

Trabajo fin de grado

Interpretabilidad en modelos de clasificación



Gregorio Blázquez Martínez

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR



Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Interpretabilidad en modelos de clasificación

Autor: Gregorio Blázquez Martínez

Tutor: Ana María González Marcos

junio 2024

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

DERECHOS RESERVADOS

© 20 de junio de 2024 por UNIVERSIDAD AUTÓNOMA DE MADRID
Francisco Tomás y Valiente, nº 1
Madrid, 28049
Spain

Gregorio Blázquez Martínez
Interpretabilidad en modelos de clasificación

Gregorio Blázquez Martínez

*A mi hermano pequeño Jesús Cipriano, por
enseñarme que hay locos que programan en Neovim.*

*El arte de la ciencia es entender,
su fruto es comunicar ese entendimiento.*

Richard Feynman

PREFACIO

Este Trabajo Fin de Grado se presenta con un enfoque formal y práctico, adecuado para un TFG de informática. Mi intención ha sido tratar el tema de la interpretabilidad de los modelos de clasificación de manera rigurosa, definiendo las ecuaciones y funciones representativas de cada modelo y técnica, pero sin olvidarme de la aplicación práctica de estos conceptos.

He procurado que el lector pueda comprender tanto los aspectos teóricos como los resultados prácticos obtenidos, reconociendo la importancia de la transparencia y la confianza en los modelos de inteligencia artificial, especialmente en el ámbito económico. Debido a las limitaciones de extensión, he tratado los experimentos de manera directa y he reducido el estado del arte al mínimo necesario, incluyendo solo lo indispensable para contextualizar y apoyar los experimentos realizados.

Además, he incluido mucha información adicional en los anexos para aquellos lectores que deseen profundizar más en los detalles y aspectos complementarios de este trabajo.

Espero que este documento proporcione una visión clara y útil sobre la interpretabilidad de los modelos de clasificación y su relevancia en el desarrollo de modelos más transparentes y fiables.

Gregorio Blázquez Martínez

AGRADECIMIENTOS

Quiero expresar mis agradecimientos a mis profesores a lo largo de estos años académicos y en especial a Ana María González Marcos, por brindarme la oportunidad de realizar este TFG y por su orientación. También quiero agradecer a Damián Álvarez y Patrizio Guagliardo por su ayuda y enseñanzas en el campo de la ciencia de datos.

Fuera del ámbito académico, deseo agradecer profundamente a mi familia, amigos y, especialmente, a mis compañeros de carrera. Gracias por ser lo mejor que me llevo de estos cinco años.

RESUMEN

El presente Trabajo Fin de Grado se centra en la interpretabilidad de los modelos de clasificación, un área crucial en la inteligencia artificial y el aprendizaje automático debido a la creciente necesidad de comprender y confiar en las decisiones tomadas por estos modelos. En este documento se abordan varios métodos de clasificación, desde modelos interpretables como la regresión logística, hasta modelos considerados cajas negras como las redes neuronales artificiales, pasando por modelos intermedios como los métodos de ensamble. En la parte central de este TFG se estudian diversas técnicas de interpretabilidad, incluyendo Partial Dependence Plots (PDP), Acumulated Local Effects (ALE), Permutation Feature Importance (PFI), Local Interpretable Model-agnostic Explanations (LIME) y Shapley Additive Explanations (SHAP).

La investigación se desarrolla implementando y evaluando modelos de clasificación en diferentes conjuntos de datos, todos ellos del ámbito económico debido a la importancia de la interpretabilidad y la necesidad de confianza en este sector. En estos experimentos, se aplican las técnicas de interpretabilidad con el fin de mostrar su utilidad práctica, estudiar estas técnicas a fondo con sus limitaciones, identificar posibles sesgos y mejorar la transparencia de los modelos. En particular, se utiliza una base de datos artificial para estudiar la consistencia de PDP y ALE en clasificación; la base de datos de riesgo de créditos alemanes para comparar FI y PFI con random forest, analizar el sesgo de género utilizando SHAP y mostrar el uso de las técnicas de interpretabilidad estudiadas. También se emplea otra base de datos sobre la satisfacción de clientes del banco Santander, para evaluar la interpretabilidad y los límites de esta. Adicionalmente, se utilizó un clasificador de análisis de sentimiento de textos financieros para evaluar interpretaciones en texto y entender el funcionamiento del clasificador.

Los experimentos realizados demuestran la utilidad de las técnicas de interpretabilidad no solo para entender el comportamiento de los modelos, sino también para detectar y corregir problemas como el sobreajuste y los sesgos de género. A través de este trabajo, se concluye que la interpretabilidad es una herramienta esencial para el desarrollo de modelos de clasificación más transparentes y fiables, y se resalta la importancia de estudiar a fondo estos métodos de manera formal para aprovecharlos al máximo y seguir avanzando.

PALABRAS CLAVE

Interpretabilidad, Modelos de clasificación, PDP, ALE, PFI, LIME, SHAP, Sesgos

ABSTRACT

The present Bachelor's Thesis focuses on the interpretability of classification models, a crucial area in artificial intelligence and machine learning due to the growing need to understand and trust the decisions made by these models. This document addresses various classification methods, ranging from interpretable models like logistic regression to black-box models such as artificial neural networks, and intermediate models like ensemble methods. The core part of this thesis studies various interpretability techniques, including Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), Permutation Feature Importance (PFI), Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive Explanations (SHAP).

The research involves implementing and evaluating classification models on different datasets, all from the economic domain due to the importance of interpretability and the need for trust in this sector. In these experiments, interpretability techniques are applied to demonstrate their practical utility, thoroughly study these techniques with their limitations, identify potential biases, and improve model transparency. Specifically, an artificial dataset is used to study the consistency of PDP and ALE in classification; the German credit risk dataset is used to compare FI and PFI with random forest, analyze gender bias using SHAP, and demonstrate the use of the interpretability techniques studied. Additionally, a customer satisfaction dataset from Santander Bank are used to evaluate interpretability and its limits. Furthermore, a sentiment analysis classifier for financial texts is used to evaluate interpretations in text and understand the classifier's functionality.

The experiments conducted demonstrate the usefulness of interpretability techniques not only to understand model behavior but also to detect and correct issues such as overfitting and gender bias. This work concludes that interpretability is an essential tool for developing more transparent and reliable classification models, emphasizing the importance of thoroughly studying these methods formally to maximize their benefits and continue advancing.

KEYWORDS

Interpretability, Classification models, PDP, ALE, PFI, LIME, SHAP, Biases

ÍNDICE

1 Introducción	1
1.1 Contexto	1
1.2 Objetivos	2
1.3 Organización	2
2 Estado del arte	3
2.1 Modelos de clasificación	3
2.1.1 Regresión logística	4
2.1.2 Árboles de decisión	5
2.1.3 Métodos de ensamblaje de modelos	7
2.1.4 Redes neuronales artificiales	8
2.2 Métodos de interpretabilidad	10
2.2.1 Partial Dependence Plot	10
2.2.2 Acumulated Local Effects Plot	11
2.2.3 Permutation Feature Importance	13
2.2.4 Local Interpretable Model-agnostic Explanations	13
2.2.5 SHapley Additive Explanations (SHAP)	15
3 Desarrollo	17
3.1 Metodología	17
3.1.1 Preprocesamiento	17
3.1.2 Entrenamiento de modelos	18
3.1.3 Medidas de precisión y evaluación	18
3.2 AutoML	18
4 Experimentos y Resultados	19
4.1 Base de datos artificial	20
4.1.1 Consistencia de PDP y ALE	20
4.2 Base de datos de riesgo de créditos alemanes	22
4.2.1 Interpretabilidad de modelos	22
4.2.2 Robustez de FI y PFI con random forest	26
4.2.3 Análisis de sesgo en función del género utilizando SHAP	28
4.3 Base de datos de satisfacción de clientes del banco Santander	32
4.3.1 Límites de la interpretabilidad	32

4.4 Clasificador de análisis de sentimientos	33
4.4.1 Resultados	33
4.4.2 Interpretación de los resultados	34
5 Conclusiones y Trabajos futuros	35
5.1 Conclusiones	35
5.2 Trabajos futuros	36
Bibliografía	38
Apéndices	39
A Complementos State of Art	41
A.1 Modelos de clasificación	41
A.1.1 Regresión lineal	41
A.1.2 Regresión logística	43
A.1.3 Redes neuronales artificiales	44
A.2 Métodos de interpretabilidad	45
A.2.1 Partial Dependence Plot	45
A.2.2 Acumulated Local Effects Plot	45
A.2.3 Feature interaction	47
A.2.4 Global surrogate	49
A.2.5 Shapley values	50
B Complementos Experimentos	53
B.1 Datos tabulares	53
B.1.1 AutoML	53
B.1.2 Consistencia de PDP y ALE	55
B.1.3 Interpretabilidad de modelos en German Risk	58
B.1.4 Robustez de FI y PFI con random forest	60
B.1.5 Análisis de sesgo en función del género utilizando SHAP	61
B.1.6 Satisfacción de clientes del banco Santander	64
B.2 Texto	65
B.2.1 Clasificador de análisis de sentimientos	66

LISTAS

Lista de ecuaciones

2.1	Definición del modelo de regresión logística	4
2.2	Definición del modelo de árbol de decisión	5
2.3	Definición de neurona	9
2.4	Definición de una red neuronal	9
2.5	Definición de la función de dependencia parcial	10
2.6	Aproximación de la función de dependencia parcial	11
2.7	Definición de la función de efectos acumulados	11
2.8	Definición de la función de minimización de lime	14
A.1	Definición del modelo de regresión lineal	41
A.2	Definición log-verosimilitud	43
A.3	Estimador por Newton-Raphson	43
A.4	Importancia de características numéricas PDP	45
A.5	Importancia de características categóricas PDP	45
A.6	Estimación de ALE	46
A.7	Estimación de ALE centrada	47
A.8	Descomposición función de dependencia parcial simple	48
A.9	Estadístico iteración dos características	48
A.10	Descomposición función de dependencia parcial individual	48
A.11	Estadístico iteración una característica	48
A.12	Definición medida R-cuadrado	49
A.13	Función característica a partir de un modelo	51
A.14	Eficiencia de los valores Shapley	51
A.15	Definición de los valores Shapley	52
A.16	Estimación de los valores Shapley	52
B.1	Cálculo específico PDP	55
B.2	Cálculo específico ALE	56

Lista de figuras

2.1	Relación precisión-interpretabilidad	4
-----	--	---

2.2	Ejemplo necesidad árbol de decisión	5
2.3	Ejemplo funcionamiento de un random forest	8
2.4	Ejemplo funcionamiento de LIME	14
2.5	Estimación de KernelSHAP	16
4.1	Comparación de PDP teórico y real	20
4.2	Comparación de ALE teórico y real	21
4.3	PFI en German Risk	23
4.4	PDP en German Risk	23
4.5	ALE en German Risk	24
4.6	SHAP Summary en German Risk	24
4.7	SHAP Dependence en German Risk	25
4.8	Random forest Feature Importances (MDI)	27
4.9	Permutation Feature Importances (test set)	27
4.10	Permutation Feature Importances (train set)	28
4.11	Distribución por género y riesgo	29
4.12	Diferencia de paridad demográfica por modelo	30
4.13	Diferencia de paridad demográfica descompuesta por características	31
4.14	Comparación PFI banco Santander	32
4.15	Interpretación LIME para texto positivo	33
4.16	Interpretación SHAP para texto positivo	34
A.1	Ejemplo comparación PDP con ALE Plots	46
B.1	Comparación de PDP teórico y real en regresión	56
B.2	Comparación de ALE teórico y real en regresión	57
B.3	Comparación de PDP teórico y real para dos variables	58
B.4	Probabilidades del modelo de regresión logística	58
B.5	Probabilidades del modelo de random forest	59
B.6	Probabilidades del modelo de red neuronal	60
B.7	PFI en German Risk con el género	61
B.8	Diferencia de paridad demográfica descompuesta por características	62
B.9	Diferencia de paridad demográfica descompuesta por características	63
B.10	Importancia de las características por permutación - Regresión logística	64
B.11	Importancia de las características por permutación - Random forest	64
B.12	Importancia de las características por permutación - Red neuronal	65
B.13	Interpretación LIME previa a cambios	66
B.14	Interpretación LIME para texto neutral	66
B.15	Interpretación SHAP para texto neutral	67
B.16	Interpretación LIME para texto negativo	67

B.17 Interpretación SHAP para texto negativo	67
--	----

Lista de tablas

2.1 Fórmulas de impureza y ganancia	6
4.1 Resultados random forest	26
4.2 Tabla de créditos por género y riesgo	29
B.1 Resultados AutoML GermanRisk	53
B.2 Medidas de precisión de regresión logística en German Risk	54
B.3 Medidas de precisión de random forest en German Risk	54
B.4 Medidas de precisión de red neuronal en German Risk	54
B.5 Resultados AutoML Santander Customer Satisfaction	54
B.6 Medidas de precisión de regresión logística en Santander Customer Satisfaction	55
B.7 Medidas de precisión de random forest en Santander Customer Satisfaction	55
B.8 Medidas de precisión de red neuronal en Santander Customer Satisfaction	55
B.9 Medidas de precisión del modelo	57
B.10 Medidas de precisión de regresión logística	59
B.11 Medidas de precisión de random forest	59
B.12 Medidas de precisión de red neuronal	60
B.13 Importancia acumulada de las características	65
B.14 Características necesarias para diferentes umbrales en red neuronal	66
B.15 Conteo de 'Helsinki' por clase	68

INTRODUCCIÓN

La interpretabilidad de los modelos de clasificación es un tema crucial en la inteligencia artificial y el aprendizaje automático. En este Trabajo Fin de Grado, la intención es tratar este tema de manera formal y rigurosa, asegurando que los modelos y técnicas discutidos estén bien definidos. Debido a la vastedad del campo, no se abordan todos los modelos y métodos de interpretabilidad posibles; en su lugar, se enfoca en aquellos que permiten una discusión detallada y precisa.

1.1. Contexto

La interpretabilidad es fundamental porque nos permite entender y confiar en las decisiones tomadas por los modelos de clasificación. En ciertos campos, esto es tan importante que puede limitar el uso de modelos de clasificación. Por ejemplo, en el ámbito económico, donde la legislación actual exige la capacidad de dar explicaciones a los clientes, imponiendo la limitación de usar actualmente modelos interpretables con peores resultados. En este proyecto vamos a realizar la parte práctica en el sector económico por este motivo. Otro sector donde es necesario la interpretación es en la salud, donde la necesidad de confianza es totalmente imperativa para ciertas decisiones con consecuencias que afectan directamente a la calidad de vida de los pacientes.

Existen diferentes tipos de métodos de interpretabilidad, que pueden ser clasificados según su aplicabilidad local o global, y según si son agnósticos al modelo o específicos de un modelo en particular. Los métodos locales explican predicciones individuales, mientras que los métodos globales proporcionan una visión general del comportamiento del modelo. Además, los métodos agnósticos pueden ser aplicados a cualquier modelo, a diferencia de los métodos específicos que se diseñan para modelos concretos.

El resultado de un método de interpretación es proporcionar una comprensión más clara de cómo el modelo toma sus decisiones, lo que es vital para identificar y corregir posibles sesgos y mejorar la transparencia. Sin embargo, evaluar la interpretabilidad presenta desafíos significativos, ya que no existe una métrica única que mida de manera concluyente cuán interpretable es un modelo. La evaluación de la interpretabilidad a menudo depende del contexto y de la audiencia específica.

1.2. Objetivos

Los objetivos de este Trabajo Fin de Grado son los siguientes:

- **Mostrar el contexto actual y la relevancia de la interpretabilidad en modelos de clasificación.** Se pretende proporcionar un marco teórico sobre los modelos de clasificación y la importancia de su interpretabilidad, enfatizando su aplicación en el análisis de datos económicos y financieros.
- **Realizar un análisis detallado del estado del arte en técnicas de interpretabilidad.** Esto incluye una revisión exhaustiva de técnicas como PFI, LIME, SHAP, PDP y ALE, y su aplicación en diferentes contextos de clasificación de datos. Nos centramos únicamente en métodos agnósticos al modelo.
- **Demostrar la utilidad práctica de las técnicas de interpretabilidad.** Aplicar estas técnicas en conjuntos de datos reales y simulados, como la base de datos de riesgo crediticio alemán y análisis de sentimiento, para ilustrar cómo estas herramientas pueden mejorar la comprensión y la precisión de los modelos.
- **Evaluar la robustez y limitaciones de las técnicas de interpretabilidad.** Mediante experimentos detallados, se busca profundizar en el funcionamiento de estas técnicas, destacando su eficacia y las áreas donde presentan desafíos o limitaciones, como se evidencia en el uso de PDP y ALE o en la comparación entre FI y PFI.
- **Identificar y mitigar sesgos en los modelos de clasificación.** Utilizar técnicas como SHAP para detectar y corregir sesgos inherentes en los datos de entrenamiento, asegurando un uso más ético y responsable de los modelos de machine learning.

1.3. Organización

Este trabajo se puede dividir en las tres siguientes partes:

- **Estudio del estado del arte.** En esta sección se proporciona una base teórica fundamental sobre los modelos de clasificación y las técnicas de interpretabilidad utilizadas en el trabajo.
- **Desarrollo práctico.** La parte central del TFG, se explica el desarrollo y se presentan los resultados obtenidos al aplicar las técnicas de interpretabilidad en varias bases de datos, destacando su utilidad práctica y sus limitaciones.
- **Conclusiones y futuros trabajos.** En esta sección se resumen los resultados más importantes obtenidos y se discuten las principales conclusiones del trabajo. Finalmente, se proponen posibles futuras líneas de estudio para seguir avanzando en el desarrollo de modelos de clasificación más transparentes y fiables.

Es importante mencionar que, aunque el cuerpo principal del trabajo se ha mantenido conciso, se ha incluido información adicional en los anexos para aquellos lectores que deseen profundizar más en los detalles y aspectos complementarios del trabajo. Para una comprensión más profunda de las técnicas de interpretabilidad discutidas, se recomienda consultar el libro de Christoph Molnar, “Interpretable Machine Learning” [1], el cual ha sido una referencia clave para este TFG y proporciona una cobertura exhaustiva del tema.

ESTADO DEL ARTE

En este estudio del arte se va a formalizar todo lo que se usará en el desarrollo práctico. Sin duda, la referencia más importante es la del autor ya mencionado Molnar [1], pero también son referencias destacadas las siguientes: [2–5].

2.1. Modelos de clasificación

En el ámbito de la informática y la ciencia de datos, los modelos de clasificación juegan un papel crucial en la resolución de problemas donde el objetivo es asignar una etiqueta categórica a una instancia dada. Estos modelos se utilizan en una amplia variedad de aplicaciones, desde la detección de spam en correos electrónicos hasta la identificación de enfermedades en diagnósticos médicos.

Los modelos de clasificación pueden variar significativamente en términos de complejidad y capacidad de interpretación. En general, existe una relación inversa entre la complejidad del modelo y su interpretabilidad. Modelos más simples, como la regresión logística, ofrecen una comprensión clara de cómo se toman las decisiones. Sin embargo, a medida que avanzamos hacia modelos más complejos, como las redes neuronales profundas, la interpretabilidad disminuye, aunque a menudo se obtienen mejoras en la precisión y la capacidad de captura de patrones complejos en los datos. Aunque depende de muchos factores, es un estándar aceptado que en general tenemos la situación mostrada en la figura 2.1.

