0.62 entre la variable ‘Duration’ y ‘Credit amount’. Además, también hay una correlación relativamente alta negativa de -0.74 entre ‘Housing_own’ y ‘Housing_rent’. Esta evaluación es relevante, dado que una alta correlación entre variables puede tener implicaciones notables tanto en la construcción del modelo como en la interpretación de sus resultados. Una correlación significativa entre variables puede alterar la percepción de la importancia de cada una dentro del modelo, conduciendo potencialmente a interpretaciones incorrectas sobre los elementos que ejercen influencia sobre el resultado objetivo. Además, una fuerte correlación impacta en la capacidad de interpretar los resultados de manera clara. Por ejemplo, con respecto a los gráficos PDP, una correlación pronunciada entre variables puede complicar la interpretación

precisa de estos perfiles debido a que presuponen la independencia entre variables, tal como se discutió la Sección 3.2. En situaciones de alta correlación, estos gráficos pueden perder su utilidad analítica, limitando su contribución al entendimiento del modelo. La identificación de este fenómeno permite tomar medidas correctivas o interpretativas adecuadas para mantener la integridad del análisis.

Con los análisis previos en mente, el siguiente paso fué el entrenamiento de un modelo de caja negra destinado a predecir la clasificación de los préstamos en términos de riesgo, como bueno o malo. En este contexto, la división del conjunto de datos se efectuó en dos segmentos principales: un conjunto destinado al entrenamiento y otro a la prueba. De esta manera, se asignó el 70 % de los datos al conjunto de entrenamiento, proporcionando así un volumen suficiente para que el modelo aprenda los patrones inherentes y ajuste adecuadamente sus parámetros. Por su parte, el conjunto de prueba, constituido por el 30 % restante de los datos, se empleó para evaluar la capacidad del modelo de generalizar a datos nuevos, no expuestos durante su entrenamiento. Una vez finalizado el entrenamiento, se procedió a contrastar las predicciones del modelo con las etiquetas originales de los datos, clasificadas previamente como préstamos de riesgo bueno o malo. Esta comparación permitió determinar posteriormente la precisión y efectividad del modelo.

## 4.2. Entrenamiento del modelo de caja negra

En la siguiente fase de este estudio, se procedió al entrenamiento de un modelo no interpretable para la predicción del riesgo crediticio. Este proceso implicó la utilización de técnicas avanzadas de ML, escogiendo un algoritmo que, si bien ofrece un alto grado de precisión y capacidad predictiva, también presenta un nivel de complejidad que dificulta la interpretación directa de sus mecanismos de decisión internos. De esta manera, se eligió el *ensemble* XGBoost como modelo clasificador de caja negra, dada su probada eficiencia en el análisis de datos transversales, superior incluso a la de algunos algoritmos de aprendizaje profundo. Esta selección se fundamentó en la capacidad de XGBoost para manejar de manera eficiente las diversas dimensiones de los datos crediticios, permitiendo capturar y modelar las complejas interacciones entre variables que son críticas para la evaluación del riesgo. Las estrategias de ensembles, que integran diversos modelos para potenciar el rendimiento predictivo y la capacidad de generalización, juegan un papel importante en el ámbito del ML(Shwartz-Ziv y Armon, 2022).

XGBoost, un ejemplo destacado de este enfoque, perfecciona iterativamente su rendimiento mediante el ajuste sucesivo a los residuos del modelo previo, utilizando para ello algoritmos de optimización de gradiente que buscan minimizar una función de pérdida específica. Este método se caracteriza por combinar múltiples "*weak learners*" (aprendices débiles) para conformar un "*strong learner*" (aprendiz fuerte), capaz de realizar prediccio-

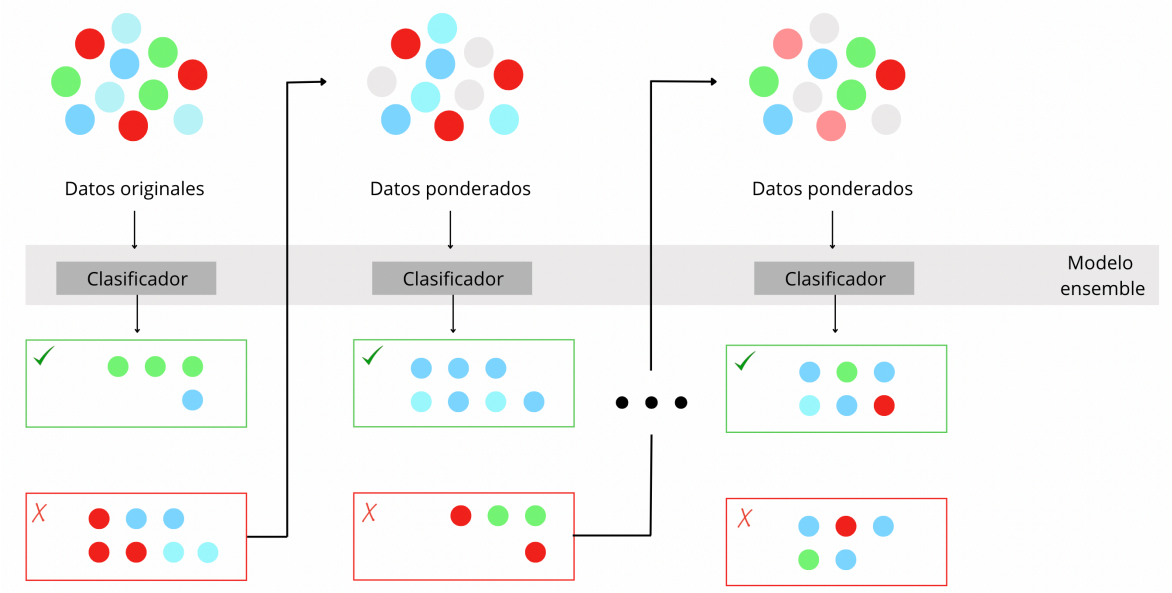



Figura 4.3: Funcionamiento de XGBoost. Elaboración propia. Adaptado de: (*XGBoost Working*, 2023)

nes con alta precisión (Chen y Guestrin, 2016). En sintesis, el funcionamiento del XGBoost se basa en el entrenamiento secuencial de árboles de decisión, cada uno enfocado en diferentes segmentos de datos, con el objetivo de corregir los errores de los árboles anteriores. La combinación de las predicciones individuales de estos árboles culmina en la generación de la predicción final del modelo. Este procedimiento de ensamble no solo contribuye a la precisión de las predicciones sino que también mejora la robustez del modelo, resultando en un sistema de evaluación del riesgo crediticio más confiable y estable ante variaciones en los datos, en comparación con métodos que dependen de un solo modelo. El algoritmo descrito previamente se ilustra en la Figura 4.3, donde se muestra el inicio del proceso de entrenamiento de XGBoost con un árbol de decisión inicial. Tras cada ciclo de entrenamiento, el modelo emplea los pesos otorgados a los datos para desarrollar y calibrar un nuevo árbol de decisión. Al concluir este proceso iterativo, se consolida un modelo conjunto (*ensemble*) integrando todos los datos y ajustes realizados. Esta secuencia de entrenamiento y actualización continua de los árboles de decisión permite incrementar la precisión predictiva del modelo conjunto, aprovechando la capacidad colectiva de los árboles individuales.

En este contexto, se procedió al entrenamiento de un modelo *ensemble* utilizando XGBoost. Para la selección óptima de sus hiperparámetros, se empleó una técnica de búsqueda en grilla, resultando en una configuración de 100 árboles, una tasa de aprendizaje de 0.05, una profundidad máxima de 3 y un mínimo de 2 muestras para dividir los nodos, como la más efectiva. Con esta configuración específica, se alcanzó una precisión del 72.7 % en el conjunto de prueba, lo cual se alinea satisfactoriamente con los objetivos de este estudio. Es importante destacar que, aunque la optimización de la precisión predictiva es relevante, el

interés primordial de esta investigación se centra en la interpretación de los resultados más que en la maximización de su precisión. Asimismo, se debe destacar que los modelos ensemble, como XGBoost, se categorizan frecuentemente como modelos de "caja negra" debido a su complejidad y la opacidad en sus mecanismos de decisión. La naturaleza de un modelo ensemble, que integra múltiples submodelos, introduce una capa adicional de complejidad y dependencias cruzadas entre las variables. Esta característica complica significativamente la tarea de discernir la influencia específica de cada variable en la predicción final. De esta manera, la combinación de varios modelos puede oscurecer las dinámicas subyacentes entre las variables de entrada y las salidas del modelo, transformando al conjunto en una entidad cuya lógica interna no es del todo transparente o directamente interpretable.