En este estado del arte, exploraremos una progresión de modelos de clasificación, comenzando con técnicas básicas y avanzando hacia métodos más sofisticados. Esta revisión abarcará desde la regresión logística y los árboles de decisión, hasta los métodos de ensamble de modelos y las redes neuronales artificiales. Destacaremos los aspectos relacionados con la interpretación que nos interesan para este TFG.

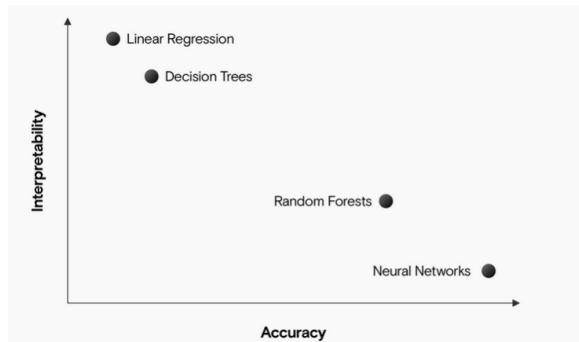


Figura 2.1: Relación precisión-interpretabilidad

2.1.1. Regresión logística

Debemos destacar la importancia de este modelo por ser altamente interpretable, lo cual hace que actualmente sea muy usado en sectores de los que ya hemos hablado donde la confianza en el modelo y las decisiones es crítica. Es un modelo muy estudiado y del que hay libros enteros hablando de los resultados formales que se conocen [6, 7]. El modelo de regresión logística se define de la siguiente manera:

Modelo de regresión logística

Supongamos un problema de clasificación binaria. La probabilidad de que la variable dependiente y sea igual a 1, dados los valores de las variables independientes x , se expresa mediante la siguiente fórmula:

$$\mathbb{P}(y = 1|x) = \frac{1}{1 + \exp(-\beta^T x)} \quad (2.1)$$

Aquí, β representa los coeficientes del modelo, y x es el vector de variables independientes, con un 1 en la primera coordenada seguido de las variables x_i . Esta formulación se adapta bien a la notación vectorial, permitiendo una representación más concisa.

Aunque no vamos a entrar en detalles, debemos comentar que la optimización de la regresión logística se basa en la maximización de la función de verosimilitud y los pesos se obtienen con el método de Newton-Raphson. Gracias a esto podemos afirmar que la regresión logística no es solo más interpretable que los modelos complejos, también es mucho más eficaz para converger rápidamente.

Interpretabilidad

En cuanto a la interpretabilidad, la regresión logística presenta ciertas complejidades debido a la naturaleza de la función logística. La interpretación de los pesos β_j no es lineal, sino aditiva debido a la multiplicación de factores exponenciales. La relación entre las probabilidades se expresa de la

siguiente manera:

$$\text{odds} = \frac{\mathbb{P}(y = 1)}{\mathbb{P}(y = 0)} = \exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)$$

Haciendo unos cálculos sencillos, interpretamos que el aumento de una unidad en x_j afecta la relación de probabilidades por un factor $\exp \beta_j$.

$$\frac{\text{odds}_{x_j+1}}{\text{odds}} = \exp(\beta_j \cdot (x_j + 1) - \beta_j \cdot x_j) = \exp \beta_j$$

Es decir, aumentar la variable x_j en una unidad incrementa la relación entre la probabilidad de que y sea 1 con la probabilidad de que y sea 0 en un factor $\exp \beta_j$.

2.1.2. Árboles de decisión

Los modelos de regresión lineal y logística fallan en situaciones donde la relación entre las características y el resultado es no lineal o donde las características interactúan entre sí. Aquí es donde los árboles de decisión destacan [8]. Estos modelos dividen los datos múltiples veces de acuerdo con ciertos valores de corte en las características. A través de estas divisiones, se crean subconjuntos diferentes del conjunto de datos, con cada instancia perteneciendo a un subconjunto. Los subconjuntos finales se llaman nodos terminales o nodos hoja, y los subconjuntos intermedios se llaman nodos internos o nodos de división. Los árboles se utilizan tanto para clasificación como para regresión.

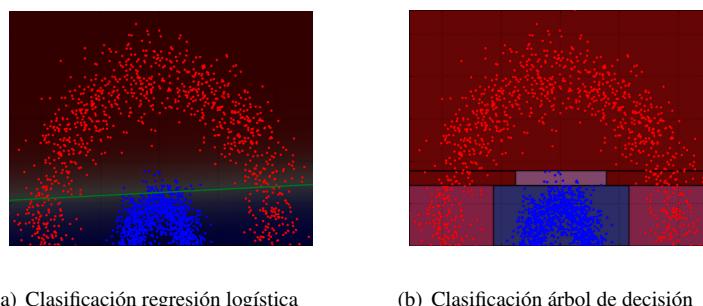


Figura 2.2: En esta figura se muestra la comparación entre la clasificación obtenida con regresión lineal 2.2(a) y la obtenida con 2.2(b) para el mismo problema.

Modelo de árbol de decisión

La relación entre el resultado \hat{y} y las características x se describe mediante la siguiente fórmula:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^M c_m \cdot \mathbb{I}(x \in R_m) \quad (2.2)$$

En un árbol de decisión, cada instancia del conjunto de datos cae exactamente en un nodo hoja (R_m). La función indicatriz $\mathbb{I}(x \in R_m)$ devuelve 1 si x está en el subconjunto R_m y 0 en caso contrario. Si una instancia cae en un nodo hoja R_l , el resultado predicho es $\hat{y} = c_l$, donde c_l es el promedio de todas las instancias de entrenamiento en el nodo hoja R_l .

Para construir el árbol de decisión, los subconjuntos de datos se generan minimizando la impureza en cada división. En tareas de regresión, se minimiza la varianza de y , mientras que en tareas de clasificación, se minimiza el índice de Gini o se maximiza la ganancia de información. Estas medidas de impureza y ganancia determinan la calidad de las divisiones y son fundamentales para el crecimiento efectivo del árbol.

En la tabla 2.1 se presentan las fórmulas utilizadas para calcular la impureza y la ganancia de información en diferentes contextos:

Impureza o Ganancia	Fórmula
Impureza de Gini	$I_G(f) = 1 - \sum_{i=1}^m f_i^2$
Ganancia de Información	$I_E(f) = -\sum_{i=1}^m f_i \log_2(f_i)$
Varianza	$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Tabla 2.1: Fórmulas de impureza y ganancia. f_i es la frecuencia del label i , m es el número de labels y n es el número de características.

Interpretabilidad

La interpretación de un árbol de decisión es sencilla: comenzando desde el nodo raíz, se avanza a los nodos siguientes y las aristas indican a qué subconjuntos o condiciones se está observando hasta llegar a un nodo hoja. A la hora de interpretar la decisión final, se juntan todas las condiciones o aristas que llevan a esa decisión unidas con una conjunción.

Importancia de las características

La importancia general de una característica en un árbol de decisión se puede calcular de la siguiente manera: se recorren todas las divisiones en las que se utilizó la característica y se mide cuánto ha reducido la varianza o el índice de Gini en comparación con el nodo padre.

Predicciones individuales

Las predicciones individuales de un árbol de decisión se pueden explicar siguiendo el camino de decisión a través del árbol y acumulando las contribuciones de cada división.

La raíz de un árbol de decisión es nuestro punto de partida. Si usáramos la raíz para hacer predicciones, predeciría la media del resultado de los datos de entrenamiento \bar{y} . Con la siguiente división,

restamos o sumamos según el siguiente nodo en el camino. Repetimos hasta el nodo terminal y obtenemos:

$$f(x) = \bar{y} + \sum_{d=1}^D \text{split.contrib}(d, x) = \bar{y} + \sum_{j=1}^p \text{feat.contrib}(j, x)$$

La predicción de una instancia individual es la media de los datos más la suma de todas las contribuciones de las D divisiones que ocurren entre el nodo raíz y el nodo terminal donde termina la instancia. Podemos observar las contribuciones de las características, las cuales pueden usarse en más de una división o no usarse en absoluto. Podemos sumar las contribuciones para cada una de las p características y obtener una interpretación de cuánto ha contribuido cada característica a una predicción.

2.1.3. Métodos de ensamblaje de modelos

Pegamos un salto en la interpretabilidad pasando a estudiar modelos cuya interpretabilidad es bastante más compleja que los anteriores pero que aún podemos considerar parcialmente interpretable.

Conceptos básicos

Ensamblajes y árboles de decisión

Los modelos de conjunto o ensamblaje [9], como Random Forest [10], Gradient Boosting, y AdaBoost, trabajan mediante la creación de un conjunto (ensamblaje) de modelos base, que generalmente son árboles de decisión. Cada modelo base se construye utilizando un subconjunto aleatorio de características y datos de entrenamiento. La predicción final se obtiene promediando (en el caso de regresión) o votando (en clasificación) las predicciones de cada modelo base.

Bagging y Random Feature Selection

El proceso de construir modelos base utilizando diferentes subconjuntos de datos se conoce como Bootstrap Aggregating (Bagging). Además, los modelos de conjunto introducen aleatoriedad al seleccionar un conjunto aleatorio de características en cada modelo base, lo que mejora la diversidad y robustez del modelo.

Podemos ver estos dos conceptos resumidos en la figura 2.3.

Interpretabilidad

A continuación, se destacan aspectos clave relacionados con la interpretabilidad de los modelos de conjunto:

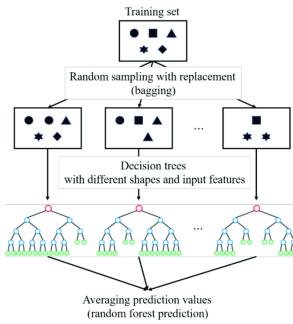


Figura 2.3: Ejemplo funcionamiento de un random forest

- **Modelo de Conjunto:** Los modelos de conjunto consisten en un conjunto de modelos base, cada uno contribuyendo a la decisión final del modelo. La interpretación directa de un solo modelo base puede ser más sencilla, pero comprender la combinación de múltiples modelos puede volverse más complejo.
- **Visualización de Modelos Base Individuales:** Aunque el conjunto completo puede ser complejo, es posible visualizar modelos base individuales dentro del modelo de conjunto. La visualización de un modelo base específico puede proporcionar una comprensión detallada de cómo se toman decisiones para conjuntos particulares de datos.
- **Feature Importance Plot:** Los modelos de conjunto extienden el concepto de importancia de las características de un solo modelo base. Calcular la importancia de las características en cada modelo base y promediar esos valores puede ofrecer una visión general de las variables más influyentes en el modelo de conjunto. Esto facilita la identificación de las características que más contribuyen a las decisiones del modelo.
- **Partial Dependence Plots:** Estos gráficos muestran cómo cambia la predicción del modelo cuando se varía una característica mientras se mantienen otras constantes. Proporcionan una comprensión más profunda de la relación entre una característica específica y la predicción del modelo.
- **Limitaciones:** Dada la complejidad del modelo de conjunto, puede ser desafiante explicar decisiones para casos individuales. Las interacciones no lineales entre características y las decisiones basadas en múltiples modelos base pueden dificultar una interpretación detallada.

A pesar de los desafíos, los modelos de conjunto siguen siendo valiosos en situaciones donde la interpretación precisa de cada predicción individual no es tan importante como la precisión general del modelo. En casos donde se requiere una interpretación más detallada, pueden considerarse técnicas específicas de interpretabilidad de modelos como las que veremos más adelante.

2.1.4. Redes neuronales artificiales

Pasamos de un modelo cuya interpretabilidad es más compleja a un modelo no interpretable. Las redes neuronales artificiales [11] están inspirados en la estructura y funcionamiento del sistema nervioso biológico. Están compuestas por unidades llamadas neuronas, organizadas en capas, y se utilizan para realizar tareas de aprendizaje automático, como clasificación y regresión.

Modelo de red neuronal

Una neurona en una red neuronal realiza una combinación lineal de sus entradas ponderadas por pesos, seguida de la aplicación de una función de activación. Formalmente, para una neurona j en una capa:

$$z_j = \sum_{i=1}^n w_{ij}x_i + b_j \quad (2.3)$$

$$a_j = f(z_j)$$

Aquí, w_{ij} son los pesos, x_i son las entradas, b_j es el sesgo (bias), $f(\cdot)$ es la función de activación, z_j es la entrada ponderada y a_j es la salida activada.

Una red neuronal se organiza en capas: capa de entrada, capas ocultas y capa de salida. La salida de una capa se convierte en la entrada de la siguiente. Para una red de L capas:

$$a_j^{(l)} = f^{(l)} \left(\sum_{i=1}^{n^{(l-1)}} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right) \quad (2.4)$$

donde l denota la capa, $n^{(l)}$ es el número de neuronas en la capa l , $w_{ij}^{(l)}$ son los pesos y $b_j^{(l)}$ son los sesgos.

Interpretabilidad de las redes neuronales

A pesar de su éxito en diversas tareas, las redes neuronales a menudo carecen de interpretabilidad directa. A diferencia de modelos más simples como la regresión lineal o logística, que proporcionan coeficientes directos para interpretar la influencia de cada característica, las redes neuronales presentan una complejidad intrínseca.

La falta de interpretabilidad en las redes neuronales se debe a su naturaleza no lineal y a la interacción no trivial entre las neuronas. A medida que aumenta la profundidad y complejidad de la red, comprender cómo cada característica contribuye a las decisiones de la red se vuelve más desafiante.

En contraste con modelos como Random Forest, donde es posible examinar la importancia de las características, las redes neuronales tienden a actuar como “cajas negras” o *modelos agnósticos*. No podemos atribuir fácilmente la predicción de la red a una característica específica.

Es crucial destacar que esta falta de interpretabilidad no invalida la utilidad de las redes neuronales. A menudo, su poder predictivo supera la necesidad inmediata de interpretación directa. Sin embargo, en contextos donde la interpretabilidad es esencial, se están desarrollando y utilizando técnicas específicas para mejorar la comprensión de las decisiones tomadas por las redes neuronales.

En la siguiente sección, exploraremos diversas técnicas de interpretabilidad de modelos que pueden aplicarse a redes neuronales para obtener una comprensión más profunda de su funcionamiento interno. Existen métodos específicos de interpretabilidad para redes neuronales, pero como no es el objetivo del TFG, nos centraremos solo en métodos agnósticos al modelo.

2.2. Métodos de interpretabilidad

Con el incremento en la complejidad de los modelos de clasificación, surge la necesidad de usar métodos que permitan interpretar las predicciones. La interpretabilidad de los modelos es esencial para garantizar la confianza en las decisiones que se toman, especialmente en aplicaciones críticas como la medicina, la justicia y las finanzas.

Los métodos agnósticos al modelo son técnicas de interpretabilidad que no dependen del tipo de modelo subyacente. Estos métodos proporcionan herramientas para entender y explicar las predicciones de cualquier modelo, ya sea simple o complejo. Como ya hemos comentado, son estos métodos los que nos interesan para este TFG.

En esta sección, revisaremos varios métodos de interpretabilidad agnósticos al modelo. Comenzaremos con el Partial Dependence Plot (PDP), que muestra la relación entre las características y la predicción de manera global, y avanzaremos hacia métodos más sofisticados como los valores Shapley y SHAP, que proporcionan explicaciones tanto globales como locales de las predicciones del modelo. Estos métodos nos permiten abrir la “caja negra” de los modelos complejos, proporcionando una comprensión más profunda y facilitando la toma de decisiones informadas.

Esta parte puede ser la más complicada de entender, pero veremos cómo se aplican estos métodos en el capítulo de resultados 4

2.2.1. Partial Dependence Plot

Partial Dependence Plot (PDP) muestra el efecto marginal que una o más características tienen en el resultado predicho de un modelo [12]. Puede indicar si la relación entre la predicción y una característica es lineal, monótona o más compleja.

En el caso de clasificación, donde el modelo de aprendizaje automático produce probabilidades, el PDP muestra la probabilidad para una cierta clase dada diferentes valores para las características en S . Para lidiar con múltiples clases, se puede dibujar una línea o gráfico por clase.

La función de dependencia parcial se define como:

$$\hat{f}_{S;PDP}(x_s) = E_{X_C} \left[\hat{f}(x_S, X_C) \right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C) \quad (2.5)$$

Donde S son las características para las cuales se debe trazar la función de dependencia parcial y C son las otras características utilizadas en el modelo de aprendizaje automático \hat{f} . Por lo general, hay una o dos características en el conjunto S para poder visualizarlo gráficamente.

Se suele estimar usando el método de Monte Carlo de la siguiente manera:

$$\hat{f}_{S;PDP}(x_s) = \frac{1}{k} \sum_{i=1}^k \hat{f}(x_S, x_C^{(i)}) \quad (2.6)$$

La función parcial nos dice, para un valor dado de las características en S , cuál es el efecto marginal promedio en la predicción. La suposición de los PDP es que las características en C no están correlacionadas con las características en S .

Ventajas y Desventajas

- **Ventajas:**

- La computación de los PDP es intuitiva: la función de dependencia parcial en un valor particular de la característica representa la predicción promedio si obligamos a todos los puntos de datos a asumir ese valor de característica.
- Si la característica para la cual se calculó el PDP no está correlacionada con las otras características, los PDP representan perfectamente cómo influye la característica en la predicción en promedio.
- Los PDP son fáciles de implementar.
- La interpretación de los PDP tiene una interpretación causal: intervenimos en una característica y medimos los cambios en las predicciones.

- **Desventajas:**

- El número máximo realista de características en una función de dependencia parcial es dos debido a la representación gráfica.
- La suposición de independencia es el mayor problema con los PD plots. Se asume que las características para las cuales se calcula la dependencia parcial no están correlacionadas con otras características.
- Los efectos heterogéneos pueden estar ocultos porque los PD plots solo muestran los efectos marginales promedio.

2.2.2. Acumulated Local Effects Plot

Los Gráficos de Efectos Locales Acumulados (ALE) muestran cómo las características influyen en la predicción de un modelo de aprendizaje automático en promedio [13]. Estos gráficos son una alternativa más rápida e imparcial a las Gráficas de Dependencia Parcial.

Los gráficos ALE promedian los cambios en las predicciones en un vecindario y los acumulan.

$$\begin{aligned}\hat{f}_{S;\text{ALE}}(x_S) &= \int_{z_{0,1}}^{x_S} \mathbb{E}_{X_C|X_S} \left[\hat{f}^S(X_S, X_C) \mid X_S = z_S \right] dz_S - \text{constante} \\ &= \int_{z_{0,1}}^{x_S} \int_{x_C} \hat{f}^S(z_S, x_C) \mathbb{P}(x_C \mid z_S) dx_C dz_S - \text{constante}\end{aligned}\tag{2.7}$$

La fórmula revela tres diferencias con respecto a los PDP:

- Promediamos los cambios en las predicciones, no las predicciones en sí. La derivada (o diferencia de intervalo) aísla el efecto de la característica de interés y bloquea el efecto de las características correlacionadas.

$$\hat{f}^S(x_S, x_C) = \frac{\partial \hat{f}(x_S, x_C)}{\partial x_S} \approx \frac{\hat{f}(x_S + \Delta, x_C) - \hat{f}(x_S, x_C)}{\Delta}$$

- Para compensar la derivación que hemos añadido, necesitamos una integral adicional sobre z . Es decir, acumulamos los gradientes locales sobre el rango de características en el conjunto S , lo que nos da el efecto de la característica en la predicción.
- Restamos una constante de los resultados para centrar el gráfico ALE y que el efecto promedio sobre los datos sea cero.

El valor del ALE se puede interpretar como el efecto principal de la característica en un cierto valor en comparación con la predicción promedio de los datos. Por ejemplo, una estimación de ALE de -2 en $x_j = 3$ significa que cuando la característica j -ésima tiene el valor 3, entonces la predicción es menor en 2 unidades en comparación con la predicción promedio.

Aunque no lo he tratado en el TFG por su complejidad, ALE también puede usarse con dos características e incluso con características categóricas si se puede establecer un orden entre ellas.

Ventajas y desventajas

Ventajas

- **No sesgados:** Funcionan bien con características correlacionadas.
- **Eficiencia computacional:** Más rápidos que los gráficos de dependencia parcial y escalan eficientemente.
- **Interpretación clara:** Interpretación centrada en cero para comprender fácilmente el efecto relativo de cambiar una característica en la predicción.

Desventajas

- **Inestabilidad con muchos intervalos:** Pueden volverse inestables con un alto número de intervalos.
- **Complejidad de la implementación:** Implementación más compleja en comparación con los gráficos de dependencia parcial.
- **Preferencia en situaciones específicas:** En situaciones con características descorrelacionadas, los gráficos de dependencia parcial pueden ser ligeramente preferibles.

2.2.3. Permutation Feature Importance

El concepto es sencillo pero extremadamente útil: medimos la importancia de una característica calculando el aumento en el error de predicción del modelo después de permutar la característica. Una característica es considerada importante si al permutar sus valores se incrementa el error del modelo, porque en este caso el modelo dependía de la característica para la predicción [14].

Algoritmo de importancia de características por permutación

- 1.– Estimar el error original del modelo $e_{\text{original}} = L(y, f(X))$ (por ejemplo, error cuadrático medio).
- 2.– Para cada característica $j = 1, \dots, p$:
 - Generar la matriz de características $X_{\text{permutado}}$ permutando la característica j en los datos X . Esto rompe la asociación entre la característica j y el resultado real y .
 - Estimar el error $e_{\text{permutado}} = L(Y, f(X_{\text{permutado}}))$ basado en las predicciones de los datos permutados.
 - Calcular la importancia de la característica mediante permutación $FI^j = e_{\text{permutado}}/e_{\text{original}}$. Alternativamente, se puede usar la diferencia: $FI^j = e_{\text{permutado}} - e_{\text{original}}$.
- 3.– Podemos ordenar las características por importancia de permutación descendente.

Ventajas y desventajas

Ventajas

- Da una visión global comprimida del comportamiento del modelo.
- Las medidas son comparables entre diferentes problemas y modelos.
- Considera todas las interacciones con otras características.
- No requiere volver a entrenar el modelo.

Desventajas

- No está claro si usar datos de entrenamiento o prueba para calcular la importancia.
- Está vinculada al error del modelo, lo que puede no ser apropiado en todos los casos.
- Requiere acceso al resultado real, para calcular el error.
- Funciona peor con características correlacionadas.

2.2.4. Local Interpretable Model-agnostic Explanations

Los modelos de sustitución local son modelos interpretables utilizados para explicar predicciones individuales de modelos de aprendizaje automático de caja negra. *Local Interpretable Model-agnostic Explanations* (LIME) fue presentado en el siguiente artículo [15] en el cual los autores proponen una implementación concreta de modelos de sustitución locales. Otras referencias interesantes son [16,17].

LIME utiliza modelos interpretables locales para aproximarse al comportamiento de un modelo

complejo alrededor de una instancia de interés como se muestra en la figura 2.4.

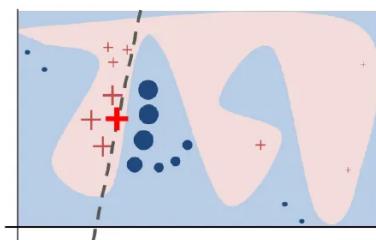


Figura 2.4: Ejemplo funcionamiento de LIME

El proceso simplificado de LIME es el siguiente:

- 1.– Se elige un punto de datos de interés x_0 .
- 2.– Se generan muestras perturbadas alrededor de x_0 .
- 3.– Se calculan las predicciones del modelo complejo en las muestras perturbadas.
- 4.– Se asignan pesos a las muestras perturbadas según su similitud con x_0 .
- 5.– Se ajusta un modelo explicativo local, como un modelo lineal, utilizando las muestras y sus pesos.
- 6.– Se utiliza el modelo local para explicar el comportamiento del modelo complejo en x_0 .

El objetivo de LIME es encontrar un modelo interpretable g que sea una buena aproximación de f en un vecindario local de x_0 . Formalmente, LIME busca resolver el siguiente problema de optimización:

$$\min_g L(f, g, \pi_x) + \Omega(g) \quad (2.8)$$

donde:

- D es el conjunto de datos generado alrededor de x_0 .
- $\pi_x(x_i)$ asigna pesos a puntos en D según su similitud con x .
- $\Omega(g)$ evalúa la complejidad de g , por ejemplo la profundidad en un árbol de decisión.
- $L(f, g, \pi_x) = \sum_i \pi_{x_i} \cdot l(f(x_i), g(x_i))$ mide la diferencia entre las predicciones de f y g en D .
- l es una función de pérdida para un punto particular.

Ventajas y desventajas

Ventajas

- Flexibilidad: Al igual que con los modelos de sustitución globales, podemos reemplazar el modelo de caja negra y utilizar el modelo local interpretable que prefiramos.
- Diversos tipos de datos: LIME es efectivo para datos tabulares, texto e imágenes, lo que amplía su utilidad.
- Medida de precisión: Como hacemos con los modelos globales tenemos medidas de precisión para hacernos a la idea de como de bien aproxima el modelo local al modelo de caja negra.
- Uso de características diferentes: Las explicaciones pueden basarse en características diferentes a las del modelo original, pudiendo facilitar las explicaciones, especialmente cuando las características originales son difíciles

de interpretar.

Desventajas

- Definición del vecindario: La definición correcta del vecindario es un desafío sin resolver, requiriendo ajustes para cada aplicación.
- Muestreo mejorable: LIME utiliza un muestreo de puntos de datos de una distribución gaussiana, ignorando la correlación entre características, lo que puede conducir a explicaciones basadas en datos improbables.
- Complejidad del modelo de explicación: La complejidad del modelo de explicación debe definirse de antemano, lo cual puede ser una limitación.
- Inestabilidad de las explicaciones: Las explicaciones pueden variar significativamente para puntos cercanos y entre repeticiones del proceso de muestreo, lo que dificulta la confianza en las explicaciones.

2.2.5. SHapley Additive Explanations (SHAP)

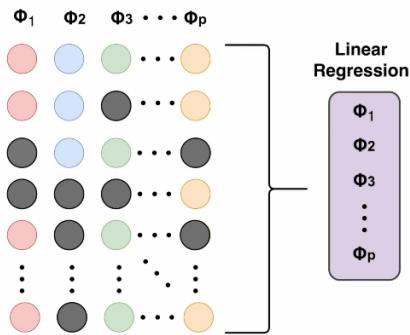
Antes de adentrarnos en SHAP, es importante mencionar que los valores Shapley, desarrollados por Lloyd Shapley [18], son la base teórica de SHAP. Aunque comprender la matemática detrás de los valores Shapley no es necesario para usar SHAP, he incluido un anexo tratando este tema y la intuición que hay detrás para aquellos interesados en la teoría que sustenta esta herramienta de interpretabilidad. También se recomiendan las siguientes referencias si se quiere profundizar más [19–23].

SHapley Additive Explanations, publicado en el siguiente artículo [24], es un método de explicación local que tiene como objetivo calcular la contribución de cada característica a una predicción específica, proporcionando así una explicación. SHAP se basa en los valores de Shapley y LIME, y puede crear visualizaciones globales al agregar las explicaciones locales. Este método tiene dos enfoques principales: KernelSHAP y TreeSHAP.

Por un lado, se fundamenta en los valores de Shapley, que representan cuánto contribuye cada variable a la predicción. Por otro lado, hereda de LIME la idea de explicar un modelo complejo a través de otro más simple, representando los valores de Shapley como una contribución aditiva de características, donde los coeficientes son los propios valores de Shapley.

KernelSHAP

KernelSHAP es un enfoque agnóstico al modelo, pero computacionalmente costoso. Crea una regresión lineal local con los valores de Shapley y estima las contribuciones de cada valor de característica a la predicción. El proceso implica la creación de coaliciones y la extracción de muestras de estas coaliciones para construir el modelo lineal ponderado con los valores Shapley como coeficientes. Sin embargo, es computacionalmente ineficiente y pone demasiado peso en casos improbables, sin considerar la correlación de los datos. En la figura 2.5 se muestra la idea detrás de la estimación de KernelSHAP.

**Figura 2.5:** Estimación de KernelSHAP

TreeSHAP

TreeSHAP, creado en 2018, es una variante específica para modelos basados en árboles de decisión. Es computacionalmente más eficiente y no se ve afectado por la correlación de variables. Aprovecha la estructura de los árboles individuales en los modelos de conjunto, como bosques aleatorios y XGBoost. A diferencia de KernelSHAP, TreeSHAP puede estimar los valores Shapley en tiempo lineal.

Ventajas y desventajas

Ventajas

- Distribución justa de la diferencia entre la predicción y la predicción promedio (propiedad de Eficiencia).
- Permite contrastar las explicaciones, comparando con subconjuntos o puntos de datos individuales en lugar de todo el conjunto.
- Base teórica sólida con propiedades como Eficiencia, Simetría, Dummy y Aditividad.

Desventajas

- Tiempo de computación elevado, generalmente se utiliza una solución aproximada en problemas del mundo real.
- No es adecuado para explicaciones dispersas (pocas características).
- Requiere acceso a los datos para calcular el valor Shapley de nuevas instancias.
- Vulnerable a instancias irreales en presencia de correlación entre características.

DESARROLLO

El desarrollo se llevó a cabo en Python utilizando varios Jupyter Notebooks para la experimentación y análisis, facilitando la documentación y visualización de los resultados a lo largo del proceso. Se emplearon las siguientes librerías para machine learning e interpretabilidad:

- **Scikit-learn (sklearn) [25]**: Para la optimización e inspección de los modelos, incluyendo la generación de Feature Importance, Permutation Feature Importance y Partial Dependence Plots.
- **ALEPython [26]**: Para generar los gráficos de ALE (Accumulated Local Effects).
- **LIME [27]**: Para interpretaciones locales de los modelos.
- **SHAP [28]**: Para interpretaciones detalladas del impacto de cada característica.
- **NLTK [29]**: Para el pipeline del análisis de texto.

3.1. Metodología

En general, previo a los experimentos y obtención de resultados que comentamos en el siguiente capítulo, para el desarrollo de este Trabajo de Fin de Grado (TFG) se ha seguido una metodología dividida en tres fases principales [30]. Realmente son seis fases, incluyendo una primera fase de estudio previo reflejada en parte en la sección 2, y una fase adicional utilizando AutoML para tener una referencia externa, como se menciona en el apartado 3.2.

3.1.1. Preprocesamiento

Se realizaron los siguientes pasos de preprocesamiento de datos para todas las bases de datos:

- **Manejo de valores faltantes**: Se imputaron los valores faltantes utilizando imputadores de k vecinos.
- **Normalización/estandarización de variables**: Las variables numéricas se normalizaron para tener una media de 0 y una desviación estándar de 1, haciendo que sean comparables en la misma escala.
- **Codificación de variables categóricas**: Las variables categóricas se codificaron utilizando One-Hot Encoding.
- **Eliminación de variables**: Mediante la visualización de las matrices de correlación de Pearson, correlation ratio y correlación de Cramer, se determinó la eliminación de variables demasiado correlacionadas entre ellas o que no aportaban información relevante para el modelo.

En el caso de la última base de datos que comentaremos también fue necesaria el balanceo de las clases, esto se hizo específicamente, como se comentará en la sección de experimentos, por estar recomendado por el proveedor de la base de datos.

3.1.2. Entrenamiento de modelos

Para esta fase, se seleccionaron tres tipos de modelos:

- **Regresión logística:** representa los modelos interpretables.
- **Random forest:** representa los modelos de ensemble.
- **Redes neuronales:** representa los modelos de caja negra.

No es el objetivo de este TFG conseguir los mejores modelos para nuestros datos, sino un medio para poder obtener el fin que queremos, estudiar la interpretabilidad. Para esta parte se ha tratado de estandarizar el proceso mediante el uso de la librería Scikit-learn. Comparando los resultados obtenidos tras la optimización de hiperparámetros [31] y umbrales, se verificó que esta simplificación no es excesiva al comparar los resultados con los obtenidos con AutoML.

3.1.3. Medidas de precisión y evaluación

Se emplearon las siguientes métricas para evaluar el rendimiento de los modelos [32]:

- **Exactitud (Accuracy):** Proporción de predicciones correctas sobre el total de casos.
- **Precisión (Precision):** Proporción de verdaderos positivos sobre el total de positivos predichos.
- **Recall (Sensibilidad):** Proporción de verdaderos positivos sobre el total de positivos reales.
- **F1-Score:** Media armónica entre la precisión y el recall, útil para balances entre ambas métricas.
- **AUC-ROC:** Área bajo la curva ROC, que mide la capacidad de un modelo para distinguir entre clases.

3.2. AutoML

Se utilizó AutoML [33] por dos motivos:

- Para reforzar el argumento planteado en la introducción, la necesidad de usar modelos más complejos y menos interpretables debido a la mejora que pueden suponer en las medidas de precisión y evaluación.
- Para tener una referencia externa que respalde que el preprocesamiento y entrenamiento de los modelos es suficientemente bueno.

En el anexo B.1.1 se muestran los resultados obtenidos con AutoML y con el desarrollo comentado en este capítulo para respaldar la correcta aplicación de la metodología comentada y para que puedan ser contrastados.

EXPERIMENTOS Y RESULTADOS

Como se ha comentado en la introducción, la parte práctica de este TFG pretende centrarse en el ámbito económico como el contexto donde reflejar la importancia de la interpretabilidad. Es por esto que, salvo la primera que sirve para una comprobación de la teoría de PDP y ALE, las bases de datos que se van a usar en este desarrollo son del ámbito económico. En particular, se han usado bases de datos etiquetadas obtenidas a través de Kaggle, y se va a proporcionar el acceso y una breve explicación de cada una de ellas con sus experimentos para dar claridad y replicabilidad.

Los experimentos se han diseñado para demostrar cómo diferentes técnicas de interpretabilidad pueden aplicarse en modelos de aprendizaje automático, haciendo énfasis en la comprensión de las decisiones que toman estos modelos. A continuación, se describen brevemente las bases de datos utilizadas y los objetivos de los experimentos:

- **Base de datos artificial:** Este primer experimento utiliza datos generados artificialmente para comprobar las teorías de PDP (Partial Dependence Plots) y ALE (Accumulated Local Effects). La simplicidad de esta base de datos no debe ser un problema pues este en este experimento vemos que es suficiente para verificar si las gráficas prácticas obtenidas coinciden con las fórmulas teóricas, llegando a mostrar ciertas limitaciones.
- **Base de datos de riesgo de crédito alemán:** Esta base de datos, que evalúa el riesgo crediticio de clientes de una institución bancaria alemana, sirve para aplicar y comparar distintas técnicas de interpretabilidad en modelos predictivos. Se busca no solo obtener modelos precisos, sino también comprender qué factores son más influyentes en las predicciones. Esta base de datos es la más explotada, pues no sólo sirve para mostrar la aplicabilidad de lo estudiado, también para realizar dos experimentos más específicos: una comparación entre FI y PFI, y un análisis de sesgo con SHAP.
- **Base de datos de Santander:** Esta base de datos se usa apenas para mostrar las limitaciones que no hemos tratado en este trabajo, y sería más interesante poder abordar este problema más en detalle.
- **Análisis sentimental:** En este caso, se emplea una base de datos sobre comentarios financieros. Los modelos de aprendizaje automático aplicados son interpretados para clasificar la connotación de estas entradas, lo que es especialmente útil en aplicaciones de procesamiento de lenguaje natural.

Cada uno de estos experimentos ha sido cuidadosamente seleccionado y diseñado para resaltar la importancia de la interpretabilidad en modelos de aprendizaje automático aplicados al ámbito económico. Se han utilizado técnicas avanzadas como PFI, PDP, ALE, LIME y SHAP, no solo para obtener resultados precisos sino para asegurar que estos resultados sean comprensibles y útiles para la toma de decisiones.

4.1. Base de datos artificial

Se generaron los datos x_1 y x_2 con una distribución uniforme entre 0 y 1, y se les asignaron clases en función de la probabilidad $f(x_1, x_2) = -2x_1^2 - 2x_2^2 + 2x_1 + 2x_2$ más un pequeño error siguiendo una distribución normal con media 0 y desviación estándar 0,1 (se ha calculado para diferentes errores, pero este es el más representativo), usando como umbral 0.5. De esta manera se pretende conseguir un problema casi ideal para comparar si en estas condiciones se cumple efectivamente que las gráficas de los PDP y ALE siguen las fórmulas descritas 2.5 y 2.7 respectivamente (los cálculos de las funciones teóricas se muestran en el anexo). Al generar los datos para este experimento, no ha sido necesario ni el preprocesamiento ni el uso de AutoML descritos en el procedimiento de desarrollo.

4.1.1. Consistencia de PDP y ALE

En este experimento, se ha entrenado una red neuronal con los datos descritos. En el anexo B.1.2 pueden comprobarse los cálculos de las funciones teóricas y que la red neuronal obtiene buenos resultados, nada sorprendente con un problema tan sencillo. Tras esto, se compararán los resultados prácticos de PDP y ALE, aplicados a este modelo, con el teórico esperado. Es importante aclarar que, aunque es una base de datos muy sencilla, es especialmente interesante por tratarse de un problema de clasificación y no regresión. Es decir, el modelo no tiene acceso a la función de probabilidad con la que se generan los datos y por tanto los resultados deberían ser peores que los que obtuviéramos si probáramos un problema de regresión con la variable objetivo la probabilidad directamente. Otro punto importante a destacar es que para ALE, debido al funcionamiento de la librería utilizada descrita en el capítulo de desarrollo 3, solo voy a mostrar las gráficas obtenidas y no voy a interpretar los resultados.

Resultados

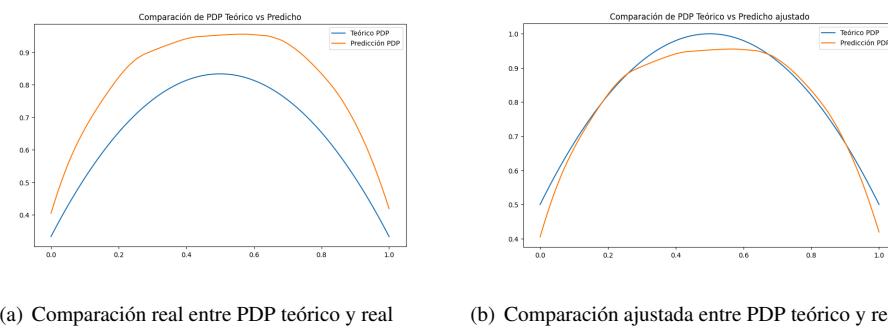


Figura 4.1: Comparación de la función teórica de PDP y la obtenida a partir del modelo. Para la gráfica sin ajustar tenemos un MSE de 0.0224 y para la ajustada un MSE de 0.001

Tal como se comentaba, los resultados no son tan buenos como en el caso de regresión. El objetivo es centrarse en los resultados para modelos de clasificación, pero en el anexo B.1.2 pueden ver por

curiosidad los resultados para el mismo problema pero con modelos de regresión, los cuales son muy buenos y sin el problema que veremos en clasificación. También se ha añadido la comparación de los resultados para dos variables en vez de una.