## 4.3. Análisis de interpretabilidad global

### 4.3.1. Importancia de las características por permutación

En el contexto de los resultados predictivos obtenidos, se procedió al análisis de interpretabilidad iniciando con la evaluación de la Importancia de Características por Permutación (PFI). Este enfoque, aplicado tanto al conjunto de entrenamiento como al de prueba, se ilustra en la Figura 4.4. Es recomendable enfocarse en el conjunto de prueba para esta evaluación, dado que representa datos no expuestos previamente al modelo, proporcionando así una perspectiva imparcial sobre su capacidad de generalización. Sin embargo, evaluar también la permutación en el conjunto de entrenamiento es interesante, ya que permite entender la forma en la que el modelo interactúa con los datos sobre los cuales fue ajustado, ofreciendo información adicional sobre su funcionamiento interno.

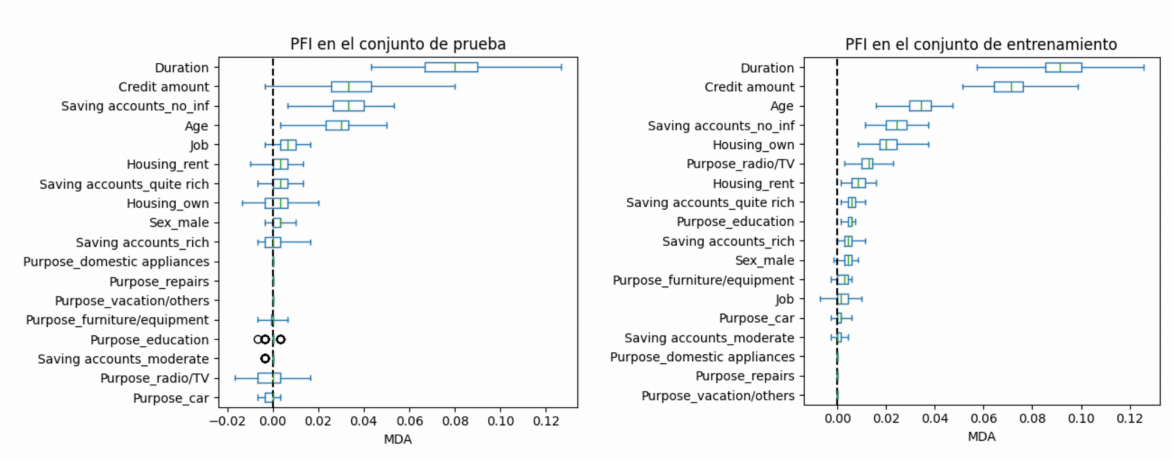



Figura 4.4: PFI obtenido con el modelo de clasificación XGBoost para el conjunto de prueba y de entrenamiento. Elaboración propia.

Los resultados de la PFI se calcularon para las 18 variables de entrada del modelo, ejecu-

tando la permutación de cada una cien veces para asegurar la obtención de resultados fiables y consistentes. La Magnitud de Disminución Media de la Precisión (MDA, por sus siglas en inglés), concepto introducido por Breiman (Breiman, 2001), se representa en el eje horizontal. En contraste, en el eje vertical, se clasifican las variables en orden ascendente según su importancia determinada tras las permutaciones. De este modo, la variable duración se posiciona como la más influyente. En pocas palabras, tras permutar los valores de 'duración' cien veces y calcular la disminución promedio en la precisión, esta se destaca como el factor más crítico en la clasificación del riesgo crediticio. Con un MDA aproximado de 0.08 en el conjunto de prueba, este resultado indica que la precisión del modelo disminuyó en un 8 % de media cuando los valores de esta variable fueron permutados al azar.

El segundo predictor de mayor relevancia, según el análisis de importancia de características en el conjunto de prueba, corresponde a la cantidad de crédito solicitado, que presenta un MDA de 0.04. Le siguen la falta de información sobre la cuenta de ahorros y la edad del solicitante. Conjuntamente, estas cuatro variables contribuyen a un MDA acumulado del 19 % en el conjunto de prueba. Bajo la suposición de independencia de las variables, la eliminación de estas cuatro variables del modelo resultaría en una reducción del 19 % en su precisión, destacando así su relevancia para la predicción del modelo y su impacto significativo en la capacidad predictiva del mismo. A pesar de que en los datos de entrenamiento, la ausencia de información sobre la cuenta de ahorros no parece ser tan determinante como en el conjunto de prueba, estas mismas cuatro variables son identificadas como las más influyentes en los dos conjuntos de datos.

Además, es importante tener en cuenta que la implementación de esta estrategia de interpretabilidad global puede conducir a variaciones en los resultados obtenidos en cada simulación. La variabilidad de estos resultados queda evidenciada en el diagrama de caja, que muestra los valores máximos, mínimos y la media de la disminución en la precisión. En el conjunto de entrenamiento, esta variabilidad suele ser menor debido a la familiaridad del modelo con los datos, pero en el conjunto de prueba, la dispersión es notablemente mayor. Se destaca que, aunque las variables 'Duratión' y 'Credit amount' se destacan como las más relevantes, también presentan la mayor volatilidad en sus MDA al alterar sus valores. Esto sugiere que las fluctuaciones en estas variables específicas tienen un impacto considerable en la estabilidad del modelo y su precisión predictiva.

Una conclusión notable que se deriva de este análisis es la prelación de variables como la duración del préstamo, la cantidad de crédito solicitado, la edad del solicitante y la ausencia de información sobre la cuenta de ahorros en la predicción del riesgo. Esta observación es coherente con los principios fundamentales de la teoría de riesgos porque la duración del préstamo y la cantidad solicitada están directamente relacionadas con la capacidad de endeudamiento y la carga financiera sobre el solicitante, influenciando así su probabilidad de incumplimiento. De igual manera, la edad puede ser indicativa de la estabilidad financiera y la experiencia en la gestión de crédito, mientras que la falta de información sobre la cuenta de

ahorros podría señalar una menor transparencia o estabilidad financiera del solicitante. Estas variables, por tanto, ofrecen perspectivas relevantes sobre la solvencia y el comportamiento financiero del individuo, justificando su relevancia en modelos que buscan evaluar el riesgo crediticio.

Finalmente, se debe recordar que este enfoque de interpretación asume que la relación entre las variables y la predicción es lineal y aditiva, lo cual puede no reflejar completamente las interacciones complejas y no lineales presentes en la realidad. Esta limitación sugiere que, aunque el análisis proporciona información relevante sobre la importancia de las variables, la comprensión completa de su influencia en las predicciones del modelo podría requerir evaluación más detallada de las dinámicas subyacentes.

### 4.3.2. Gráficos de dependencia parcial y Efectos Locales Acumulados

Procediendo con el análisis de interpretabilidad, se examinaron las Gráficos de Dependencia Parcial (PDP) y los Efectos Locales Acumulados (ALE) para las variables identificadas como más significativas en el análisis PFI del conjunto de prueba. La implementación mediante la librería 'scikit-learn' plantea retos en términos de interpretabilidad, especialmente debido al escalado de datos efectuado durante el preprocesamiento para el entrenamiento del modelo de caja negra. Este ajuste previo complica la percepción directa del impacto que las variaciones en los datos originales podrían tener en las predicciones del modelo, dificultando la interpretación de la relación entre cambios en las entradas y variaciones en las salidas. A pesar de esta complejidad, tanto los PDP como los ALE logran evidenciar la naturaleza de la relación entre cada variable y la predicción del modelo, ilustrando si esta relación es, por ejemplo, lineal, exponencial o de otro tipo. De estas visualizaciones se pueden derivar conclusiones sobre el impacto de cada variable sobre el resultado final, permitiendo una comprensión más detallada de sus relaciones.

En la Figura 4.5 se presentan los gráficos PDP, los cuales ilustran la relación marginal entre cada variable y la salida del modelo. Se muestran un total de cinco gráficas, cada una representando el efecto marginal de una variable sobre la salida del modelo. En el eje $x$ se encuentran los valores de entrada de cada variable, mientras que en el eje $y$ se muestra el impacto en la salida del modelo.