En la figura 4.1 se muestra la comparación entre el resultado teórico y el real obtenido. El resultado más importante que obtenemos, es la necesidad de ajustar la función teórica con una constante para comparar los resultados. El resultado obtenido para modelos de clasificación se asemeja mucho al efecto esperado pero difiere en una constante, a diferencia del caso de regresión. En la figura 4.2 se muestran la gráfica teórica y la obtenida por el modelo para ALE.

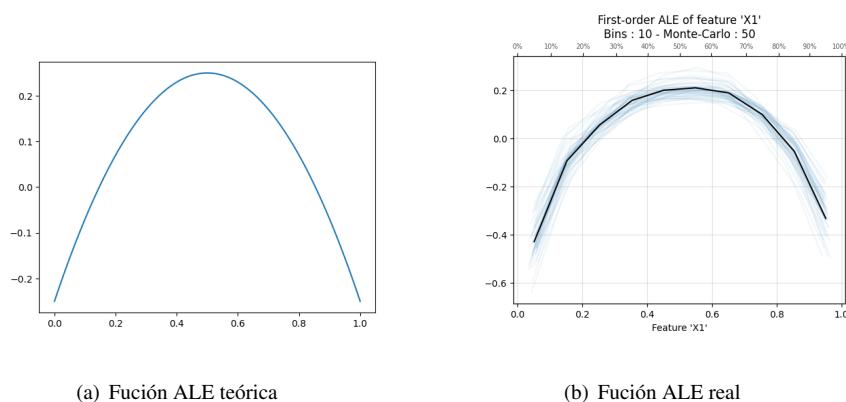


Figura 4.2: Comparación de ALE teórico y real

Interpretación de los resultados

Este experimento ha mostrado una clara falta de consistencia entre el resultado teórico, conocido en un problema tan simple, y el resultado real obtenido por el modelo. Esta diferencia cobra especial relevancia al tratarse de un problema tan básico, muy alejado de situaciones más realistas donde la relación entre variables es desconocida y la complejidad es considerablemente mayor. Es crucial destacar que, si bien el efecto general está reflejado en los resultados obtenidos, como se demuestra por la similitud entre las funciones teóricas y prácticas de PDP y ALE, el error en la constante sugiere una representación inexacta de la función real por parte del modelo. Esta discrepancia puede atribuirse a las particularidades de los problemas de clasificación, donde la naturaleza discreta de las clases introduce complejidades adicionales en la relación entre las características de entrada y la salida. Por ende, aunque los modelos puedan capturar parcialmente la relación subyacente, la falta de continuidad en la función objetivo puede dificultar la representación precisa de esta relación.

En conclusión, si bien los resultados ofrecen información valiosa sobre la capacidad del modelo para aprender y generalizar a partir de los datos de entrenamiento, es esencial reconocer las limitaciones y la falta de coherencia entre los resultados prácticos y teóricos. Este análisis puede servir para futuros trabajos y mejorar la comprensión de los modelos de aprendizaje automático en problemas de clasificación complejos.

4.2. Base de datos de riesgo de créditos alemanes

La base de datos German Risk es un conjunto de datos utilizado para evaluar el riesgo crediticio de clientes de una institución bancaria alemana. La base original contiene 1000 registros y 20 características que incluyen tanto variables numéricas como categóricas, para este TFG se usa una simplificación con las 10 características más importantes, disponible en este [enlace](#). El objetivo principal es clasificar si un cliente es un buen o mal riesgo crediticio (variable Risk) para tomar decisiones más informadas y solventar las implicaciones que puede tener dar malos créditos, por ejemplo impagos.

Las características de la base de datos son las siguientes.

- **Age:** Edad del cliente.
- **Sex:** Género del cliente (masculino o femenino).
- **Job:** Tipo de trabajo del cliente.
- **Housing:** Tipo de vivienda del cliente (propia, alquilada, gratis).
- **Saving accounts:** Cantidad de ahorro del cliente.
- **Checking account:** Cantidad en la cuenta corriente del cliente.
- **Credit amount:** Cantidad del crédito solicitado o monto.
- **Duration:** Duración del crédito en meses.
- **Purpose:** Propósito del crédito (por ejemplo, coche, electrodomésticos, etc.).
- **Risk:** Variable objetivo que indica si el cliente es un buen o mal riesgo crediticio.

Tras realizar el preprocesamiento ya comentado en la el apartado de la sección de desarrollo 3.1.1, se procedieron a realizar tres experimentos diferentes, cada uno con un objetivo diferente. Es importante destacar que se ha mantenido la característica del género pues vamos a tratar este sesgo y cómo la interpretabilidad puede ayudarnos a enfrentarlo en el experimento 4.2.3.

4.2.1. Interpretabilidad de modelos

Este experimento consiste en aplicar las diferentes técnicas de interpretabilidad estudiadas en este TFG para comprobar su utilidad y comparar los resultados en varios modelos (regresión logística, random forest y red neuronal). Vamos a mostrar estas técnicas de manera muy concisa con el objetivo de mostrar su utilidad, pero sin extendernos en los detalles.

En la figura 4.3 podemos ver el resultado de aplicar PFI a los tres modelos, esta técnica no da información muy específica pero es fundamental para ver qué importancia le están dando los modelos a cada característica y es un primer paso para entender el funcionamiento del modelo. Lo más destacable de estos resultados lo veremos más en detalle en el experimento 4.3, la capacidad de la red neuronal para usar más variables en comparación con los otros modelos.

El siguiente paso natural es intentar entender cómo afectan las variables más importantes, según

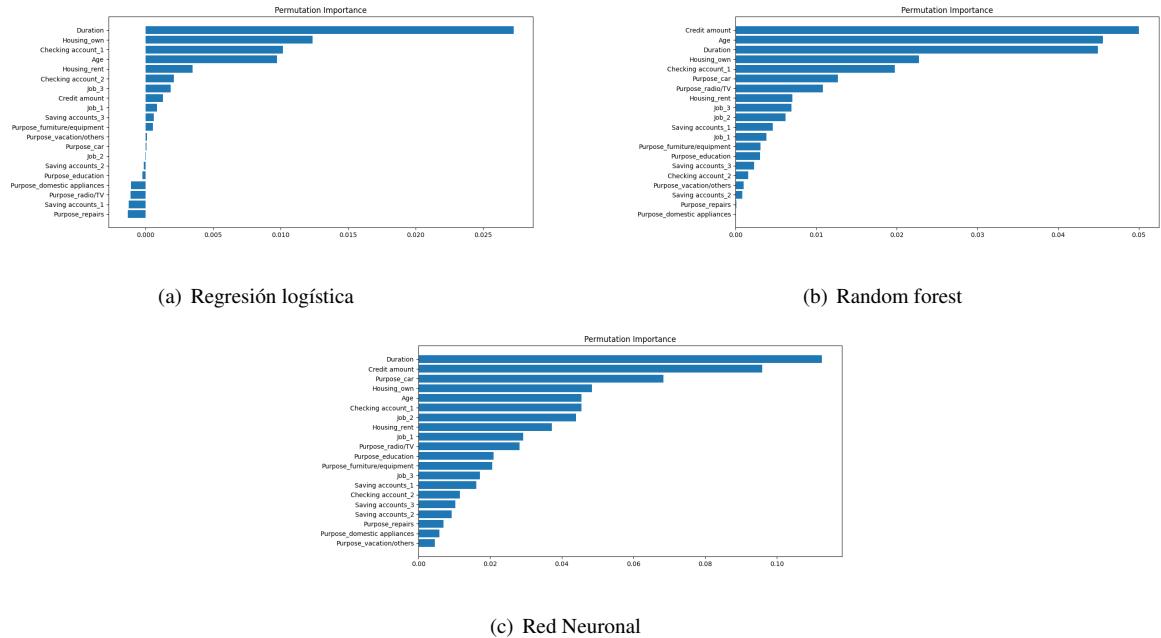


Figura 4.3: Permutación de la Importancia de las Características (PFI) para diferentes modelos.

lo visto anteriormente, nos interesamos por las características edad, cantidad prestada, duración, si está pagando un alquiler y si tiene muchos ahorros, mostradas en la figura 4.4. Dado que la regresión logística es interpretable, a partir de ahora vamos a aplicar estas técnicas sólo al modelo random forest y a la red neuronal.

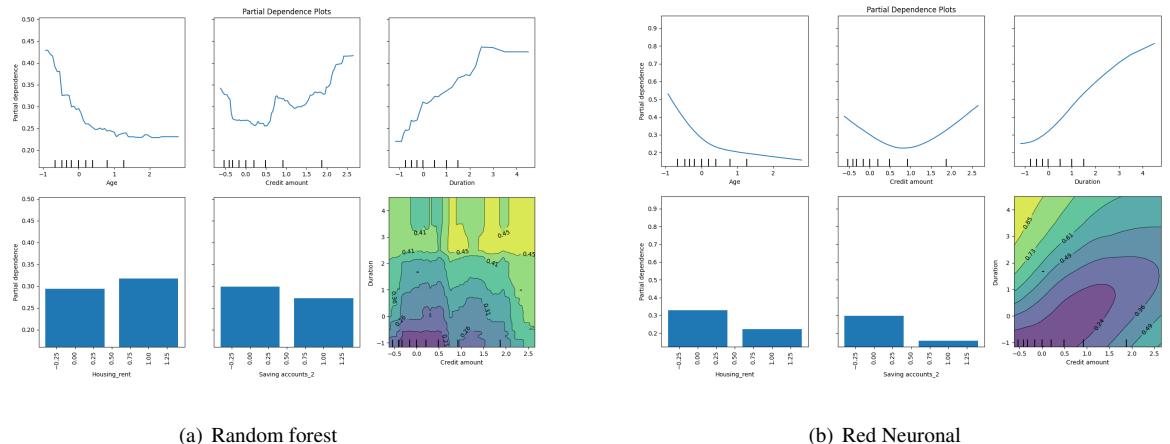


Figura 4.4: Partial Dependence Plot (PDP) de las características más importantes para diferentes modelos.

Vemos que los resultados para la red neuronal son más suaves que los obtenidos por random forest, esto se debe al funcionamiento interno de cada modelo, pero lo más interesante es que efectivamente hay similitudes entre ambos modelos, especialmente en cuanto a las variables edad, cantidad y duración del crédito. También destaca la diferencia en el mapa de calor de duración y crédito juntos.

Como alternativa a PDP se presenta la posibilidad de usar los ALE Plots, en la figura 4.5 vemos el resultado de estos gráficos para las tres variables más importantes. Es muy interesante observar la gran similitud que hay entre estos resultados independientemente de los modelos, mostrando ALE como una posibilidad que puede ser más consistente que PDP, habiendo suavizado los efectos del random forest al promediarlos y con una mayor coherencia al tomar intervalos para su cálculo como comentamos en el estudio del arte.

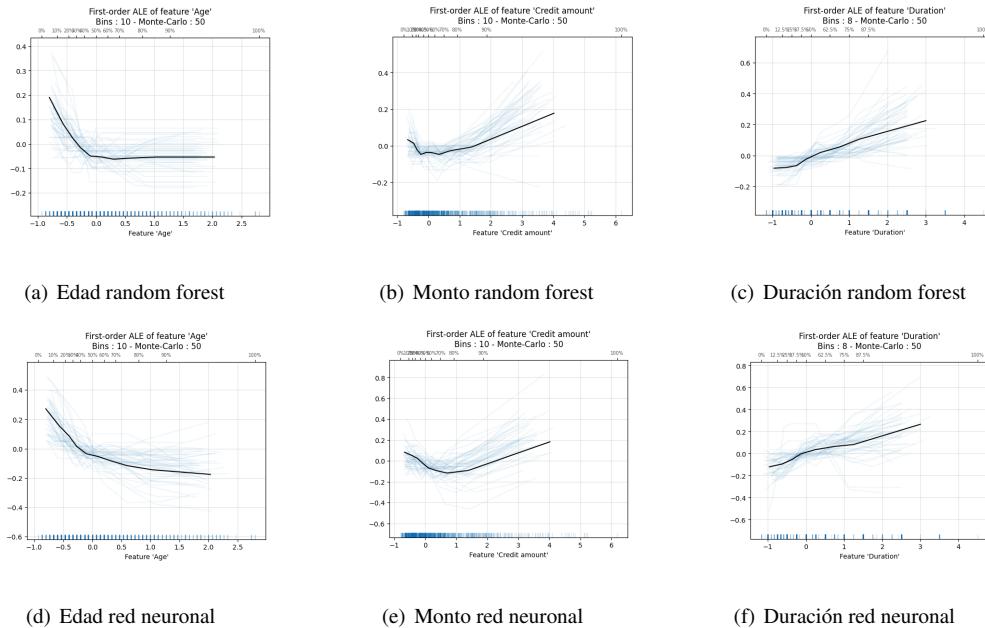


Figura 4.5: Gráficos de efectos acumulativos locales para las tres variables más importantes.

Finalmente quiero acabar destacando la enorme utilidad de SHAP, el cual no sólo resulta más interesante por toda la base matemática que tiene que lo hace un método más consistente y robusto, también por lo mucho que ha sido desarrollado estos años ofreciendo una gran cantidad de herramientas que lo convierten sin duda en la técnica de interpretabilidad más versátil estudiada en este TFG. En esta gráfica al valor SHAP se representa en el eje de abcisas y el valor de la variable con una escala de colores.



Figura 4.6: Summary Plot de SHAP de las características más importantes para diferentes modelos.

SHAP presenta muchas posibilidades, intentando unificar las diferentes técnicas de interpretabilidad más importantes actualmente. De esta manera SHAP nos ofrece posibilidades como los Bar Plots, una alternativa a PFI, aprovechando la consistencia que le otorga su base matemática. A continuación, se muestra en la figura 4.7 los gráficos de dependencia de SHAP que sustituyen a los PDP y ALE mostrados.

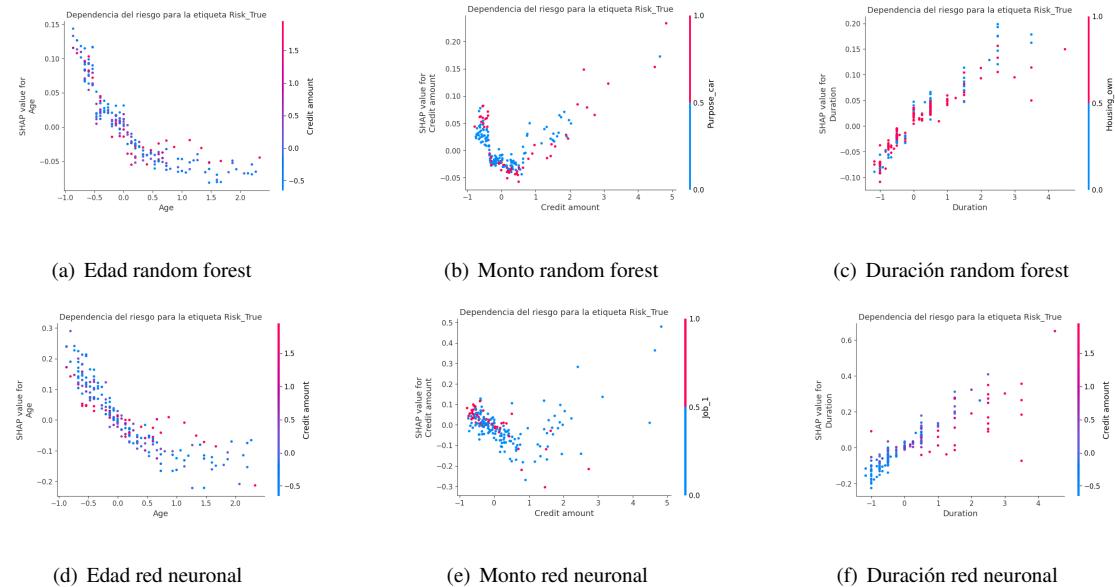


Figura 4.7: Dependence Plot de SHAP para las características más importantes para diferentes modelos.

En la figura 4.7 se representa en el eje de ordenadas los valores SHAP y en el eje de abcisas el valor de la variable a estudiar, igual que se hace en los PDP y ALE Plots. SHAP añade información extra, con una escala de colores muestra el valor de la variable que SHAP encuentra más relevante para entender la dependencia de la variable a estudiar.

Interpretación de los resultados

En este experimento hemos mostrado la aplicabilidad de las técnicas estudiadas con datos reales, y hemos comentado algunos detalles clave. Aunque hay mucho más que se podría comentar, se ha intentado ser lo más conciso posible para no extenderse demasiado y mostrar únicamente lo más relevante y sin entrar en detalles específicos sobre la base de datos. Como comentarios finales a resaltar, debemos destacar los siguientes:

- Hemos utilizado PFI como una primera herramienta para estimar qué variables son las que nos interesan estudiar. Además, es fácilmente comparable entre los distintos modelos.
- Para aquellos modelos no interpretables hemos utilizado los PDP para aproximar el efecto de las variables de manera individualizada, o incluso por pares. Hemos podido observar las diferencias que puede haber con esta técnica entre los diferentes modelos.
- Hemos presentado los ALE Plots como una técnica más consistente y una alternativa a los PDP. Es con es-

ta técnica donde hemos asemejado las interpretaciones para los distintos modelos, permitiendo que podamos comparar mejor los detalles que sí varían entre ellos.

- Finalmente hemos mostrado algunas de las aplicaciones de SHAP. Hay muchas más posibilidades que no hemos presentado, pero con lo poco mostrado hemos sido capaces de destacar la enorme utilidad y el gran potencial de esta herramienta. Mostraremos en el experimento 4.2.3 como SHAP puede aportar mucho más, en este caso para estudiar el sesgo de los modelos.

En resumen, cada técnica de interpretabilidad aporta una perspectiva única sobre el funcionamiento interno de los modelos. Mientras que los modelos de caja negra como las redes neuronales y los random forests suelen ser más precisos, técnicas como SHAP y ALE pueden ayudar a desentrañar su complejidad y hacerlos más transparentes.

4.2.2. Robustez de FI y PFI con random forest

En este experimento se compararán dos medidas de interpretabilidad: el Feature Importance (FI) específico de un random forest, basado en la impureza de los árboles de decisión descrita en 2.1.2, y la Permutation Feature Importance (PFI), una técnica agnóstica al modelo. El propósito de este experimento es exemplificar algunos problemas o limitaciones del FI y explorar el uso de PFI como una alternativa más robusta.

El experimento consiste en lo siguiente: se introdujeron dos variables aleatorias (una numérica y otra categórica) sin relación con la variable objetivo a la base de datos original del crédito alemán. Posteriormente, se entrenó un modelo de random forest sobre estos datos, obteniendo un modelo con la precisión descrita en la tabla 4.1. Es importante destacar que este modelo muestra signos de sobreajuste evidentes. A partir de este modelo, se calcularon los valores de FI y PFI, de los cuales se extraen ciertas conclusiones sobre la importancia de las variables en el modelo.

Métrica	Valor
Exactitud entrenamiento	1.000
Exactitud test	0.760
Precisión	0.809
Recall	0.866
F1 Score	0.837

Tabla 4.1: Resultados random forest

Resultados

La importancia de las características basada en la disminución media de la impureza (MDI) clasifica las características numéricas como las más importantes, en el anexo B.1.4 se detallan los motivos de así sea. Como resultado, en la figura 4.8 podemos observar que la variable *random_num* no predictiva

se clasifica como una de las características más importantes. La variable *random_cat* también adquiere importancia.

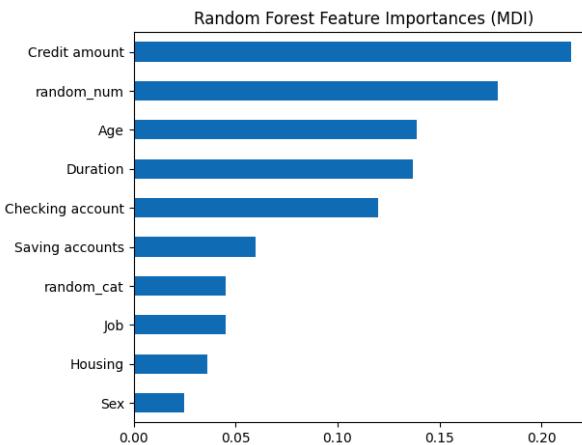


Figura 4.8: Random forest Feature Importances (MDI)

Este problema surge de dos limitaciones de las importancias de características basadas en la impureza:

- Las importancias basadas en la impureza están sesgadas hacia características de alta cardinalidad.
- Las importancias basadas en la impureza se calculan en estadísticas del conjunto de entrenamiento y, por lo tanto, no reflejan la capacidad de las características para ser útiles para hacer predicciones que se generalicen al conjunto de prueba.

Como alternativa, las importancias por permutación (PFI) se calculan en el conjunto de test. En la figura 4.9 podemos observar como estos valores son más coherentes que los comentados anteriormente. Las características aleatorias tienen importancias muy bajas, como se esperaba, y no tenemos el gran sesgo hacia las características de alta cardinalidad comentado (dos de las tres variables más importantes son categóricas con pocos posibles valores).

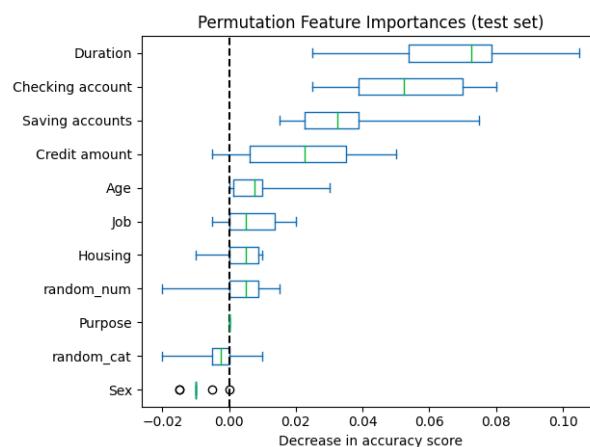


Figura 4.9: Permutation Feature Importances (test set)

Esto revela que *random_num* y *random_cat* obtienen un ranking de importancia significativamente más bajo. Un detalle muy importante que puede generar dudas, es que resultado obtendríamos si calculamos los valores de importancia por permutación en el conjunto de entrenamiento. Hemos obtenido resultados más coherentes, pero hemos cambiado tanto el método como el conjunto donde se calculan los valores. En la figura 4.10 se muestran estos valores calculados sobre el conjunto de entrenamiento, mostrando que claramente tenemos un modelo sobre ajustado al ser relevante para predecir las dos variables aleatorias que hemos introducido sin relación con la variable objetivo. La importancia de las dos variables aleatorias sigue siendo menor que la obtenida por FI, pero tenemos un error considerable.

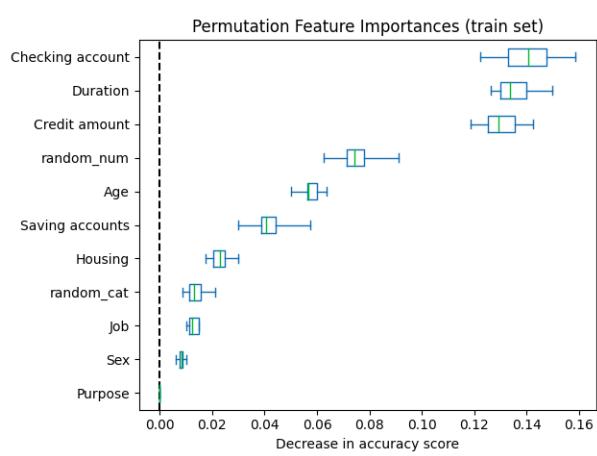


Figura 4.10: Permutation Feature Importances (train set)

Interpretación de los resultados

Los resultados muestran que la importancia de características basada en la impureza puede inflar la importancia de características numéricas y que sufre especialmente con el sobreajuste. En contraste, Permutation Feature Importance proporciona una medida más robusta y generalizable de la importancia de las características, revelando que las variables aleatorias no predictivas tienen importancias cercanas a cero. Sin embargo, se enfatiza que no se pretende desacreditar el uso de FI, sino resaltar la importancia de comprender las limitaciones de cada método. En contextos donde se requiera una interpretación más intuitiva y generalizable, especialmente para audiencias no especializadas, PFI puede ser una opción más adecuada.

4.2.3. Análisis de sesgo en función del género utilizando SHAP

Al principio de esta base de datos se comentó que se mantenía la variable que representa el género para este experimento. En este experimento, exploramos el sesgo en función del género en modelos de clasificación de riesgo crediticio utilizando la biblioteca SHAP (SHapley Additive exPlanations). El

objetivo es analizar si los modelos de aprendizaje automático muestran sesgo en función del género y cómo la interpretabilidad de los modelos puede ayudar a identificar y abordar dicho sesgo.

Si usamos la base de datos original con la variable del género presente, todos los modelos entrenados utilizan esta variable para discriminar los datos (puede comprobarse los PFI de los modelos en el anexo B.1.5). A partir de aquí nos surge la siguiente pregunta: ¿si quitamos esta variable nuestros modelos seguirán discriminando en función del género o no? Gracias a SHAP y, en particular al siguiente enlace de su documentación [Explaining quantitative measures of fairness](#), podemos comprobar esto fácilmente analizando la distribución de los valores de SHAP para diferentes grupos.

Antes de mostrar los resultados, en la tabla 4.2 y en la figura 4.11 podemos observar las distribuciones de las cantidades de préstamos y cantidades solicitadas en función del género y el riesgo. Viendo estos datos, observamos que la distribución de créditos peligrosos son similares, pero tenemos menos cantidad de información sobre mujeres y además la distribución de la cantidad prestada en total se ve muy afectada por esto. A partir de estos datos podemos intuir que el modelo puede penalizar a un grupo social por falta de datos de manera errónea, pues no debemos permitir un sesgo de este tipo.

Género	Riesgo	Número de Créditos	Cantidad Media Crédito	Cantidad Total Crédito
Femenino	False	201	2.555,98	513.751
	True	109	3.471,18	378.359
Masculino	False	499	3.158,45	1.576.069
	True	191	4.204,60	803.079

Tabla 4.2: Créditos por género y riesgo

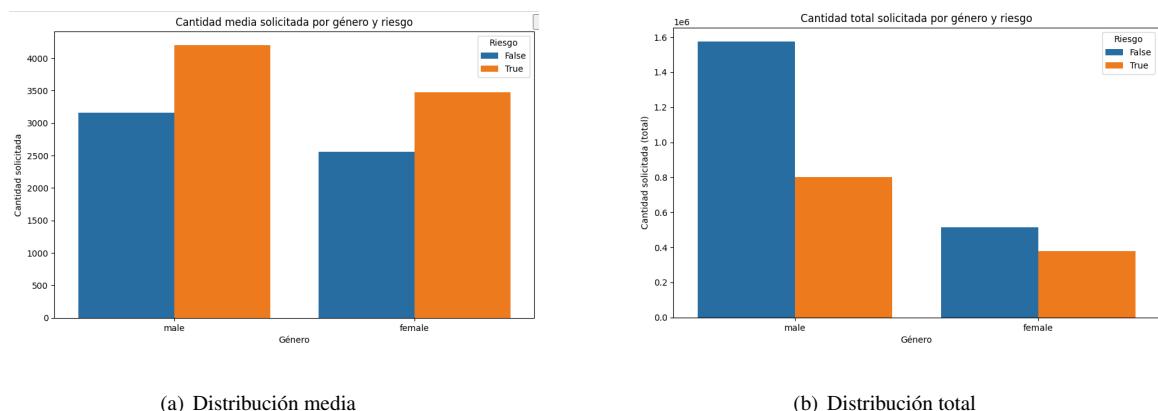


Figura 4.11: Distribución cantidad de crédito prestada por género y riesgo.

Resultados

Tras eliminar la característica del género, podemos fácilmente aplicar SHAP a los modelos entrenados para observar si hay diferencias en la distribución de estos valores en función del género. En

particular, SHAP permite esto con el método *group_difference_plot*. En la figura 4.12 observamos la diferencia en la paridad demográfica calculada por SHAP.

Observamos que el modelo de random forest exhibe un sesgo significativo en función del género. La diferencia de paridad demográfica (Demographic Parity Difference) entre mujeres y hombres es de $-0,23$, lo que indica una clasificación desigual del riesgo crediticio: en promedio, ser hombre reduce en un factor de $-0,23$ la probabilidad de ser clasificado como mal crédito, es decir, el modelo favorece a los hombres. Por otro lado, la red neuronal no muestra un sesgo significativo en función del género, con una diferencia de paridad demográfica cercana a 0. Esto sugiere que la red neuronal clasifica el riesgo crediticio de manera más equitativa para mujeres y hombres.

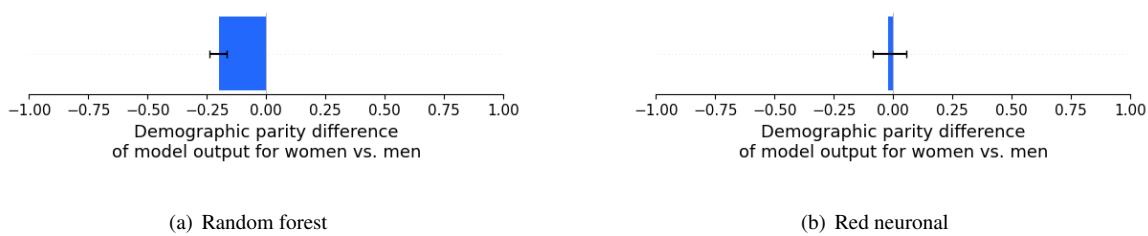


Figura 4.12: Diferencia de paridad demográfica por género para random forest y red neuronal.

Como comentario adicional, y gracias a que SHAP permite fácilmente mostrar la descomposición de esta diferencia en la paridad demográfica desglosada por características, observamos en la figura 4.13 las características con los valores más altos. Viendo esta descomposición, que puede observarse en su totalidad en el anexo B.1.5, podemos suponer que el árbol de decisión se ha sobreajustado a los datos y por eso sesga con prácticamente todas las características en contra de la mujer. En el caso de la red neuronal, se observa que algunas características tienen valores SHAP negativos y otras positivos, lo que lleva a una suma cercana a cero. Esto indica que el modelo balancea bien las contribuciones de diferentes características, evitando un sesgo significativo hacia cualquier género. Esta capacidad de las redes neuronales para equilibrar las influencias de varias características podría explicar por qué muestra una mejor equidad en comparación con el random forest.

Interpretación de los resultados

Estos resultados destacan la importancia de evaluar y abordar el sesgo en los modelos de aprendizaje automático, especialmente en aplicaciones sensibles como la clasificación de riesgo crediticio. La interpretabilidad de los modelos, utilizando herramientas como SHAP, puede ayudar a identificar y corregir sesgos potenciales, promoviendo la equidad y transparencia en las decisiones automatizadas.

En el caso del random forest, el sesgo significativo en contra de las mujeres sugiere que este modelo podría estar aprendiendo patrones discriminatorios a partir de los datos de entrenamiento. Esto puede estar influenciado por la presencia de características correlacionadas con el género que el modelo utiliza indirectamente para tomar decisiones. La descomposición de SHAP muestra qué

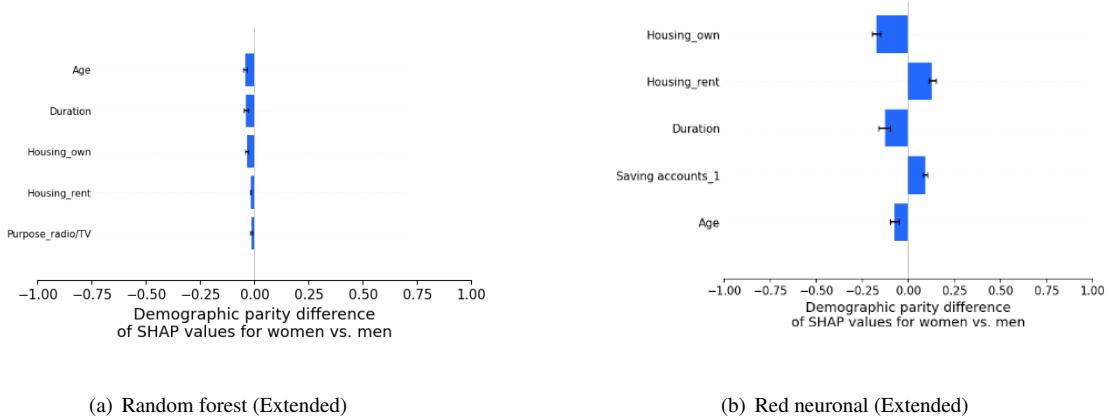


Figura 4.13: Diferencia de paridad demográfica descompuesta por características para random forest y red neuronal.

características específicas están contribuyendo más a este sesgo, lo cual es crucial para entender y mitigar el problema.

Por otro lado, la red neuronal muestra una mejor equidad en la clasificación del riesgo crediticio. Esto puede deberse a la capacidad de las redes neuronales para aprender representaciones más complejas y generalizar mejor a partir de los datos, diluyendo el efecto de las características que podrían inducir sesgo. La diferencia en la paridad demográfica cercana a cero indica que el modelo trata de manera más justa a ambos géneros.

En resumen, este análisis subraya la necesidad de utilizar métodos interpretables como SHAP para evaluar la equidad de los modelos de aprendizaje automático. Al identificar y cuantificar el sesgo, se pueden tomar medidas correctivas para asegurar que los modelos sean justos y equitativos, contribuyendo a una toma de decisiones más responsable en el ámbito del riesgo crediticio.

4.3. Base de datos de satisfacción de clientes del banco Santander

La base de datos utilizada en este experimento proviene de un concurso de Kaggle organizado por Santander en 2016 para predecir si un cliente está insatisfecho ([enlace a la base de datos](#)). Para esta base de datos se ha realizado un balanceo de los datos utilizando la librería *SMOTE*, como se sugiere en el concurso para manejar el desbalanceo de clases. Se ha escogido esta base de datos por tener más de 300 características.

4.3.1. Límites de la interpretabilidad

En este trabajo, se analiza cómo las redes neuronales superan a *random forest* y regresión logística en capacidad predictiva al utilizar un mayor número de variables. Sin embargo, esto complica la interpretabilidad del modelo, de una manera que no se aborda en profundidad en este TFG. Los resultados de precisión de los modelos se encuentran en el Anexo B.1.6.

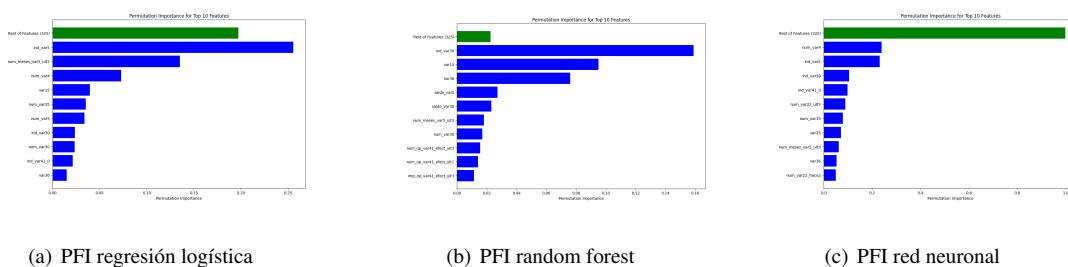


Figura 4.14: Comparación de la importancia de las características por permutación entre los modelos. Se muestran las importancias de las diez características más importantes en azul y en verde la suma del resto de características.

A partir de los resultados mostrados en la figura 4.14, podemos observar: En el caso de la regresión logística, no enfrentamos el mismo problema con el número de variables debido a su naturaleza interpretable, además, la suma de la importancia del resto de las variables es relativamente baja (0,2). Para random forest, esta suma del resto de características es insignificante (0,02), permitiendo que se pueda mantener la interpretabilidad al considerar únicamente las diez variables más importantes. Sin embargo, en el Anexo B.1.6 se puede comprobar que es la red neuronal la que obtiene una mejora significativa en la precisión. Para este modelo, no es posible simplificar tanto sin perder rendimiento, lo cual presenta un desafío no abordado en este TFG: cómo mantener la interpretabilidad cuando se requiere comprender muchas variables, algo difícil especialmente para personas no especializadas.

El objetivo de este experimento es mostrar cómo el problema de la interpretabilidad no es tan sencillo como podría parecer con bases de datos más accesibles, y sirve como punto de partida para ideas sobre futuros trabajos en este tema.

4.4. Clasificador de análisis de sentimientos

Al igual que el resto de experimentos, este se centra en el ámbito económico, en particular se ha utilizado la siguiente base de datos disponible en el siguiente [enlace](#). Este conjunto de datos contiene textos financieros en inglés clasificados como positivos, neutrales o negativos.

Para preprocesar los datos, se implementaron las siguientes etapas:

- **Tokenización:** Separación de oraciones en palabras individuales.
- **Eliminación de palabras vacías:** Remoción de palabras muy repetidas y sin valor, como “a”, “by”, “for”, etc.
- **Lematización:** Conversión de las palabras a sus formas base o lemas.
- **Reconstrucción:** Reensamblaje del texto preprocesado en cadenas de palabras.

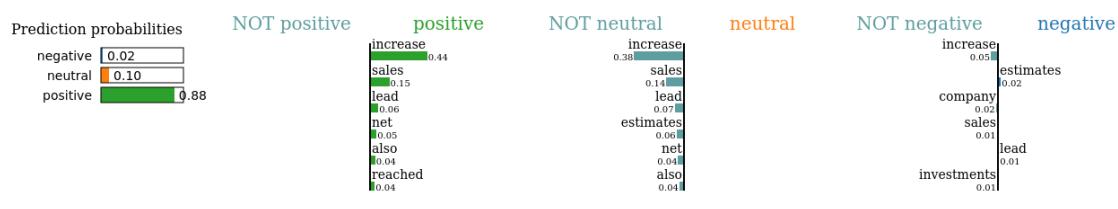
Cabe destacar que la función de preprocesamiento fue ajustada tras la primera versión del modelo. Originalmente, se observó que se tomaban como indicadores de sentimientos números o años, lo cual no era adecuado. Gracias a la interpretación con LIME, se identificó y corrigió este problema, eliminando números, signos de dólar (\$) y referencias a euros (EUR) del texto. Se muestra en el anexo B.2.1 la salida de LIME previa a estos cambios que sirvió para detectar el error.

Se entrenó un clasificador random forest utilizando el texto preprocesado. El objetivo fue interpretar un modelo capaz de clasificar la connotación de las oraciones, no obtener buenos resultados de clasificación. Al tener una base de datos con solo 5.000 entradas, estos resultados están totalmente sesgados por los datos como vamos a poder observar.