Se observa que la variable 'Duration' ejerce una influencia significativa en las predicciones, evidenciando una relación decreciente negativa. En otras palabras, a medida que aumenta la duración del crédito, la probabilidad de calificar el riesgo como bueno tiende a disminuir. Esto puede reflejar que préstamos con términos más largos podrían estar asociados con un mayor riesgo, según el modelo. Por el contrario, la variable 'Age' refleja una tendencia creciente, indicando que la probabilidad de una calificación de riesgo favorable mejora con el aumento de la edad de los solicitantes, hasta alcanzar un nivel en el cual el efecto tiende a nivelarse, lo que podría interpretarse como una asociación en el modelo de una mayor esta-

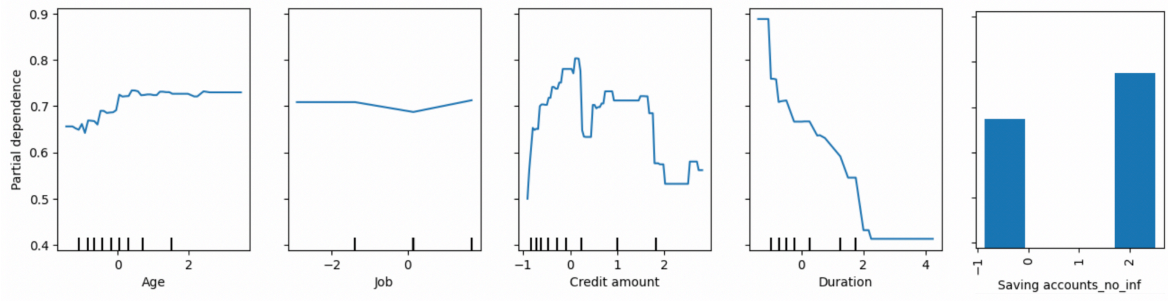



Figura 4.5: PDP generado para las variables más importantes acorde a PFI. Elaboración propia

bilidad o fiabilidad financiera con clientes de mayor edad. Las marcas verticales en la base del gráfico señalan la distribución de 'Age' dentro del conjunto de datos, ilustrando la frecuencia de las diferentes edades de los solicitantes. Una concentración más densa de marcas en los valores inferiores evidencia que existe una proporción más significativa de solicitantes jóvenes.

Contrastando con 'Age', la variable 'Job' muestra una influencia relativamente constante a través de su rango, indicando que la ocupación del solicitante, de acuerdo a la codificación incluida en la etapa de pre-procesado, no varía significativamente la salida del modelo. Esto podría interpretarse como que la ocupación del solicitante no es un predictor significativo de la variable objetivo en el modelo, o que cualquier efecto que tenga es constante independientemente de los cambios en 'Job'. El PDP para 'Credit amount' muestra una correlación no lineal entre la cantidad de crédito solicitada y las predicciones del modelo, con patrones que indican una influencia diferencial a lo largo de varios rangos de crédito. Esto implica que cambios en la cantidad de crédito impactan de forma variable las predicciones, lo que podría reflejar percepciones distintas de riesgo asociadas con diferentes niveles de crédito solicitado. Las oscilaciones observadas podrían señalar interacciones entre 'Credit amount' y otras variables, añadiendo una capa de complejidad al comportamiento predictivo del modelo y sugiriendo que la relación entre la cantidad de crédito y la probabilidad de un buen riesgo crediticio no es consistente a través de todo el espectro de cantidades de crédito.

Con respecto a la variable binaria 'Saving accounts_no_inf', la escala del eje x en la gráfica de dependencia parcial ha sido ajustada como resultado del preprocesamiento, con valores de -1 representando casos en los cuales hay información disponible sobre las cuentas de ahorro del solicitante (equivalente a 'Saving accounts_no_inf' = 0), y valores de 2 para casos sin información disponible (equivalente a 'Saving accounts_no_inf'= 1). Se observa que la probabilidad de aprobación del crédito es más baja cuando hay información disponible sobre las cuentas de ahorro del solicitante (valor -1 en el eje x), mientras que es más alta cuando tal información no está disponible (valor 2 en el eje x). Esto podría interpretarse como que no tener información sobre las cuentas de ahorro está asociado con una mayor probabilidad de que el crédito sea valorado positivamente, lo cual es contraintuitivo y podría

necesitar una investigación adicional para entender este comportamiento en los datos.

Como conclusión, los PDP revelan el efecto discriminante de las variables explicativas al mostrar las variaciones en la salida del modelo ante cambios específicos en una sola variable, lo cual es necesario para comprender las relaciones en el análisis predictivo. Sin embargo, la presencia de colinealidad entre las variables puede perturbar las relaciones capturadas por este enfoque. En caso de una alta correlación entre variables, los PDP pueden ofrecer información incorrecta, por lo que una alternativa recomendable son los gráficos ALE.

Los gráficos ALE, que se muestran en la figura 4.6, más precisa al incorporar la dependencia entre las variables, ilustrando el impacto localizado de una variable mientras se controlan las demás. Aunque los PDP y ALE pueden ser teóricamente similares cuando no hay colinealidad o interacción entre las variables, en la práctica, pueden exhibir diferencias significativas, como se evidencia en esta figura. A diferencia de los PDP, los ALE consideran las interacciones locales entre las variables, proporcionando una interpretación contextualizada de sus efectos. Por ejemplo, el gráfico ALE para la variable 'Duration' presenta una tendencia negativa, indicando que un aumento en la duración del crédito está asociado con una disminución en el efecto sobre la predicción del modelo. La línea azul muestra la trayectoria del efecto ALE de la duración sobre la salida del modelo, mientras que el área sombreada representa el intervalo de confianza del 95 %, lo cual proprociona una idea de la certeza estadística de estos efectos estimados. En comparación con el PDP de 'Duration' visto anteriormente, ambos muestran una relación negativa, pero el ALE ofrece una visión más localizada del efecto, resaltando la variación del impacto de la duración del crédito en diferentes segmentos de su distribución. Por ejemplo, podría interpretarse que para valores bajos de 'Duration', hay mayor certeza sobre el impacto (como lo sugiere el intervalo de confianza más estrecho), pero conforme aumenta la 'Duration', el efecto negativo en la predicción se vuelve más pronunciado, conun intervalo de confianza más amplio. Esta interpretación se alinea con la noción de que préstamos con términos más largos representan un mayor riesgo.

El gráfico ALE para la variable 'Age' sugiere que existe una correlación positiva entre la edad de los solicitantes de crédito y la probabilidad de recibir una calificación de crédito positiva hasta alcanzar un cierto umbral de edad. Esta tendencia puede reflejar la percepción de un riesgo crediticio menor asociado con solicitantes más maduros, posiblemente debido a una mayor estabilidad financiera o historial de crédito más largo. Sin embargo, el efecto de la edad sobre la probabilidad de aprobación se estabiliza más allá de este punto, lo que indica que incrementos adicionales en la edad no influyen significativamente en la evaluación del riesgo crediticio del modelo. Este resultado puede interpretarse como que el modelo no atribuye automáticamente un riesgo menor a solicitantes de mayor edad más allá de una cierta edad. El intervalo de confianza es amplio para edades menores, lo que muestra una alta incertidumbre en estas estimaciones iniciales. A medida que aumenta la edad, este intervalo se vuelve más estrecho, reflejando una mayor confianza en las predicciones del modelo para solicitantes de mayor edad.

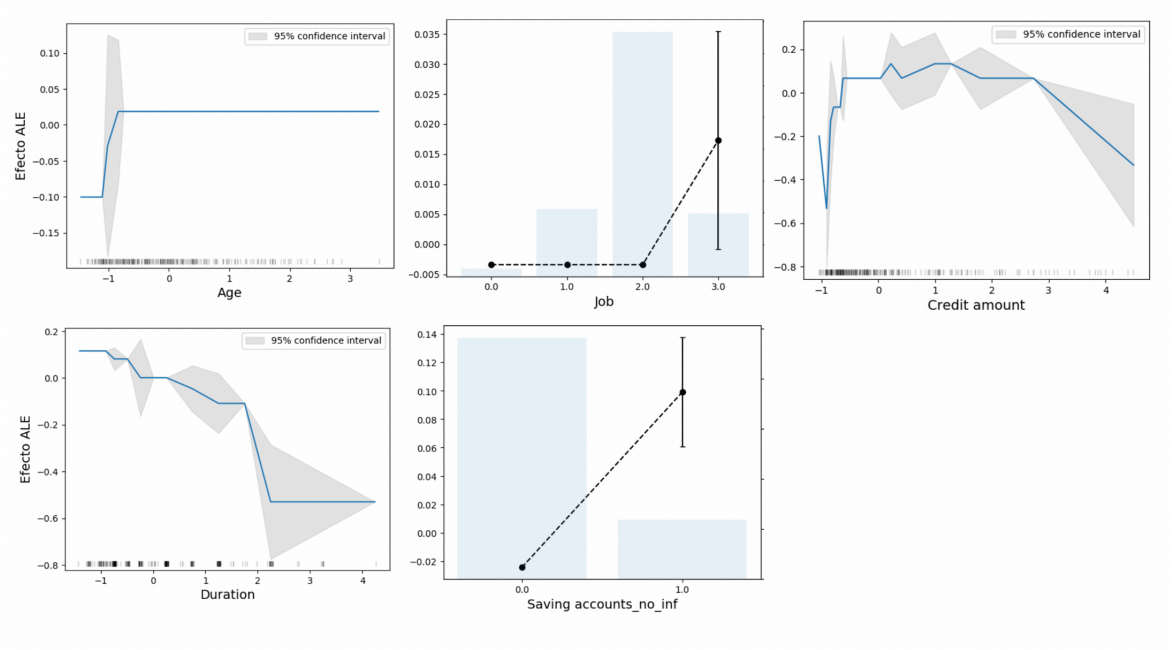



Figura 4.6: Gráfico ALE generado para las variables más importantes acorde a PFI. Elaboración propia.