Tras el preprocesamiento de datos de texto y el entrenamiento del clasificador, se aplicaron LIME y SHAP para interpretar los resultados del clasificador de análisis de sentimientos.

4.4.1. Resultados

Aplicamos LIME y SHAP de manera local, para interpretar clasificaciones de un texto en particular. Esta es la manera más sencilla de interpretar los resultados y, repitiendo este proceso para ejemplos de las distintas posibles salidas, somos capaces de generalizar estas interpretaciones. A continuación, se muestran los resultados de LIME para una frase clasificada como positiva en la figura 4.15.



Text with highlighted words

The company also estimates the already carried out investments to lead to an increase in its net sales for 2010 from 2009 when they reached EUR 141.7 million

Figura 4.15: Interpretación LIME para un texto clasificado como positivo.

Al igual que vemos en los resultados obtenidos con SHAP en la figura 4.16, parece un buen resultado que ambos métodos muestran como el modelo aumenta la probabilidad de la clase positiva en función de palabras con connotaciones positivas como puede ser *increase*, al mismo tiempo que disminuyen la probabilidad del resto de clases. También hay que destacar palabras como *sales*, en el contexto de esta oración positiva aumentan la probabilidad, si bien en el anexo B.2.1 ya mencionado se puede ver que también se tienen en cuenta con el contexto negativo.

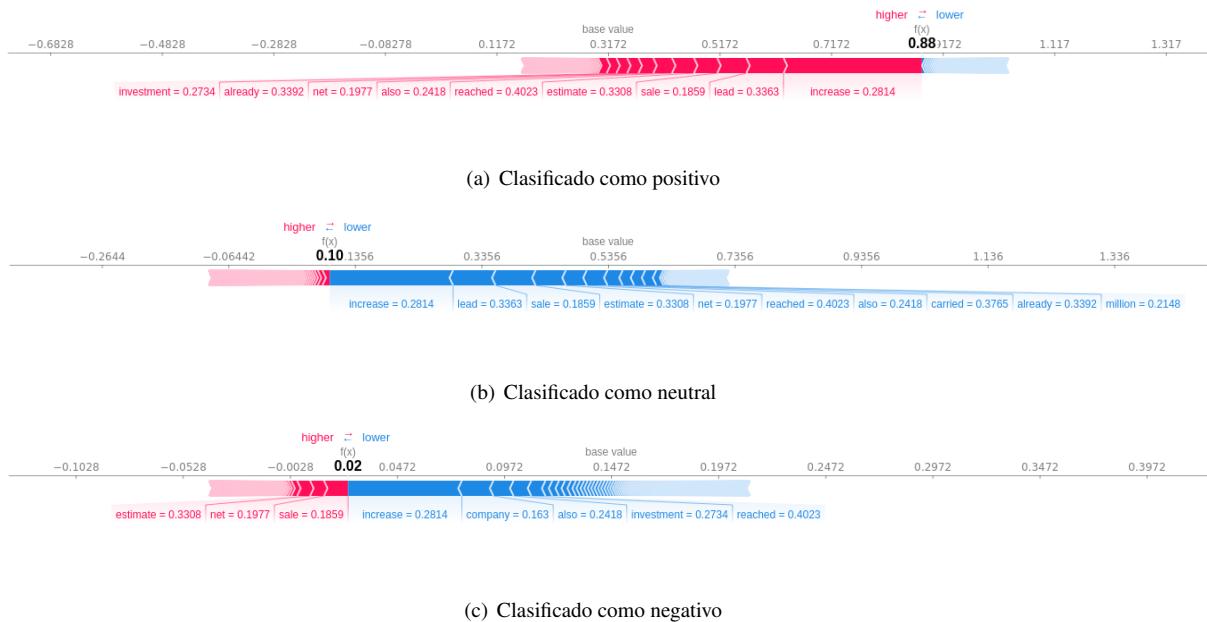


Figura 4.16: Interpretación SHAP para un texto clasificado como positivo.

4.4.2. Interpretación de los resultados

El uso de LIME y SHAP ha sido clave para entender el funcionamiento del modelo, identificar errores en el preprocesamiento y mejorarlo. En el anexo B.2.1 se muestran los resultados de aplicar LIME y SHAP para ejemplos de textos clasificados como neutrales o negativos. A partir de la interpretación de estas explicaciones se generalizó la siguiente interpretación del modelo de manera global:

El modelo de análisis de sentimientos tiende a clasificar inicialmente las oraciones como neutrales. A partir de ahí, busca indicios de connotaciones positivas o negativas. Palabras específicas pueden aumentar la probabilidad de una clase particular, como *increase* y *sales* que indican positividad, mientras que *declined* sugiere negatividad. Además, hay palabras cuyo impacto depende del contexto, como *lead* o *reached*, que pueden tener connotaciones tanto positivas como negativas.

La interpretabilidad con LIME y SHAP ha sido clave para entender el funcionamiento del modelo, identificar errores en el preprocesamiento y mejorar su rendimiento. También ha permitido detectar problemas de sobreajuste, como con la palabra *Helsinki*, que el modelo asocia incorrectamente con una connotación neutral debido a su distribución en el conjunto de datos.

CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

En este trabajo se ha mostrado el contexto formal del que partimos sobre modelos de clasificación y a partir de esto se ha explorado el uso de técnicas de interpretabilidad de manera teórica y práctica mediante la aplicación a diferentes conjuntos de datos económicos. También se ha utilizado una base de datos artificial para mostrar una cierta falta de robustez de PDP y ALE, y una base de datos sobre clasificación de texto para extender el uso de LIME y SHAP a este tipo de problemas. A lo largo de este documento, se han planteado varios objetivos específicos que han guiado la investigación. A continuación, se presenta un resumen del grado de consecución de dichos objetivos:

- 1.– Mostrar el contexto actual de los modelos de clasificación y la importancia de su interpretabilidad. Este objetivo se abordó en la introducción y en el capítulo de estado del arte, donde se detallaron los conceptos y la relevancia de la interpretabilidad en modelos de machine learning.
- 2.– Realizar un estudio formal de la interpretabilidad de modelos de clasificación, incluyendo un análisis del estado del arte. Este objetivo se logró mediante la revisión de técnicas de interpretabilidad como LIME, SHAP, PDP y ALE, y su aplicación en diversos contextos.
- 3.– Demostrar la utilidad práctica de las técnicas de interpretabilidad estudiadas y sus aplicaciones. Esto se hizo aplicando dichas técnicas a casos prácticos como la base de datos German Credit Risk y el análisis de sentimiento, mostrando cómo mejoran la comprensión y la calidad de los modelos.
- 4.– Profundizar en conceptos específicos de interpretabilidad mediante experimentos detallados. Se realizaron experimentos con PDP y ALE para evaluar su robustez, se compararon FI y PFI con random forest, y se analizaron sesgos de género utilizando SHAP, cumpliendo así este objetivo.

Los experimentos realizados han arrojado resultados significativos:

- En el experimento de PDP y ALE, se observó una discrepancia notable entre los resultados teóricos y los obtenidos por el modelo, resaltando la complejidad de los problemas de clasificación y la necesidad de mejorar la precisión de la representación del modelo.
- La comparación entre FI y PFI con random forest reveló que PFI es una medida más robusta y generalizable, especialmente en presencia de sobreajuste, mientras que FI puede inflar la importancia de características numéricas.
- El análisis de sesgo de género mostró que los modelos pueden aprender patrones discriminatorios a partir de los datos de entrenamiento, pero herramientas interpretativas como SHAP pueden ayudar a identificar y mitigar

estos sesgos.

- En el estudio de la base de datos de satisfacción de clientes del banco Santander, se destacó que las redes neuronales superan a los modelos de random forest y regresión logística en precisión predictiva, aunque a costa de una menor interpretabilidad. En particular, el aumento de las variables a tener en cuenta es un todavía reto que enfrentar.
- El clasificador de análisis de sentimientos, mediante el uso de LIME y SHAP, permitió entender y mejorar el modelo, identificando errores de preprocesamiento y problemas de sobreajuste.

En conclusión, todos los objetivos planteados al inicio de este trabajo han sido alcanzados satisfactoriamente, permitiendo no solo una mejor comprensión de los modelos utilizados, sino también una mejora en su rendimiento y utilidad práctica.

5.2. Trabajos futuros

Existen varias líneas de investigación y desarrollo que podrían ampliarse a partir de este trabajo:

- Extensión de técnicas de interpretabilidad a otros tipos de modelos y datos para validar su generalizabilidad.
- Desarrollo de nuevas técnicas de interpretabilidad que ofrezcan explicaciones más precisas y útiles.
- Implementación de estas técnicas en entornos productivos, especialmente en sectores críticos como la banca y las finanzas.
- Estudio del impacto de la interpretabilidad en la confianza y aceptación de los usuarios finales.
- Integración de técnicas de interpretabilidad con herramientas de AutoML para facilitar la adopción de modelos complejos.
- Creación de herramientas automáticas para abordar problemas con un gran número de variables y mantener la interpretabilidad.

Estas direcciones futuras no solo profundizan en la investigación iniciada, sino que también abren nuevas oportunidades para mejorar y aplicar las técnicas de interpretabilidad en machine learning en diversos contextos y aplicaciones.

BIBLIOGRAFÍA

- [1] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2020. Leer.
- [2] A. Goldstein, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," 2014. Leer.
- [3] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," 2016. Leer.
- [4] M. R. Muñoz, "Estudio de la explicabilidad de modelos predictivos para la covid-19," Junio 2021. Trabajo Fin de Máster, Máster en Big Data y Data Science: ciencia e ingeniería de datos, Escuela Politécnica Superior, Universidad Autónoma de Madrid.
- [5] A. N. Juscafresa, "An introduction to explainable artificial intelligence with lime and shap." https://deposit.ub.edu/dspace/bitstream/2445/192075/1/tfg_nieto_juscafresa_aleix.pdf, 2022. Treball final de grau, Grau de Matemàtiques, Departament de Matemàtiques i Informàtica, Universitat de Barcelona.
- [6] J. David W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Wiley Series in Probability and Statistics, Wiley, 3rd ed., 2013.
- [7] J. M. Hilbe, *Logistic Regression Models*. Chapman and Hall/CRC, 2009.
- [8] J. R. Quinlan, *Induction of Decision Trees*, vol. 1. Springer, 1986.
- [9] T. G. Dietterich, *Ensemble Methods in Machine Learning*, vol. 1857. Springer, 2000.
- [10] M. Schott, "Random forest algorithm for machine learning." <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9fe> April 25 2019. Accedido: Junio 2024.
- [11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. Accedido: Junio 2024.
- [12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [13] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 82, no. 4, pp. 1059–1086, 2020.
- [14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," 2016. Accedido: Junio 2024.
- [16] M. T. Ribeiro, "Lime - local interpretable model-agnostic explanations." <https://homes.cs.washington.edu/marcotcr/blog/lime/>, 2016. Accedido: Junio 2024.
- [17] DeepFindr, "Explainable ai explained! | #3 lime." https://www.youtube.com/watch?v=d6j6bofhj2M&t=478s&ab_channel=DeepFindr, 2021. Accedido: Junio 2024.

- [18] L. Shapley, “A value for n-person games,” in *Contributions to the Theory of Games* (H. W. Kuhn and A. W. Tucker, eds.), vol. 2 of *Annals of Mathematics Studies*, pp. 307–317, Princeton University Press, 1953.
- [19] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent individualized feature attribution for tree ensembles,” *Proceedings of the National Academy of Sciences*, 2018. Accedido: Junio 2024.
- [20] Kie, “Shapley additive explanations (shap).” https://www.youtube.com/watch?v=VB9uv-x0gtg&t=219s&ab_channel=KIE, 2021. Accedido: Junio 2024.
- [21] C. Molnar, “Characteristics of shapley values.” <https://stats.stackexchange.com/q/389261>, 2019. Accedido: Junio 2024.
- [22] M. L. TV, “Understanding the shapley value.” https://www.youtube.com/watch?v=9OFMRiAVH-w&ab_channel=MachineLearningTV, 2021. Accedido: Junio 2024.
- [23] DeepFindr, “Explainable ai explained! | #4 shap.” https://www.youtube.com/watch?v=9haIOp1EIGM&t=490s&ab_channel=DeepFindr, 2021. Accedido: Junio 2024.
- [24] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, 2017. Accedido: Junio 2024.
- [25] “Scikit-learn: Machine learning in python.” <https://scikit-learn.org/>. Accedido: Junio 2024.
- [26] “Alepython: Accumulated local effects (ale) plots.” <https://github.com/blange/aleplot>. Accedido: Junio 2024.
- [27] “Lime package.” <https://github.com/marcotcr/lime>. Accedido: Junio 2024.
- [28] “Shap package.” <https://shap.readthedocs.io/>. Accedido: Junio 2024.
- [29] “Nltk package.” <https://www.nltk.org/>. Accedido: Junio 2024.
- [30] M. Sahakyan, Z. Aung, and T. Rahwan, “Explainable artificial intelligence for tabular data: A survey,” *IEEE Access*, vol. 9, pp. 135392–135422, 2021.
- [31] A. Rosebrock, “Introduction to hyperparameter tuning with scikit-learn and python.” <https://pyimagesearch.com/2021/05/17/introduction-to-hyperparameter-tuning-with-scikit-learn-and-python/>, 2021. Accedido: Junio 2024.
- [32] N. S. Chauhan, “Model evaluation metrics in machine learning.” <https://towardsdatascience.com/model-evaluation-metrics-in-machine-learning-e41a3c1fd78a>, May 2020. Accedido: Junio 2024.
- [33] “H2o automl: Automatic machine learning.” <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html>. Accedido: Junio 2024.

APÉNDICES

COMPLEMENTOS STATE OF ART

En este capítulo del anexo se muestran algunos conceptos del estudio de modelos o métodos de interpretabilidad que no son necesarios para entender este TFG, pero que pueden servir para saciar la curiosidad y que, tras entender lo tratado en este TFG, pueden servir como un complemento que no requiere mucho esfuerzo digerir.

A.1. Modelos de clasificación

A.1.1. Regresión lineal

Introducción

La regresión lineal es una técnica fundamental en ciencias de la computación para modelar y analizar relaciones entre variables. Aunque no voy a centrarme en esto, quiero formalizar algunos detalles matemáticos importantes para comprender la esencia del método y su aplicación en problemas de la vida real. La regresión lineal es una herramienta poderosa para prever y entender patrones, y su utilidad se extiende a áreas como el aprendizaje automático, la inteligencia artificial y el análisis de datos.

Conceptos básicos

Modelo lineal

En esencia, la regresión lineal busca ajustar una línea (o hiperplano en dimensiones superiores) que represente la mejor aproximación lineal a los datos observados. La relación entre la variable dependiente (y) y las variables independientes (x_1, x_2, \dots, x_n) se modela como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (\text{A.1})$$

Donde (β_0) es el término de intercepción, ($\beta_1, \beta_2, \dots, \beta_n$) son los coeficientes asociados a las

variables independientes y (ϵ) es el término de error.

Obtención de pesos por mínimos cuadrados

Aunque podemos simplificar los detalles matemáticos, es esencial entender cómo se obtienen los pesos. El objetivo es encontrar valores para $(\beta_0, \beta_1, \dots, \beta_n)$ que minimicen el error entre las predicciones del modelo y los valores reales observados. Esto se logra aplicando el método de mínimos cuadrados, que implica minimizar la suma de los cuadrados de los residuos, es decir, la diferencia entre las predicciones y los valores reales. Puede demostrarse que bajo ciertas hipótesis sobre nuestros datos la fórmula para calcular los coeficientes por mínimos cuadrados es: $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Donde:

- \mathbf{X} es la matriz de datos de las variables independientes.
- \mathbf{Y} es el vector de datos de las variables dependientes reales.
- $\hat{\beta}$ es el vector de coeficientes estimados.

Implementación en ciencias de la computación

Aprendizaje automático

En el contexto del aprendizaje automático, la regresión lineal se utiliza para crear modelos predictivos. Por ejemplo, puede emplearse para prever el rendimiento de un algoritmo en función de diversos parámetros, o para estimar el tiempo de ejecución de un programa.

Análisis de datos

En el análisis de datos, la regresión lineal puede emplearse para identificar relaciones entre variables, proporcionando información valiosa sobre cómo ciertos factores afectan otros.

Mínimos cuadrados y teorema de Gauss-Markov

El proceso de obtención de los pesos por mínimos cuadrados asegura que los coeficientes $(\beta_0, \beta_1, \dots, \beta_n)$ minimizan la suma de los cuadrados de los residuos. Este método no solo proporciona estimadores insesgados, sino que, según el Teorema de Gauss-Markov, también garantiza que estos estimadores sean los mejores estimadores lineales insesgados (BLUE, por sus siglas en inglés) en términos de varianza.

Teorema de Gauss-Markov

Sea $\check{\beta}$ un estimador lineal insesgado de β en nuestro modelo lineal con $E[\epsilon] = 0$ y $\text{Var}[\epsilon] = \sigma^2 I$. Entonces, la varianza de $\check{\beta}$ es mayor o igual a la varianza de $\hat{\beta}$, nuestro estimador de mínimos cuadrados, para todos los coeficientes. Formalmente,

$$\text{Var}[\check{\beta}_i] \geq \text{Var}[\hat{\beta}_i] \quad \forall i = 0, 1, \dots, n$$

Este teorema destaca la eficiencia de los estimadores obtenidos por mínimos cuadrados en términos de varianza.

Consideraciones prácticas y desafíos

Aunque la regresión lineal es una herramienta poderosa, hay desafíos en su aplicación. La elección de variables adecuadas, la validación del modelo y la interpretación de los resultados son aspectos importantes a considerar. Además está limitada a modelos muy semejantes a los lineales, es decir, modelos muy sencillos.

Conclusión e interpretabilidad

La regresión lineal, a pesar de su simplicidad, es una técnica valiosa en ciencias de la computación. Desde el análisis de datos hasta el aprendizaje automático, su versatilidad y facilidad de interpretación la convierten en una herramienta esencial para comprender y modelar relaciones entre variables en problemas informáticos del mundo real. La combinación de mínimos cuadrados y el Teorema de Gauss-Markov proporciona una base sólida para obtener estimadores eficientes y precisos.

A.1.2. Regresión logística

Máxima verosimilitud

La optimización de la regresión logística se basa en la maximización de la función de verosimilitud. Esta función se define como el producto de las probabilidades de cada observación, asumiendo independencia entre ellas. Equivalentemente, se minimiza la log-verosimilitud, dada por:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^k y^{(i)} \boldsymbol{\beta}^\top \mathbf{x}^{(i)} - \log(1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}^{(i)})) \quad (\text{A.2})$$

Obtención de los pesos

La obtención de los pesos en la regresión logística implica el uso del método de Newton-Raphson. La actualización de los pesos se realiza mediante la siguiente fórmula:

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - [\mathbf{X} \mathbf{W}^{(s)} \mathbf{X}^\top]^{-1} \mathbf{X} (\mathbf{Y} - \hat{\mathbf{Y}}^{(s)}) \quad (\text{A.3})$$

Aquí, \mathbf{W} es una matriz diagonal que contiene los productos de probabilidad y su complemento para cada observación $p(x^{(i)})(1 - p(x^{(i)}))$.

El método de Newton-Raphson es eficaz para converger rápidamente hacia el óptimo local de la función de verosimilitud en la regresión logística. Sin embargo, es importante señalar que, aunque este método es eficiente, puede no converger en casos de separación perfecta o multicolinealidad extrema en los datos.