En cuanto a 'Credit Amount', los gráficos PDP y ALE difieren de manera notable en la forma y la magnitud de la influencia sobre la salida del modelo, destacando la necesidad de utilizar ALE sobre PDP para entender las influencias reales de esta variable en el modelo.La correlación de 0.62 con 'Duration', ilustrada en la Figura 4.2, podría estar distorsionando los resultados del PDP, pero los ALE permiten observar una visión más ajustada de la realidad. El gráfico ALE muestra un patrón más complejo, que podría interpretarse como la existencia de puntos donde incrementos en la cantidad de crédito solicitada afectan de manera diferente la probabilidad de aprobación. Inicialmente, se observa que incrementos en la cantidad de crédito se correlacionan con un aumento en la probabilidad de otorgar un crédito clasificado como bueno. Esto podría interpretarse como una percepción de solvencia o fiabilidad asociada con solicitudes de crédito de mayor envergadura, hasta cierto punto. Tras alcanzar un determinado umbral de 'Credit Amount', la probabilidad de aprobación desciende. Esta disminución podría reflejar una política de evaluación de riesgo más conservadora para cantidades de crédito más altas, donde el modelo podría empezar a considerar esos préstamos como potencialmente riesgosos. Asimismo, este resultado podría ser un indicativo de que el modelo detecta un umbral donde el nivel de endeudamiento se vuelve preocupante, independientemente de otros indicadores positivos que pueda tener el solicitante. En general, el intervalo de confianza es más estrecho al principio, lo que implica una menor incertidumbre en la influencia de cantidades de crédito más bajas sobre la clasificación de riesgo. A medida que el monto de crédito aumenta, el intervalo de confianza aumenta, mostrando mayor incertidumbre en las estimaciones del modelo para cantidades solicitadas considerablemente

altas.

La discrepancia entre los ALE y los PDP para la variable 'Job' también es significativa. Mientras que los PDP sugieren una influencia marginal o inexistente de las distintas categorías laborales en la probabilidad de otorgamiento de crédito, el análisis ALE revela un patrón distinto. Las categorías iniciales (etiquetadas como 0, 1 y 2 en el eje x) muestran un impacto mínimo en la predicción del modelo. No obstante, la inclusión en la categoría 'Job' 3, que representa profesiones altamente especializadas, muestra una asociación pronunciada con un incremento en la probabilidad de aprobación de crédito. Esta categoría sugiere una percepción de mayor estabilidad financiera y habilidad económica, la cual tiene un reflejo positivo en la evaluación del riesgo crediticio por parte del modelo. Además, la amplia barra de error para la categoría 'Job' 3 muestra una variabilidad o incertidumbre considerable en la estimación de su impacto. Estos resultados destacan la importancia de priorizar los resultados proporcionados por los ALE frente a los PDP cuando se interpreta el papel de 'Job' en las decisiones crediticias, ya que ofrecen una perspectiva más detallada del fenómeno estudiado.

En relación con la variable binaria 'Saving accounts_no_inf', el análisis mediante gráficos ALE muestra comportamientos distintivos en función de la disponibilidad de información sobre las cuentas de ahorro. Se observa un efecto marginal cuando existe información disponible (valor 0 en el eje x), lo cual implica que la sola existencia de datos de ahorros no altera sustancialmente las probabilidades de aprobación de crédito del modelo. Por otro lado, la falta de información (valor 1 en el eje x) incrementa la influencia sobre la probabilidad de aprobación, alineándose con las observaciones del PDP, aunque en promedio la influencia es reducida en los casos sin información de cuentas de ahorros. La amplitud del intervalo de confianza indica una variación significativa en la estimación del efecto de esta característica, señalando que la ausencia de información sobre las cuentas de ahorro introduce un grado de incertidumbre en la valoración del riesgo crediticio.

Resulta contraintuitivo, pero el análisis ALE sugiere que la ausencia de información sobre las cuentas de ahorro de los solicitantes de crédito incrementa la probabilidad de aprobación por parte del modelo. Esto podría reflejar una dinámica en la que la falta de datos específicos conduce a una valoración más favorable debido a mecanismos implícitos del modelo o a patrones en los datos de entrenamiento. Tal vez el modelo asocia la falta de información con características positivas no explícitas o bien se ajusta a tendencias históricas en los datos donde la ausencia de esta información no ha sido un indicador de riesgo crediticio elevado. Esta interpretación resalta la necesidad de complementar los análisis de intrepretabilidad con el fin de entender la lógica subyacente desde diversas perspectivas, con el fin de evitar conclusiones erróneas y fundamentar las decisiones crediticias en criterios justos y equitativos.

### 4.3.3. Modelos subrogados

Para continuar con el análisis de interpretación global del modelo XGBoost, se optó por la utilización de un modelo subrogado, que replica de forma simplificada la estructura del modelo original manteniendo sus atributos más relevantes para el análisis. Un árbol de decisión fué empleado como tal modelo subrogado, entrenado para emular las predicciones del modelo XGBoost utilizando el mismo conjunto de entrenamiento. Los hiperparámetros utilizados para entrenar este árbol de decisión fueron: una profundidad máxima de 100, un número mínimo de muestras requerido para ser una hoja del nodo de 2, el criterio utilizado para medir la calidad de una división de 'gini', y un número mínimo de muestras requeridas para dividir un nodo interno de 2. La fiabilidad del modelo subrogado se ha medido a través del coeficiente de determinación R cuadrado, reflejando la proporción de la variabilidad de las predicciones originales del XGBoost que el modelo subrogado logra explicar. Se ha obtenido un valor de 0.78 para este coeficiente, lo que indica una muy buena capacidad para aproximar las predicciones del modelo de caja negra al utilizar los mismos datos de entrenamiento. El modelo subrogado, representado en la Figura 4.8, proporciona una visualización de la toma decisiones del árbol de clasificación. El punto de partida del árbol comienza en la parte superior con un nodo raíz en el que se presenta la totalidad de la muestra y a partir de este el árbol comienza a tomar decisiones y a dividir los datos hasta llegar a las hojas, donde se muestran las predicciones finales.

Este árbol inicia su análisis discriminando entre riesgo crediticio bueno y malo basándose en la variable de duración, con un punto crítico de 1.128 que corresponde a un período de 34.49 meses en la escala original. Si la duración es menor o igual a este umbral, el árbol dirige su atención a la situación habitacional del solicitante, específicamente si alquila su vivienda. En el caso afirmativo, se vuelve a considerar la duración del préstamo. Por otro lado, si no es arrendatario, el árbol prioriza el monto del crédito solicitado como el siguiente factor determinante. Cuando la duración del préstamo supera el umbral de 34.49 meses, el criterio de evaluación cambia a la edad del solicitante, fijando una demarcación en 26 años, que corresponde al valor desescalado de -0.796. Para aquellos individuos menores de 26 años, la información sobre la cuenta de ahorros se convierte en el factor decisivo. En cambio, para los individuos mayores de este límite de edad, la posesión de una vivienda surge como el siguiente aspecto crítico. Las variables que inicialmente dirigen el flujo de decisiones del árbol son, por tanto, la duración del crédito, la edad, la situación de vivienda del solicitante, la cantidad de crédito solicitado y la disponibilidad de información sobre ahorros. En etapas subsecuentes, el árbol puede integrar factores adicionales como el propósito del crédito, el género y la ocupación del solicitante. El árbol es muy profundo ya que se aproxima al modelo

Se ha llevado a cabo un análisis de importancia de variables utilizando la métrica de impureza 'gini', con el fin de determinar las variables con mayor influencia en las divisiones del árbol de decisión. La importancia de las variables, obtenida a través del modelo de árbol

de decisión subrogado, se presenta en la Figura 4.7. En el gráfico, las variables se ordenan de menor a mayor importancia en el eje vertical, mientras que en el eje horizontal se representa la medida de importancia de estas variables según el criterio 'gini'. Se considerará más importante una variable si es utilizada frecuentemente para tomar decisiones. De esta manera, las tres variables de mayor importancia son la cantidad de crédito, la duración y la edad. La cantidad de crédito es la característica más determinante, con una relevancia que supera el 0.3, lo que destaca su efecto preponderante en la toma de decisiones del modelo. La duración ocupa el segundo lugar, con una importancia un poco mayor a 0.25, mostrando que el tiempo del préstamo también es un factor relevante para el modelo. Siguiendo en importancia, está la edad del solicitante, con un valor ligeramente superior a 0.2, mostrando asimismo su relevancia en las predicciones del modelo. Otras variables, como las relacionadas con la vivienda, las cuentas de ahorro, el propósito del crédito, el empleo y el sexo, proporcionan una importancia reducida en la capacidad predictiva del modelo.