A.1.3. Redes neuronales artificiales

Algoritmo de retropropagación (Backpropagation)

El algoritmo de retropropagación es utilizado para entrenar una red neuronal. Dado un conjunto de entrenamiento con entradas X y salidas deseadas Y , la retropropagación ajusta los pesos y sesgos minimizando una función de pérdida. El descenso por gradiente se utiliza para la optimización. Las actualizaciones de pesos se realizan según:

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \alpha \frac{\partial J}{\partial w_{ij}^{(l)}}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \alpha \frac{\partial J}{\partial b_j^{(l)}}$$

donde α es la tasa de aprendizaje y J es la función de pérdida.

Funciones de activación comunes

- **Sigmoide:**

$$f(z) = \frac{1}{1 + e^{-z}}$$

- **ReLU (Rectified Linear Unit):**

$$f(z) = \max(0, z)$$

- **Tangente hiperbólica (tanh):**

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

Propiedades de las redes neuronales

- **Universalidad Aproximada:** Una red neuronal con una sola capa oculta puede aproximar cualquier función continua en un dominio compacto.
- **No Convexidad:** El espacio de parámetros de una red neuronal suele ser no convexo, lo que puede llevar a mínimos locales en lugar de globales.
- **Regularización:** Técnicas como la regularización $L1$ y $L2$ se utilizan para evitar el sobreajuste durante el entrenamiento.

A.2. Métodos de interpretabilidad

A.2.1. Partial Dependence Plot

Importancia de características basada en PDP

Existe una medida simple de la importancia de características basada en la dependencia parcial. La motivación básica es que un PDP plano indica que la característica no es importante, y cuanto más varía el PDP, más importante es la característica. Para características numéricas, la importancia se define como la desviación de cada valor de la característica respecto a la curva promedio:

$$I(x_S) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K \left(\hat{f}_S(x_S^{(k)}) - \frac{1}{K} \sum_{k=1}^K \hat{f}_S(x_S^{(k)}) \right)^2} \quad (\text{A.4})$$

Aquí, $x_S^{(k)}$ son los K valores de la característica X_S . Para características categóricas, tenemos:

$$I(x_S) = \frac{\max_k \hat{f}_S(x_S^{(k)}) - \min_k \hat{f}_S(x_S^{(k)})}{4} \quad (\text{A.5})$$

Esta forma de aproximar la desviación se llama regla del rango. El denominador (cuatro) proviene de la distribución normal estándar: en la distribución normal, el 95 % de los datos están dentro de dos desviaciones estándar alrededor de la media. Entonces, el rango dividido por cuatro da una estimación aproximada que probablemente sea menor que la varianza real.

Esta medida de importancia basada en PDP debe interpretarse con precaución. Captura solo el efecto principal de la característica y pasa por alto posibles interacciones entre características. Una característica podría ser muy importante según otros métodos pero la PDP podría ser plana si la característica afecta la predicción principalmente a través de interacciones con otras características. Otra desventaja de esta medida son que las características con solo una instancia se ponderan de la misma manera en el cálculo de la importancia que un valor con muchas instancias.

A.2.2. Acumulated Local Effects Plot

Motivación e intuición

Cuando las características de un modelo de aprendizaje automático están correlacionadas, la gráfica de dependencia parcial no es fiable. Promediamos las características seleccionadas sobre todos los posibles valores del resto de características, pudiendo dar lugar a datos artificiales no realistas, por ejemplo una casa de $30m^2$ con 10 habitaciones.

Podemos evitar estas instancias improbables promediando sobre la distribución condicional de la

característica, lo que significa que en un valor de la cuadrícula de x_1 , promediamos las predicciones de las instancias con un valor similar de x_1 . Se denominan Gráficas Marginales o M-Plots pero no son suficiente pues aunque la característica que estamos estudiando sea irrelevante mostraría el efecto que produce la variable correlacionada a la predicción.

Los gráficos ALE solucionan el problema al calcular, también basándose en la distribución condicional de las características, las diferencias en las predicciones en lugar de promediarlas. Esto nos da el efecto de la característica y sin mezclarla con los efectos de características correlacionadas. En la figura A.1 mostramos un ejemplo en el cual vemos claramente el problema de los PDP al promediar sobre todos los posibles valores de una variable.

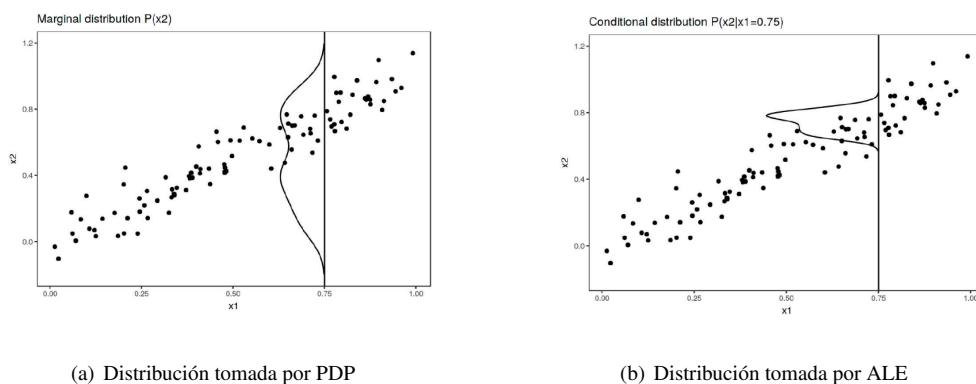


Figura A.1: En esta figura se muestra la comparación entre los valores tomados para obtener PDP A.1(a) y ALE A.1(b) para el mismo problema con dos variables correlacionadas.

Estimación de ALE con una característica numérica

- 1.- **Definir cuantiles:** Seleccionar cuantiles z_k para dividir la característica de interés en intervalos.
- 2.- **Dividir la característica:** Crear intervalos $[z_{k-1}, z_k]$ utilizando los cuantiles seleccionados.
- 3.- **Calcular predicciones:** Calcular las predicciones del modelo para cada intervalo.
- 4.- **Calcular las diferencias acumuladas:** Para cada intervalo $[z_{k-1}, z_k]$: Calcular la diferencia en predicciones: $f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})$, donde x_j es el valor de la característica de interés y n_j es el número de instancias en el intervalo.
- 5.- **Acumular las diferencias:** Acumular las diferencias calculadas a lo largo de los intervalos.

Fórmula Simplificada del ALE Plot:

$$\hat{f}_{j;\text{ALE}}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} [f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})] \quad (\text{A.6})$$

Donde:

- $\hat{f}_{j;\text{ALE}}(x)$ es la estimación ALE para la característica x_j .
- $n_j(k)$ es el número total de instancias en la característica de interés.

- k_j es el número de intervalos definidos por los cuantiles z_k .
- $N_j(k)$ es el conjunto de instancias en el intervalo $[z_{k-1}, z_k]$.

Esta fórmula expresa cómo se calcula la estimación ALE acumulando las diferencias de predicciones en cada intervalo y promediándolas.

Este efecto se centra para que el efecto promedio sea cero:

$$\hat{f}_{j;\text{ALE}}(x) = \hat{\bar{f}}_{j;\text{ALE}}(x) - \frac{1}{k} \sum_{i=1}^k \hat{f}_{j;\text{ALE}}(x_j^{(i)}) \quad (\text{A.7})$$

A.2.3. Feature interaction

Aristóteles dijo “El todo es mayor que la suma de sus partes”. Cuando las características interactúan entre sí en un modelo de predicción, la predicción no puede expresarse como la suma de los efectos de las características, ya que el efecto de una característica depende del valor de la otra característica.

Intuición

Cuando un modelo de aprendizaje automático utiliza dos características para hacer predicciones, estas se descomponen en cuatro términos: un término constante, un término para la primera característica, un término para la segunda característica y un término para la interacción entre ambas.

En una situación sin interacción, como en el ejemplo siguiente:

Ubicación	Tamaño	Predicción
buena	grande	300,000
buena	pequeña	200,000
mala	grande	250,000
mala	pequeña	150,000

La predicción se descompone en términos constante, de tamaño y de ubicación, sin necesidad de un término de interacción.

En cambio, en presencia de interacción, como en el siguiente ejemplo:

Ubicación	Tamaño	Predicción
buena	grande	400,000
buena	pequeña	200,000
mala	grande	250,000
mala	pequeña	150,000

La descomposición incluye un término adicional para la interacción entre tamaño y ubicación (+100,000 si la casa es grande y está en una buena ubicación), ya que en este caso, la diferencia en la predicción entre una casa grande y pequeña depende de la ubicación.

Teoría: estadístico H de Friedman

Interacción de dos características

En el caso de dos características, podemos descomponer la función de dependencia parcial de la siguiente manera (asumiendo que las funciones de dependencia parcial están centradas en cero):

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k) \quad (\text{A.8})$$

Si dos características no interactúan, la función de dependencia parcial bidireccional se expresa como la suma de las funciones de dependencia parcial individuales.

El estadístico de fuerza de la interacción entre dos características (j y k) se mide mediante el estadístico H_{jk}^2 :

$$H_{jk}^2 = \frac{\sum_{i=1}^n [PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)})]^2}{\sum_{i=1}^n PD_{jk}^2(x_j^{(i)}, x_k^{(i)})} \quad (\text{A.9})$$

Este estadístico mide la diferencia entre la función de dependencia parcial observada y la descompuesta sin interacciones.

Interacción de una característica con el resto

Si una característica no interactúa con las demás, la función de predicción $\hat{f}(x)$ se expresa como la suma de la función de dependencia parcial individual y la función de dependencia parcial que involucra a todas las demás características salvo la característica j :

$$\hat{f}(x) = PD_j(x_j) + PD_{-j}(x_{-j}) \quad (\text{A.10})$$

El estadístico de fuerza de la interacción entre una característica (j) y todas las demás características se mide mediante el estadístico H_j^2 :

$$H_j^2 = \frac{\sum_{i=1}^n [\hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)})]^2}{\sum_{i=1}^n \hat{f}^2(x^{(i)})} \quad (\text{A.11})$$

Este estadístico mide la diferencia entre la función de predicción observada y la descompuesta sin interacciones.

Ventajas y desventajas del estadístico H de Friedman

Ventajas

- Tiene una fuerte bases teórica.
- Interpretación significativa: la interacción se define como la proporción de varianza explicada por la interacción.
- Adimensional y siempre entre 0 y 1, lo que permite comparaciones entre características y modelos.
- Detecta todo tipo de interacciones, incluso con más de dos características.

Desventajas

- Computacionalmente costoso: en el peor de los casos $2n^2$ llamadas para el estadístico bidireccional y $3n^2$ para el estadístico total.
- No revela la naturaleza específica de las interacciones.
- Inaplicable a entradas de píxeles en clasificadores de imágenes.
- Dependen de la suposición de independencia al cambiar el orden de las características.

A.2.4. Global surrogate

Un modelo de sustitución global es un modelo interpretable para aproximar las predicciones de un modelo de caja negra. Podemos obtener conclusiones sobre el modelo de caja negra interpretando las predicciones del modelo de sustitución. Es un método agnóstico, ya que no requiere información sobre el funcionamiento interno del modelo de caja negra, solo se necesita acceso a los datos y a la función de predicción.

No vamos a comentar mucho de este método pues no es lo que estamos buscando en este TFG, pero queremos mostar una forma de medir qué tan bien el modelo interpretable aproxima al modelo de caja negra es la medida R-cuadrado:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (\hat{y}_*^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2} \quad (\text{A.12})$$

donde $\hat{y}_*^{(i)}$ es la predicción para la i -ésima instancia del modelo interpretable, $y^{(i)}$ es la predicción del modelo de caja negra y \bar{y} es la media de las predicciones del modelo de caja negra. SSE representa la suma de los cuadrados del error y SST la suma de los cuadrados total. La medida R^2 se puede interpretar como el porcentaje de varianza que captura el modelo de sustitución. Si R^2 está cerca de 1 (bajo SSE), entonces el modelo interpretable aproxima muy bien el comportamiento del modelo de caja negra.

Ventajas y desventajas

Ventajas

- Flexibilidad: Se puede usar cualquier modelo interpretable, permitiendo cambiar modelos de sustitución según las necesidades.
- Intuitivo: Fácil de entender e implementar.
- Medida R^2 : Permite medir la calidad de los modelos de sustitución para aproximar predicciones de la caja negra.

Desventajas

- Conclusiones del modelo, no de los datos: Se extraen conclusiones sobre el modelo, ya que nunca ve el resultado real.
- Punto de corte para R^2 : No está claro cuál es el umbral aceptable para la cercanía entre el modelo de sustitución y la caja negra.
- Dependencia del conjunto de datos: La interpretación puede variar para diferentes subconjuntos de datos.
- Modelos interpretables: La elección del modelo interpretable como sustituto tiene sus propias ventajas y desventajas, dependiendo de la perspectiva sobre la interpretabilidad intrínseca.

A.2.5. Shapley values

La idea principal es la siguiente, explicar una predicción a través de la teoría de juegos, asumiendo que cada característica es un jugador en un juego donde la predicción es el pago. Los valores Shapley nos dicen cómo distribuir de manera justa el pago entre las características.

Intuición

Imaginemos que tú y un amigo participáis en un concurso y ganais el primer premio de \$10,000. Tu amigo sugiere una división equitativa, pero tú argumentas que mereces una parte más grande. Para resolver esto, imaginemos que pudierais volver a participar en el concurso por separado.

En la segunda tanda, quedas en segundo lugar y ganas \$7,500, mientras que tu amigo queda en tercer lugar y gana \$5,000. Si ninguno jugara, ganaríais \$0. Aquí es donde entran en juego los valores Shapley. La pregunta es cómo dividir de manera justa el premio, y una forma de hacerlo es calcular la contribución marginal esperada de cada jugador.

Por ejemplo, tú podrías unirte a una coalición de solo tu amigo, llevándolo del tercer lugar al primer lugar y aumentando el premio en \$5,000. También podrías unirte a una coalición sin jugadores y aumentar el premio en \$7,500. Estas son tus contribuciones marginales, y el promedio de estas contribuciones nos da una contribución marginal esperada de \$6,250.

Siguiendo un proceso similar, calculamos la contribución marginal esperada para tu amigo, que resulta ser \$3,750. Al final, tú recibes \$6,250 y tu amigo recibe \$3,750, conocidos como los Shapley values. Estos valores se consideran una forma justa de dividir el premio, y aunque hemos calculado esto para un equipo de 2 jugadores, la teoría de Shapley se puede generalizar para equipos de cualquier

tamaño.

Teoría

Consideramos un conjunto de jugadores de tamaño N y una función v que asigna subconjuntos de jugadores a números reales: $v : 2^N \rightarrow \mathbb{R}$, donde $v(\emptyset) = 0$. Esta función v se denomina función característica y representa la suma total de pagos que los miembros de una coalición S pueden obtener mediante la cooperación.

Función valor

La función característica v determina la interpretación de la contribución de las características. Podemos usar la función $v_x(S)$, que representa la predicción para los valores de características en el conjunto S marginalizados sobre las características que no están incluidas en S :

$$v_x(S) = \int \hat{f}(x_1, \dots, x_n) d\mathbb{P}_{x \notin S} - E_X(\hat{f}(X)) \quad (\text{A.13})$$

Aquí, realizamos múltiples integraciones para cada característica que no está contenida en S . Un ejemplo concreto sería evaluar la predicción para la coalición S que consiste en los valores de características x_1 y x_3 :

$$v_x(S) = v_x(\{x_1, x_3\}) = \int_R \int_R \hat{f}(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2 X_4} - E_X(\hat{f}(X))$$

Shapley demostró que los Shapley Values (ϕ_j) son únicos en su capacidad para satisfacer cuatro axiomas que definen un pago justo: Eficiencia, Simetría, Dummy y Aditividad.

Axiomas de los Shapley Values

- **Eficiencia:** Las contribuciones de las características deben sumar la diferencia entre la predicción para x y el promedio.

$$\sum_{j=1}^n \phi_j = \hat{f}(x) - E_X(\hat{f}(X)) \quad (\text{A.14})$$

- **Simetría:** Las contribuciones de dos valores de características j y k deben ser iguales si contribuyen igualmente a todas las coaliciones posibles.

Si $v(S \cup \{x_j\}) = v(S \cup \{x_k\})$ para todas las $S \subseteq \{x_1, \dots, x_n\} \setminus \{x_j, x_k\}$, entonces $\phi_j = \phi_k$.

- **Jugador Nulo:** Una característica j que no cambia el valor predicho, independientemente de la coalición a la que se añada, debe tener un valor Shapley de 0.

Si $v(S \cup \{x_j\}) = v(S)$ para todas las $S \subseteq \{x_1, \dots, x_n\}$, entonces $\phi_j = 0$.

- **Aditividad:** Para un juego con pagos combinados $v + v^+$, los Shapley Values respectivos son la suma de los valores individuales $\phi_j + \phi_j^+$.

Valores Shapley

Se definen mediante la función característica v de los jugadores en S :

$$\phi_j(v) = \sum_{S \subseteq \{x_1, \dots, x_n\} \setminus \{x_j\}} \frac{j!(n-|S|-1)!}{n!} (v(S \cup \{x_j\}) - v(S)) \quad (\text{A.15})$$

Donde S es un subconjunto de las características utilizadas en el modelo, x es el vector de valores de características de la instancia a explicar, y n es el número de características.

Teorema sobre Shapley Values

Shapley demuestró la existencia y unicidad de los valores Shapley al satisfacer los axiomas establecidos como definición de un pago justo.

Estimación

Para calcular el Valor Shapley exacto, todas las posibles coaliciones de valores de características deben evaluarse con y sin la característica j . Sin embargo, esta solución exacta se vuelve problemática con un gran número de características, ya que el número de coaliciones posibles aumenta exponencialmente. Una posible aproximación con muestreo Monte Carlo es la siguiente:

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^M (\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)) \quad (\text{A.16})$$

Donde $\hat{f}(x_{+j}^m)$ es la predicción para x , pero con un número aleatorio de valores de características reemplazados por valores de características de un punto de datos aleatorio z , excepto por el valor respectivo de la característica j . El vector x_{+j}^m el valor de la característica j se toma de x y el vector x_{-j}^m no.

Estimación aproximada de Shapley para un valor de característica individual:

- Para cada $m = 1, \dots, M$:
 - Seleccionar una instancia aleatoria z de la matriz de datos X
 - Elegir una permutación aleatoria de los valores de características
 - Ordenar la instancia x : $x_o = (x_1; \dots; x_j; \dots; x_n)$
 - Ordenar la instancia z : $z_o = (z_1; \dots; z_j; \dots; z_n)$
 - Construir dos nuevas instancias:
 - ◊ $x_{+j}^m = (x_1; \dots; x_{j-1}; x_j; z_{j+1}; \dots; z_n)$
 - ◊ $x_{-j}^m = (x_1; \dots; x_{j-1}; z_j; z_{j+1}; \dots; z_n)$
 - Calcular la diferencia Shapley: $\phi_j^m = \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m)$
- Calcular el Valor Shapley como el promedio: $\hat{\phi}_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_j^m$

COMPLEMENTOS EXPERIMENTOS

En este capítulo del anexo se muestran algunas ideas que pueden aclarar dudas o que permiten profundizar en los experimentos con el fin de potenciar los resultados obtenidos. Al igual que el resto de los anexos, no es necesario para comprender el cuerpo principal del TFG, pero puede resultar interesante.