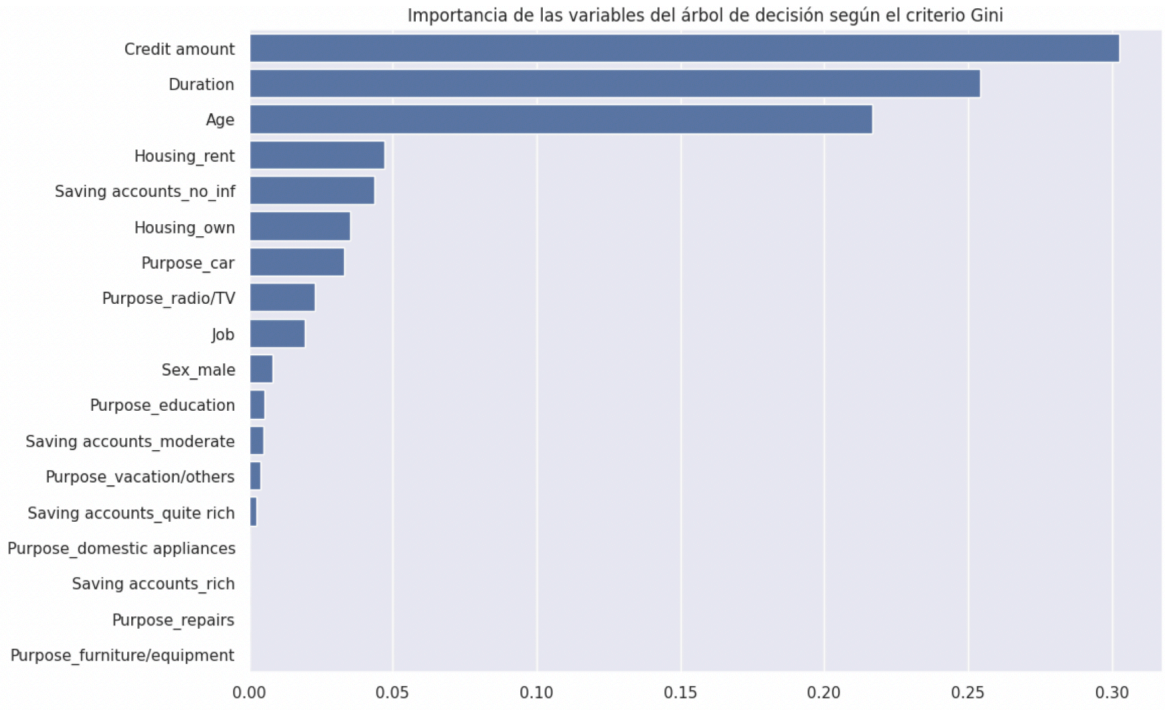



Figura 4.7: Importancia de las variables del modelo subrogado según el criterio gini. Elaboración propia.

El análisis utilizando el modelo subrogado concuerda con los resultados de PFI realizados en la sección 4.3.1, indicando una jerarquía similar en la influencia de las variables, aunque no en un orden idéntico. Esto se debe a que el árbol de decisión se aproxima al modelo, pero no es igual. La importancia otorgada a la cantidad de crédito, la duración y la edad, como factores principales, alinea ambos análisis, lo que financieramente tiene sentido ya que estos elementos son pilares fundamentales en la evaluación del riesgo crediticio. El monto solicitado en un crédito actúa como un indicio inmediato del nivel de compromiso financiero que

el cliente está dispuesto a asumir. Un crédito mayor generalmente lleva asociado un mayor riesgo de incumplimiento, ya que implica una mayor carga en términos de intereses y pagos mensuales. Por su parte, la duración del préstamo tiene implicaciones en la carga financiera a lo largo del tiempo; préstamos con plazos más cortos pueden sugerir que el solicitante tiene la confianza y la capacidad de cumplir con sus obligaciones financieras de manera más rápida, lo que puede interpretarse como un indicador positivo de solidez financiera.

La edad del solicitante, por otro lado, aporta al modelo una estimación de la experiencia en la gestión de finanzas personales. Con la edad, a menudo viene un historial más largo y diversificado de crédito, lo que puede brindar a los prestamistas un indicador sobre el comportamiento de pago histórico del solicitante. Una edad más avanzada también puede correlacionarse con un mayor desarrollo profesional y, por ende, con una mayor estabilidad en los ingresos. Además, factores como la vivienda y la información sobre ahorros son coherentes con la evaluación tradicional del riesgo, ya que proporcionan perspectiva sobre la solvencia y la capacidad de ahorro del individuo. Por ejemplo, poseer una vivienda puede indicar un grado de estabilidad y compromiso a largo plazo. Aunque estos resultados son coherentes con el análisis de PFI y el R cuadrado obtenido indica una buena adaptación del modelo subrogado, es importante recordar que no reflejan el comportamiento exacto del modelo. El modelo subrogado es, después de todo, una simplificación y puede no capturar todas las sutilezas del modelo XGBoost. No obstante, se ha determinado que el nivel de precisión alcanzado es indicativo de un modelo adecuado, el cual puede ser considerado como una base confiable para la toma de decisiones crediticias.

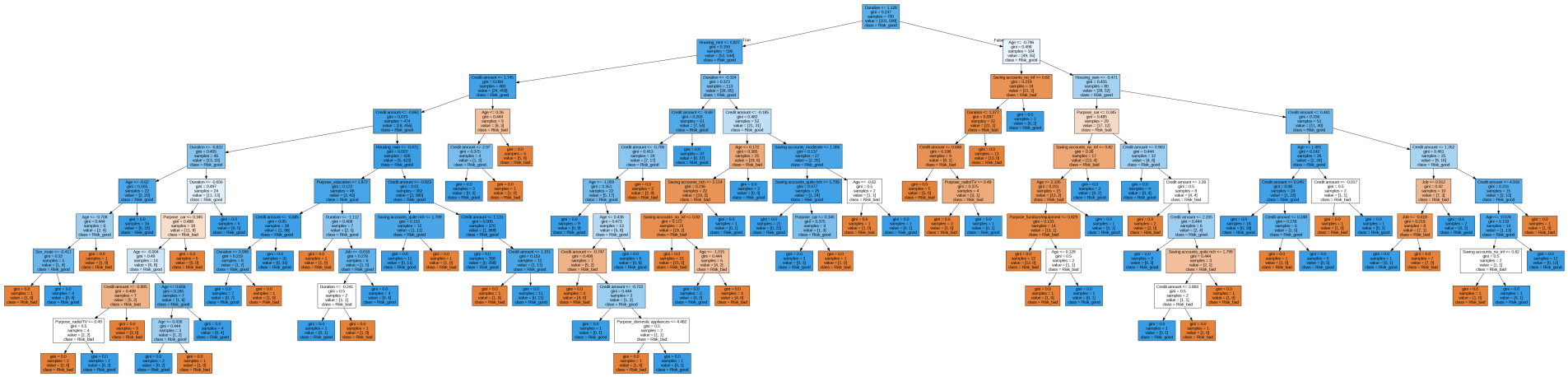


Figura 4.8: Modelo subrogado generado para aproximar el modelo XGBoost. Elaboración propia.

### 4.3.4. Prototipos y críticas

El proceso de interpretación global del modelo incluye un análisis mediante prototipos y críticas. Los prototipos representan instancias del conjunto de datos que capturan la distribución principal y las tendencias generales, mientras que las críticas ilustran las excepciones o las singularidades dentro de los datos. A través del algoritmo Protodash y como se observa en la Figura 4.9, se seleccionaron diez prototipos y críticas, cada uno con un peso asignado que refleja su grado de representatividad. Los pesos asignados a los prototipos son comparativamente más elevados, con un destacado 41 % de importancia en el caso del prototipo correspondiente al índice 264, lo que destaca su relevancia en la muestra (enmarcado en rojo en la figura). Esta metodología permite resaltar tanto los casos típicos como aquellos que se desvían de la norma, lo que permite obtener una evaluación completa del modelo.

| | Age | Job | Credit amount | Duration | Sex_male | Housing_own | Housing_rent | Saving accounts_moderate | Saving accounts_no_inf | Saving accounts_quite rich | Saving accounts_rich |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **321** | 32.0 | 2.0 | 1938.0 | 24.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **342** | 25.0 | 2.0 | 3213.0 | 18.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| **682** | 33.0 | 2.0 | 1478.0 | 15.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **220** | 26.0 | 1.0 | 625.0 | 12.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **264** | 32.0 | 1.0 | 1231.0 | 10.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **208** | 21.0 | 1.0 | 6568.0 | 24.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **81** | 47.0 | 2.0 | 1213.0 | 15.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| **82** | 24.0 | 1.0 | 1568.0 | 18.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **328** | 31.0 | 2.0 | 4473.0 | 36.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **563** | 37.0 | 2.0 | 12389.0 | 36.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

| | Purpose_car | Purpose_domestic appliances | Purpose_education | Purpose_furniture/equipment | Purpose_radio/TV | Purpose_repairs | Purpose_vacation/others | Pesos |
|---|---|---|---|---|---|---|---|---|
| **321** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.00 |
| **342** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.00 |
| **682** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.13 |
| **220** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.04 |
| **264** | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.41 |
| **208** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.04 |
| **81** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.11 |
| **82** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.04 |
| **328** | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.18 |
| **563** | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.07 |

Observación más representativa (41%)

Figura 4.9: Prototipos generados para las observaciones objeto de estudio. Elaboración propia.

La observación de prototipo más destacada según el algoritmo Protodash corresponde a un individuo de 32 años, solicitante de un crédito por 1231 marcos (equivalentes a aproximadamente 637 euros), con un término de devolución de 10 meses. Este prototipo, que posee vivienda propia, se caracteriza por ser hombre, carecer de experiencia laboral previa y no residir en la localidad de referencia. Asimismo, registra un saldo bajo en su cuenta de ahorros y proyecta utilizar el préstamo para la compra de un vehículo. En el análisis de la tabla de prototipos y críticas, se percibe una diversidad en la característica ‘Age’, predominando los individuos con edades cercanas a los 30 años. Por su parte, las variables ‘Credit Amount’ y

'Duration' muestran un rango extenso y diverso de valores. Resulta notorio que, dentro de la categoría 'Sexmale', la representación femenina es mínima, con un único prototipo femenino que posee un 4 % de peso, reflejando así la proporción de género presente en la base de datos general (aproximadamente el 30 %).

En el conjunto de datos representativo se observa una predominancia de propietarios de viviendas, lo que refleja la tendencia de propiedad del 70 % presente en la base de datos completa. Además, la característica de saldo bajo en las cuentas de ahorro de los solicitantes es una constante, alineada con el 60 % de la población de la base de datos original que presenta este atributo. Es destacable que casi la mitad del peso total asignado a los prototipos se concentra en aquellos solicitantes que buscan financiamiento para la adquisición de un vehículo. Esta inclinación se complementa con la observación de que una proporción significativa de los solicitantes se inclina hacia la compra de bienes de consumo electrónico, como televisores o radios, consistentes con el 34 % y 28 %, respectivamente, de los propósitos de préstamo en la base de datos completa. Además, la categorización laboral en este conjunto representativo muestra una frecuencia en las categorías 1 y 2, mientras que las categorías 0 y 3 aparecen con menos regularidad. Este patrón de ocupación podría sugerir una correlación entre ciertos tipos de empleo y la evaluación de la solvencia crediticia. Estos patrones destacan la importancia de considerar factores demográficos y socioeconómicos al evaluar el riesgo crediticio, así como de revisar críticamente la equidad de género en la modelización de datos financieros.

### 4.3.5. Resumen de los resultados

Tras aplicar cinco técnicas distintas de interpretabilidad a nivel global, es importante resumir los resultados más relevantes para entender mejor el funcionamiento del modelo de predicción de riesgo crediticio. A diferencia de un enfoque local, que se centra en casos individuales, este análisis global permite obtener una visión general de la relación entre las diferentes características y la clasificación final del riesgo crediticio.

En el análisis global de prototipos destacados, se observa una diversidad significativa en las características de edad, predominando solicitantes cercanos a los 30 años. Los montos de crédito y la duración del préstamo varían considerablemente, reflejando una amplia gama de necesidades y capacidades financieras entre los solicitantes. La representación de género muestra una predominancia masculina, con una limitada inclusión femenina que refleja una proporción de aproximadamente el 30 % en la base de datos general. La mayoría son propietarios de vivienda y poseen saldos bajos en cuentas de ahorro, indicando una capacidad de ahorro reducida. Además, la distribución de las categorías laborales en este grupo muestra una mayor frecuencia en ciertas ocupaciones de poca experiencia laboral.

El análisis a través de la permutación de variables y la observación de prototipos y críticas revelan que las variables más influyentes en la evaluación del riesgo crediticio son la duración

del crédito, el monto del crédito solicitado, la edad del solicitante, y la ausencia de información sobre ahorros. Por otro lado, el modelo subrogado, diseñado para simplificar y replicar el comportamiento del modelo de caja negra, confirma la importancia de estas mismas variables. La estructura y decisiones del árbol de decisión subrogado destacan especialmente la cantidad de crédito solicitado, la duración del préstamo y la edad del solicitante. Esta coherencia entre los métodos subrogado y de permutación de variables refuerza la validez de estas variables como determinantes clave en la toma de decisiones crediticias, consolidando su papel central en la predicción y manejo del riesgo crediticio. Los gráficos PDP y ALE, aportan un mayor valor en cuanto a la relación entre las variables de entrada y la salida del modelo. Estos métodos complementan los resultados previos, proporcionando información más detallada de la incluencia de cada variable en el resultado final del modelo.

En el análisis de la cantidad de crédito solicitado PDP muestra una relación no lineal, indicando que distintos niveles de crédito afectan de forma variada la percepción del riesgo. El análisis ALE refuerza este punto al mostrar que, aunque los incrementos iniciales en el monto del crédito aumentan la probabilidad de aprobación, esta probabilidad disminuye después de alcanzar un cierto límite. Este patrón, que refleja un incremento inicial seguido de una disminución en la probabilidad de aprobación con el aumento en la cantidad solicitada, responde a criterios financieros establecidos para la evaluación de solvencia. Un monto solicitado dentro de un rango moderado puede ser percibido como indicativo de una necesidad financiera justificada y una planificación responsable, reduciendo así el riesgo para el prestamista. Por el contrario, solicitudes de crédito por montos elevados pueden ser interpretadas como señales de un posible sobreendeudamiento o de inversión en proyectos de alta incertidumbre, incrementando la probabilidad de impago. Para la variable duración, los análisis sugieren que cuanto mayor es la duración del crédito, menor es la probabilidad de obtener una calificación de riesgo positiva. Este fenómeno responde a principios financieros, donde créditos de mayor duración representan un riesgo prolongado para el prestamista debido al potencial incremento en la incertidumbre financiera del prestatario a lo largo del tiempo, lo que puede afectar su capacidad de cumplimiento.

Respecto a la edad, tanto el análisis PDP como el ALE muestran una correlación inicialmente positiva con la calificación de crédito favorable, que se estabiliza al alcanzar un umbral de edad. En el gráfico de ALE para edad, el intervalo de confianza es inicialmente amplio para edades menores, reflejando una alta incertidumbre en estas estimaciones iniciales. A medida que la edad aumenta, el intervalo se estrecha, indicando una mayor confianza en las predicciones del modelo para solicitantes de edad avanzada. Este cambio en el intervalo de confianza se explica por la percepción de mayor riesgo asociado a solicitantes jóvenes, quienes a menudo carecen de un historial crediticio sólido y experiencia financiera. Además, este intervalo de confianza más amplio en las edades cercana a los 30 puede estar asociado a que hay una mayor cantidad de datos en ese intervalo, expuesto en los prototipos. En contraste, los solicitantes de mayor edad suelen presentar una estabilidad financiera acumulada que reduce

el riesgo desde la perspectiva del prestamista. Además, la mejora en la confianza del modelo con la edad avanzada refleja una menor incertidumbre en las predicciones, apoyándose en historiales crediticios más largos y comportamientos de pago consistentes.

En relación con la variable trabajo, el método ALE proporciona más información que el PDP. Destaca que las categorías de trabajo altamente especializadas están vinculadas con una mayor probabilidad de aprobación de crédito, lo cual refleja una percepción de mayor estabilidad financiera y competencia económica, factores que se traducen en una evaluación positiva del riesgo crediticio. Sin embargo, es importante tener en cuenta que el modelo podría estar sesgado ya que no dispone de suficientes datos sobre categorías laborales más especializadas. En los prototipos y críticas la mayoría de los datos están asociados a trabajos poco especializados, y la categoría 3 no está representada. La falta de representación equitativa en los datos puede llevar a sobreestimar o subestimar la importancia de ciertas profesiones en la toma de decisiones crediticias, afectando la precisión y la equidad del modelo. Finalmente, la ausencia de información detallada sobre las cuentas de ahorro presenta un desafío interesante en el análisis del riesgo crediticio. Aunque normalmente la presencia de ahorros actúa como un indicador de solvencia, el análisis ALE sugiere paradójicamente que la falta de esta información puede resultar en una valoración más favorable. Esto puede deberse a la forma en que el modelo ha sido entrenado, interpretando los patrones de datos históricos donde la ausencia de detalles sobre ahorros no necesariamente se ha asociado con un mayor riesgo crediticio. Este resultado sugiere que la representación insuficiente de esta variable podría llevar a conclusiones erróneas sobre la solvencia de un solicitante. La variable más representativa en el conjunto de datos, que indica ahorro bajo, podría estar influyendo también en esta percepción, destacando el efecto de la escasez de datos completos y detallados en la precisión y la justicia de las evaluaciones crediticias.

# Capítulo 5

# Conclusiones

La interpretabilidad de los modelos de ML es un tema de gran relevancia, especialmente en sectores de alto riesgo como el financiero, donde se requiere claridad e imparcialidad no solo por ética, sino también por cumplimiento legal. En este trabajo de fin de grado, se han empleado cinco técnicas de interpretabilidad global para entender las decisiones de un modelo de XGBoost, diseñado para predecir una clasificación de préstamos clasificados como buenos o malos. Estas técnicas, que incluyen la Importancia de Características Permutadas (PFI), los Gráficos de Dependencia Parcial (PDP), los Efectos Locales Acumulados (ALE), modelos subrogados, y análisis de prototipos y críticas, han permitido identificar las variables clave y su influencia en las predicciones del modelo, así como comprender con mayor profundidad los datos utilizados. Con este propósito, se seleccionó una base de datos referencial en el ámbito financiero, se entrenó un modelo de "caja negra" para clasificar los préstamos según su nivel de riesgo y se realizaron las predicciones pertinentes. Posteriormente, se aplicaron las técnicas de interpretabilidad global mencionadas anteriormente, facilitando un análisis detallado del modelo tanto desde una perspectiva técnica como financiera. Este enfoque integral no solo permitió comprender el funcionamiento interno del modelo, sino que también posibilitó el análisis de coherencia entre las predicciones del modelo y los principios fundamentales de la gestión de riesgos financieros.

El análisis mostró que variables como el monto del crédito y la duración del préstamo son relevantes para la evaluación del riesgo crediticio. Se constató que los préstamos con montos elevados generalmente conllevan mayores riesgos debido al incremento del riesgo financiero en caso de incumplimiento por parte del prestatario. Por otro lado, los préstamos a corto plazo son percibidos como menos riesgosos, dado que el periodo de exposición al riesgo de impago es más corto. En contraste, los préstamos a largo plazo presentan menores probabilidades de aprobación, reflejando la incertidumbre acerca de la estabilidad financiera futura del solicitante. Además, la edad del solicitante se destaca como un factor determinante; los análisis indican que las personas más jóvenes suelen ser financieramente más vulnerables, afectando así su evaluación de riesgo.

Además, los datos muestran que individuos en torno a los 30 años con bajos ahorros y empleos de baja o media cualificación predominan en la muestra utilizada para el entrenamiento del modelo, lo que proporciona una visión general de los patrones predominantes en los datos. Los análisis ALE aportan una perspectiva adicional, destacando relaciones no lineales entre las variables y las evaluaciones de riesgo. Por ejemplo, para la variable del monto del crédito, se observa que incrementos iniciales en el monto del préstamo aumentan la probabilidad de aprobación hasta cierto punto, tras el cual, aumentos adicionales resultan en una disminución de la probabilidad. Esta relación no lineal sugiere que existe un punto óptimo de solicitud de crédito que los prestamistas consideran manejable y menos arriesgado. Similarmente, para la duración del préstamo, los análisis ALE muestran que créditos más largos incrementan progresivamente el riesgo percibido, lo cual coincide con principios financieros básicos que asocian mayor duración con incertidumbre aumentada y mayores desafíos en prever la estabilidad financiera a largo plazo del solicitante.

Sin embargo, el modelo de riesgo crediticio no siempre muestra resultados consistentes, lo que destaca su complejidad interna. Un ejemplo notable es que la falta de información detallada sobre las cuentas de ahorro a veces resulta en evaluaciones de crédito positivas, contrario a lo que se esperaría. Este resultado inesperado podría indicar que el modelo está sobrevalorando otros elementos del perfil del solicitante. Este fenómeno podría estar relacionado con la distribución desigual presente en la base de datos entre las distintas clases de las variables categóricas. El modelo podría no estar extrayendo información valiosa de variables como las cuentas de ahorro o el tipo de empleo, especialmente cuando hay más datos de ciertas clases en unas variables que en otras. Además, se han identificado ciertas limitaciones de las técnicas reportadas en la interpretación de los datos, particularmente en relación con el escalado de variables. Para el entrenamiento de modelos, es necesario modificar la escala original de los datos, lo que afecta directamente las técnicas de interpretabilidad utilizadas y a los resultados obtenidos, dificultando su interpretación. Otro aspecto problemático es la alta correlación entre variables clave, como el monto del crédito y la duración del préstamo. Esta correlación afecta negativamente a la técnica de interpretabilidad utilizada PDP. Los modelos PDP asumen que las variables son independientes entre sí, por lo que cuando existen correlaciones significativas, los resultados pueden no ser confiables. A pesar de los desafíos inherentes, los modelos de *scoring* crediticio continúan siendo herramientas indispensables para las instituciones financieras, agilizando significativamente el proceso de evaluación de crédito y facilitando la labor de los banqueros. Resulta interesante observar que estos modelos logran identificar eficazmente las variables tradicionalmente consideradas más relevantes en la banca, tales como el monto del crédito, la duración del préstamo y la edad del solicitante. Estas capacidades no solo optimizan los procesos, sino que también refuerzan la precisión, transparencia y eficiencia en la toma de decisiones financieras.

En este contexto, para futuros trabajos, sería beneficioso que los modelos de ML incorporasen nuevas variables como el historial de residencia, los pagos de utilidades y el uso de

tarjetas de crédito o préstamos en otros bancos. Aunque la adquisición de estos datos puede presentar desafíos debido a su sensibilidad, ofrecen una perspectiva más amplia del comportamiento financiero del solicitante y pueden revelar patrones de comportamiento significativos que mejoren la predicción de riesgos crediticios. Una estrategia efectiva para recopilar estos datos podría ser permitir que los solicitantes los proporcionen de manera voluntaria, mejorando así sus posibilidades de obtener financiación. Además, dada la complejidad que introduce la inclusión de numerosas variables categóricas en los modelos, como "Purpose" y "Savings account", que aumentan significativamente la dimensionalidad del conjunto de datos, sería útil explorar técnicas de reducción de dimensionalidad. Aplicar métodos como el Análisis de Componentes Principales (PCA) podría ser ventajoso para reducir la dimensionalidad manteniendo la mayor cantidad de información posible. Finalmente, se recomienda investigar técnicas de agrupación que permitan gestionar el desequilibrio en la distribución de las clases de variables categóricas, que puede influir negativamente en la fiabilidad de los modelos. La implementación de métodos como el sobremuestreo sintético podría contribuir a equilibrar estas distribuciones, potenciando la precisión y la capacidad de generalización de los modelos, y promoviendo decisiones más justas.

## Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

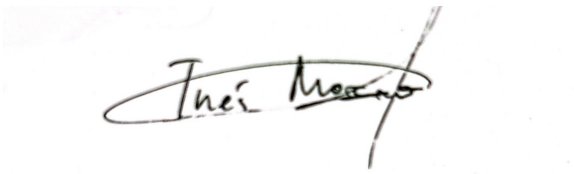
**ADVERTENCIA:** Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Inés Moreno Rabanal, estudiante de Doble Grado en ADE y Business Analytics (E-2 + Analytics) de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Interpretabilidad Global de Modelos de Machine Learning para la Gestión de Riesgos Financieros: Un Enfoque Agnóstico al Modelo", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
2. Interpretador de código: Para realizar análisis de datos preliminares.
3. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
4. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
5. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
6. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
7. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
8. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 19 de abril de 2024

Firma: 

# Referencias


Agarwal, G. (2022). Explainable ai (xai) - permutation feature importance. *Medium*. (Retrieved from `https://medium.com/@gauravagarwal_14599/explainable-ai-xai-permutation-feature-importance-7be0da6f0600`)

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., … Herrera, F. (2023, November). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805.

Allen, G. I., Gan, L., y Zheng, L. (2023). Interpretable machine learning for discovery: Statistical challenges\& opportunities. *arXiv preprint arXiv:2308.01475*.

Alonso, A., y Carbo, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost.

Alonso, A., y Carbó, J. M. (2022). Accuracy of explanations of machine learning models for credit decision.

Apley, D. W., y Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086.

Ayad, O. M., Hegazy, A.-E. F., y Dahroug, A. (2023). A proposed model for loan approval prediction using explainable artificial intelligence. En *2023 eleventh international conference on intelligent computing and information systems (icicis)* (pp. 166–173).

Bermúdez, L., Anaya, D., y Belles-Sampera, J. (2023). Explainable ai for paid-up risk management in life insurance products. *Finance Research Letters*, *57*, 104242.

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., … Wojewnik, P. (2021). Enabling machine learning algorithms for credit scoring–explainable artificial intelligence (xai) methods for clear understanding complex predictive models. *arXiv preprint arXiv:2104.06735*.

Bodensteiner, J. (2020, 5 de Octubre). Comparison of ale and pdp. En *Limitations of interpretable machine learning methods* (cap. 6).

Bodria, F., Giannotti, F., Guidotti, R., y et al. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, *37*, 1719–1778. Descargado de `https://doi.org/10.1007/s10618-023-00933-9`

doi: 10.1007/s10618-023-00933-9

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., y Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. En *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1721–1730).

Cascarino, G., Moscatelli, M., y Parlapiano, F. (2022). Explainable artificial intelligence: interpreting default forecasting models based on machine learning. *Bank of Italy Occasional Paper*(674).

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cirqueira, D., Nedbal, D., Helfert, M., y Bezbradica, M. (2020). Scenario-based requirements elicitation for user-centric explainable ai: A case in fraud detection. En *International cross-domain conference for machine learning and knowledge extraction* (pp. 321–341).

Commission, E. (2019). Ethics guidelines for trustworthy ai. *High-Level Expert Group on Artificial Intelligence*. Descargado de `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

Company, M. . (2023). The top trends in tech. *McKinsey Digital*. Descargado de `https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech`

Dankers, C. (2020). Pfi: Training vs. test data. En *Limitations of interpretable machine learning methods* (cap. 11).

Demajo, L. M., Vella, V., y Dingli, A. (2020). Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*.

Ferreira, L. (2023). *German credit data with risk.* Descargado de `https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk`

Fisher, A., Rudin, C., y Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, *20*(177), 1–81.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H., y Popescu, B. E. (2008). Predictive learning via rule ensembles.

Fritz-Morgenthal, S., Hein, B., y Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible ai. *Frontiers in artificial intelligence*, *5*, 779799.

Gartner, I. (2023). Gartner unveils top predictions for it organizations and users in 2024 and beyond. *Gartner Newsroom*. Descargado de `https://www.gartner.com/en/newsroom/press-releases/2023-10-17-gartner-unveils-top-predictions`

-for-it-organizations-and-users-in-2024-and-byond

Goldstein, A., Kapelner, A., Bleich, J., y Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Greenwell, B. M., Boehmke, B. C., y McCarthy, A. J. (2018). A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*.

Güntay, L., Bozan, E., Tığrak, Ü., Durdu, T., y Özkahya, G. E. (2022). An explainable credit scoring framework: A use case of addressing challenges in applied machine learning. En *2022 ieee technology and engineering management conference (temscon europe)* (pp. 222–227).

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., y Aggarwal, C. (2019). Efficient data representation by selecting prototypes with importance weights. En *2019 ieee international conference on data mining (icdm)* (pp. 260–269).

Hall, P., Ambati, S., y Phan, W. (2017). *Ideas on interpreting machine learning.* `https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/`.

Hashimoto, R., Miura, K., y Yoshizaki, Y. (2023). *Application of machine learning to a credit rating classification model: Techniques for improving the explainability of machine learning* (Inf. Téc.). Bank of Japan.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., … Hussain, A. (2023). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 1–30.

Hassija, V., Chamola, V., y Mahapatra, A. e. a. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, *16*, 45–74. Descargado de `https://doi.org/10.1007/s12559-023-10179-8` doi: 10.1007/s12559-023-10179-8

Hastie, T., Tibshirani, R., y Friedman, J. (2009). *The elements of statistical learning* (Second ed.). Springer.

Hooker, G. (2004). Discovering additive structure in black box functions. En *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 575–580).

Jones, H., y Chen, S. (2021, 03 de Agosto). Surrogate models: A comfortable middle ground? *Society of Actuaries*. (Additional article information: 6 minutos de lectura)

Karpagam, G., Varma, A., Samrddhi, M., y cols. (2022). Understanding, visualizing and explaining xai through case studies. En *2022 8th international conference on advanced computing and communication systems (icaccs)* (Vol. 1, pp. 647–654).

Khan, M. A., Khan, M. A., Khan, M. A., Khan, M. U., Khan, M. S., y Khan, M. F. (2022). The impact of artificial intelligence on the future of work: A review and research agenda. *Sensors*, *22*(8), 469.

Kim, B., Khanna, R., y Koyejo, O. O. (2016). Examples are not enough, learn to criticize!

criticism for interpretability. *Advances in neural information processing systems*, *29*.

Kłosok, M., Chlebus, M., y cols. (2020). *Towards better understanding of complex machine learning models using explainable artificial intelligence (xai): Case of credit scoring modelling*. University of Warsaw, Faculty of Economic Sciences Warsaw.

Kuo, C. (2022). *How is the partial dependent plot computed?* https://medium.com/dataman-in-ai/how-is-the-partial-dependent-plot-computed-8d2001a0e556. (Retrieved from https://medium.com/dataman-in-ai/how-is-the-partial-dependent-plot-computed-8d2001a0e556)

*Ley de inteligencia artificial de la unión europea.* (2024). https://artificialintelligenceact.eu/es/ai-act-explorer/. (Accedido: 2024-04-07)

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub.

Moreno, I. (2023). *Interpretabilidad de modelos de ml en riesgo de crédito.* https://github.com/inesmorenor/Interpretabilidad_Modelos_ML_Riesgo_de_Credito. (Accedido: 2023-04-17)

Okoli, C. (2024, 13 de February). Statistical inference using machine learning and classical techniques based on accumulated local effects (ale). *SKEMA Business School – Université Côte d'Azur, Paris*.

Rdusseeun, L., y Kaufman, P. (1987). Clustering by means of medoids. En *Proceedings of the statistical data analysis based on the l1 norm conference, neuchatel, switzerland* (Vol. 31).

Ronaghan, S. (2020, 5 de Noviembre). An introduction to protodash: An algorithm to better understand datasets and machine learning. *Towards Data Science*. Descargado de https://towardsdatascience.com/an-introduction-to-protodash-an-algorithm-to-better-understand-datasets-and-machine-learning-613c24b23719

Salter, R. I. (2023). *Explainable artificial intelligence and its applications in behavioural credit scoring.*

Shwartz-Ziv, R., y Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

Tejero, J. (2022). Unwrapping black box models a case study in credit risk. *Revista de Estabilidad Financiera*(Autumn).

Toloşi, L., y Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994.

Trusted AI. (2019). *Aix360: Ai explainability 360 toolkit by trusted ai.* Descargado de https://github.com/Trusted-AI/AIX360 (Accedido: 2023-04-08)

*XGBoost Working.* (2023, February). https://www.geeksforgeeks.org/xgboost/.

Yi, Z., Liang, Z., Xie, T., y Li, F. (2023). Financial risk prediction in supply chain finance

based on buyer transaction behavior. *Decision Support Systems*, *170*, 113964.