B.1. Datos tabulares

B.1.1. AutoML

German Risk

Model	AUC	LogLoss	AUCPR	Error	RMSE	MSE
StackedEnsemble	0.757113	0.520414	0.876334	0.41741	0.417903	0.174643
GLM	0.750525	0.528015	0.862553	0.471849	0.419759	0.176198
GBM	0.740081	0.535874	0.843844	0.414478	0.422193	0.178247
DeepLearning	0.734128	0.556461	0.854328	0.422739	0.428824	0.18389

Tabla B.1: Tabla de resultados con AutoML GermanRisk (1K) (9 características)

En la tabla B.1 se muestran los resultados de los mejores representantes de los diferentes tipos de modelos obtenidos con AutoML. En las tablas B.2, B.3 y B.4 se muestran los resultados propios obtenidos tras aplicar el preprocesamiento, entrenamiento de modelos con optimización de hiperparámetros y umbrales. Podemos ver que los resultados son buenos, dandonos la confianza para aplicar la parte de interpretabilidad que nos interesa.

Santander Customer Satisfaction

En la tabla B.5 se muestran los resultados de los mejores representantes de los diferentes tipos de modelos obtenidos con AutoML. En las tablas B.6 B.7 y B.8 se muestran los resultados propios obtenidos tras aplicar el preprocesamiento, entrenamiento de modelos con optimización de hiperparámetros

Umbral: 0.56				
	precision	recall	f1-score	support
Clase 0	0.72	0.96	0.82	700
Clase 1	0.58	0.13	0.21	300
Exactitud	0.71			

Tabla B.2: Medidas de precisión de regresión logística después de optimizar los umbrales en German Risk.

Umbral: 0.38				
	precision	recall	f1-score	support
Clase 0	0.84	0.89	0.86	700
Clase 1	0.70	0.60	0.65	300
Exactitud	0.80			

Tabla B.3: Medidas de precisión de random forest después de optimizar los umbrales en German Risk.

Umbral: 0.42				
	precision	recall	f1-score	support
Clase 0	0.84	0.85	0.85	700
Clase 1	0.65	0.63	0.64	300
Exactitud	0.79			

Tabla B.4: Medidas de precisión de red neuronal después de optimizar los umbrales en German Risk.

Model	AUC	LogLoss	AUCPR	Error	RMSE	MSE
StackedEnsemble	0.838	0.134	0.189	0.302	0.186	0.034
GBM	0.834	0.136	0.180	0.299	0.186	0.035
GLM	0.796	0.147	0.140	0.309	0.190	0.036
DeepLearning	0.771	0.154	0.138	0.333	0.190	0.036

Tabla B.5: Tabla de resultados con AutoML Santander Customer Satisfaction (76K) (310 características)

y umbrales. Podemos ver que los resultados son buenos, dándonos la confianza para aplicar la parte de interpretabilidad que nos interesa.

Umbral: 0.57				
	precision	recall	f1-score	support
Clase 0	0.81	0.86	0.83	14629
Clase 1	0.85	0.80	0.82	14576
Exactitud	0.83			

Tabla B.6: Medidas de precisión de regresión logística después de optimizar los umbrales en Santander Customer Satisfaction.

Umbral: 0.6				
	precision	recall	f1-score	support
Clase 0	0.80	0.89	0.84	14629
Clase 1	0.87	0.77	0.82	14576
Exactitud	0.83			

Tabla B.7: Medidas de precisión de random forest después de optimizar los umbrales en Santander Customer Satisfaction.

Umbral: 0.4				
	precision	recall	f1-score	support
Clase 0	0.88	0.87	0.88	14629
Clase 1	0.87	0.88	0.88	14576
Exactitud	0.88			

Tabla B.8: Medidas de precisión de red neuronal después de optimizar los umbrales en Santander Customer Satisfaction.

B.1.2. Consistencia de PDP y ALE

Cálculos funciones teóricas

Definimos la función de probabilidad que usamos para este problema como $f(x_1, x_2) = -2x_1^2 - 2x_2^2 + 2x_1 + 2x_2$ y aplicamos las fórmulas de PDP y ALE definidas en 2.5 y 2.7.

Para PDP y la variable x_1 , usando que la distribución la conocemos y es uniforme, tenemos lo siguiente:

$$f_{1;PDP}(x_1) = E_{X_2} [f(x_1, X_2)] = \int_0^1 f(x_1, X_2) d\mathbb{P}(X_2)$$

Es decir, tenemos:

$$f_{1;PDP}(x_1) = \int_0^1 (-2x_1^2 - 2x_2^2 + 2x_1 + 2x_2) dx_2 = -2x_1^2 - \frac{2}{3} + 2x_1 + 1 \quad (\text{B.1})$$

$$\begin{aligned} f_{1;\text{ALE}}(x_1) &= \int_{z_{0,1}}^{x_1} \mathbb{E}_{X_2|X_1} [f^1(X_1, X_2) | X_1 = z_1] dz_1 - \text{constante} \\ &= \int_{z_{0,1}}^{x_1} \int_{x_2} f^1(z_1, x_2) \mathbb{P}(x_2 | z_1) dx_2 dz_1 - \text{constante} \end{aligned}$$

Siendo:

$$f^1(x_1, x_2) = \frac{\partial f(x_1, x_2)}{\partial x_1} = \frac{\partial (-2x_1^2 - 2x_2^2 + 2x_1 + 2x_2)}{\partial x_1} = -4x_1 + 2$$

Es decir, obtenemos gracias a conocer la función teórica:

$$f_{1;\text{ALE}}(x_1) = \int_{x_1-\epsilon}^{x_1+\epsilon} (-4x_1 + 2) = -2x_1^2 + 2x_1 + C \quad (\text{B.2})$$

Esta constante es para centrar el efecto de ALE y puede comprobarse que es $\frac{1}{2}$.

Resultados de regresión

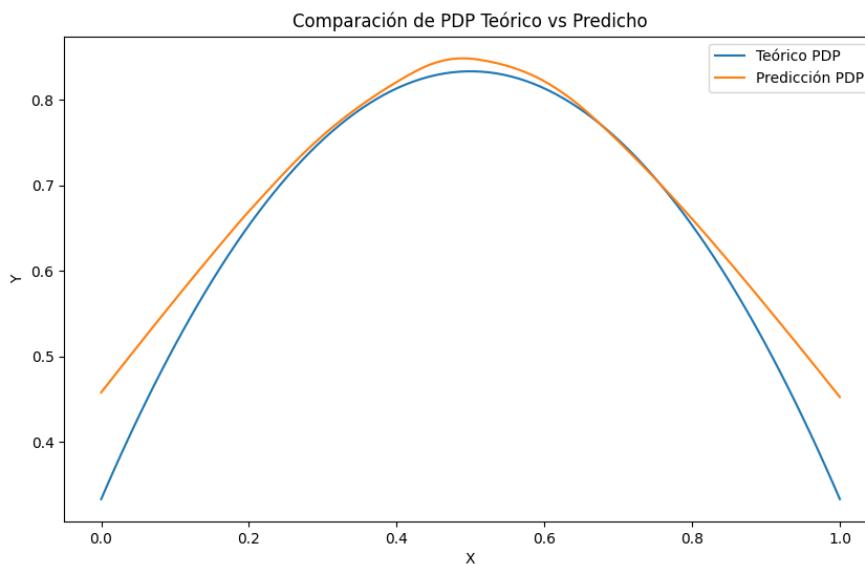


Figura B.1: Comparación función teórica de PDP y real para un modelo de regresión sobre la función de probabilidad.

Los resultados de la Figura B.1 muestran una alta precisión con un bajo error cuadrático medio (MSE) de 0.0018 para PDP. Esto indica que el modelo de regresión es capaz de capturar de manera efectiva la relación subyacente en los datos.

De manera similar, la Figura B.2 muestra la comparación de ALE con un MSE de 0.0073, reafirmando la capacidad del modelo de regresión para representar adecuadamente la función de probabilidad.

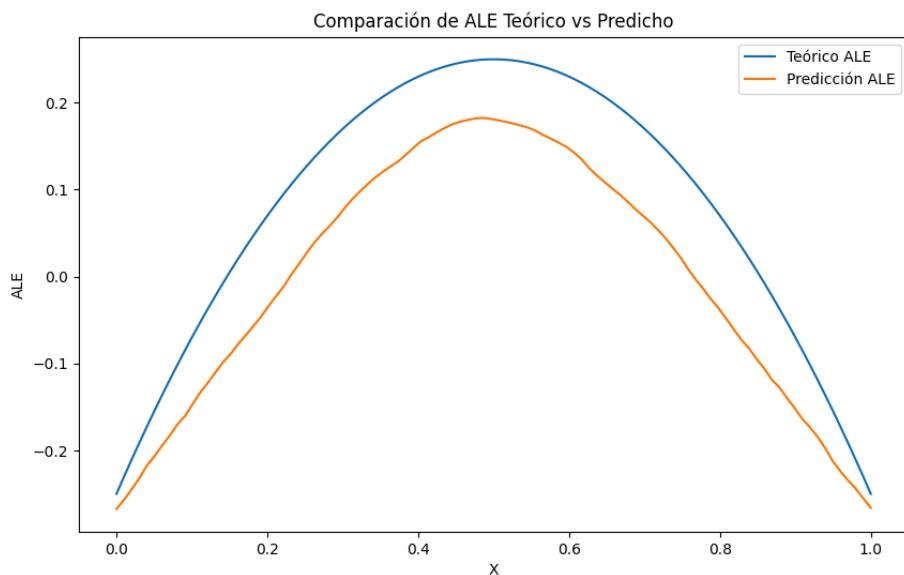


Figura B.2: Comparación función teórica de ALE y real para un modelo de regresión sobre la función de probabilidad.

Errores cuadráticos medios (MSE):

- MSE para PDP en regresión: 0.0018240328036249458
- MSE para ALE en regresión: 0.007308896594081433

Resultados precisión red neuronal

	precision	recall	f1-score	support
Clase 0	0.80	0.69	0.74	259
Clase 1	0.90	0.94	0.92	741
Exactitud	0.88			

Tabla B.9: Medidas de precisión de la red neuronal.

Como se muestra en la Tabla B.9, la red neuronal muestra una precisión y recall bastante bueno para ambas clases (especialmente para la clase 1), y una exactitud global de 0.88. Esto demuestra que el modelo es competente en clasificar correctamente las instancias en ambas clases.

Resultados de clasificación para dos variables

La Figura B.3 muestra la comparación del PDP teórico y real para dos variables. Se observa que, aunque ambos muestran una relación general similar, hay variaciones debido a la naturaleza discreta del problema de clasificación. Esto destaca la complejidad adicional en la representación precisa de la relación entre variables en un contexto de clasificación.

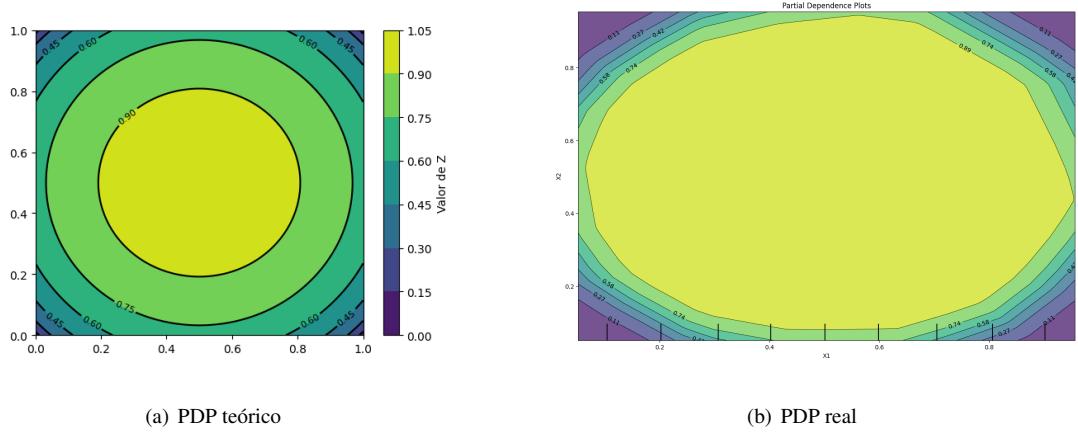


Figura B.3: Comparación de la función teórica de PDP y la obtenida a partir del modelo de clasificación para dos variables.

B.1.3. Interpretabilidad de modelos en German Risk

En la figura B.4 se muestra la representación de la distribución de las probabilidades obtenidas por el modelo de regresión logística a los datos de la base de datos. En la tabla B.10 se muestra la precisión del modelo sobre la base de datos completa, no se muestra sólo sobre el conjunto de test por tener tan pocos elementos en él.

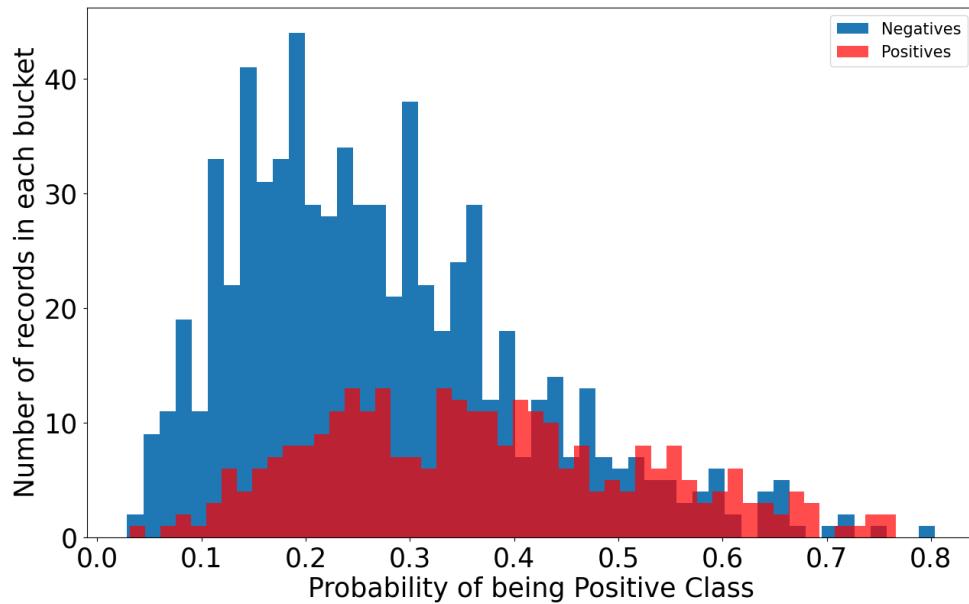


Figura B.4: Probabilidades de los datos tras aplicar el modelo de regresión logística.

En la figura B.5 se muestra la representación de la distribución de las probabilidades obtenidas por el modelo de random forest. En la tabla B.11 se muestra la precisión del modelo sobre la base de datos.

Umbral: 0.56				
	precision	recall	f1-score	support
Clase 0	0.72	0.96	0.82	700
Clase 1	0.58	0.13	0.21	300
Exactitud	0.71			

Tabla B.10: Medidas de precisión de regresión logística después de optimizar los umbrales.

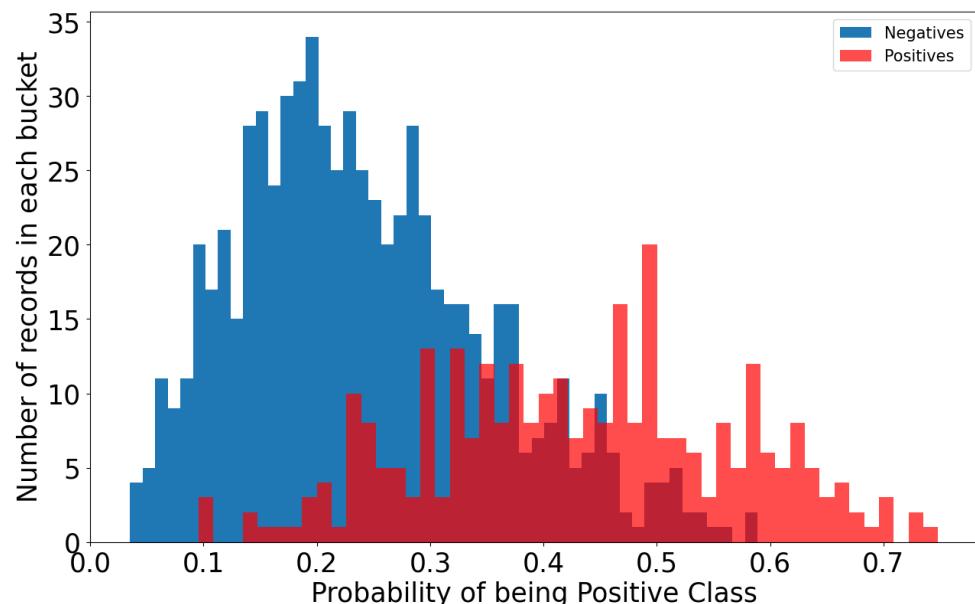


Figura B.5: Probabilidades de los datos tras aplicar el modelo de random forest.

Umbral: 0.38				
	precision	recall	f1-score	support
Clase 0	0.84	0.89	0.86	700
Clase 1	0.70	0.60	0.65	300
Exactitud	0.80			

Tabla B.11: Medidas de precisión de random forest después de optimizar los umbrales.

En la figura B.6 se muestra la representación de la distribución de las probabilidades obtenidas por el modelo de red neuronal. En la tabla B.12 se muestra la precisión del modelo sobre la base de datos.

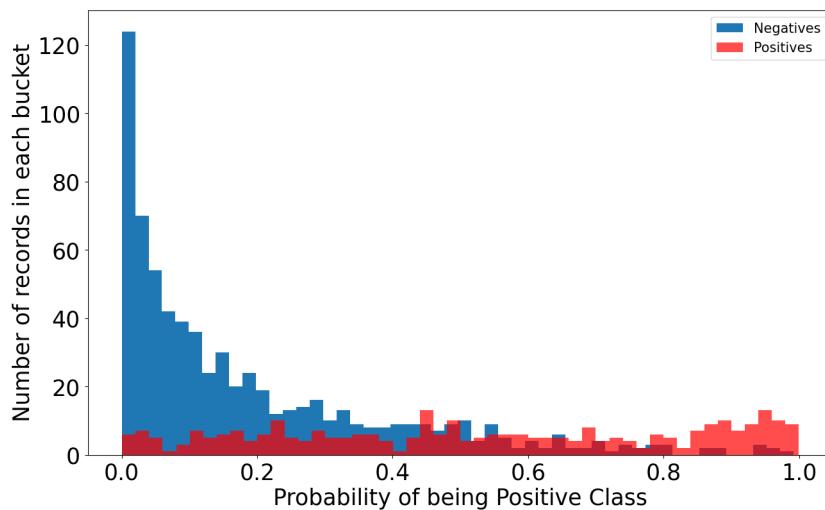


Figura B.6: Probabilidades de los datos tras aplicar el modelo de red neuronal.

Umbral: 0.42				
	precision	recall	f1-score	support
Clase 0	0.84	0.85	0.85	700
Clase 1	0.65	0.63	0.64	300
Exactitud	0.79			

Tabla B.12: Medidas de precisión de red neuronal después de optimizar los umbrales.

B.1.4. Robustez de FI y PFI con random forest

Feature Importance basado en la impureza en un random forest tiende a dar más importancia a las variables numéricas por varias razones clave:

- Alta cardinalidad: Las variables numéricas suelen tener una alta cardinalidad, lo que significa que pueden tomar una gran cantidad de valores distintos. Esto le da al modelo más oportunidades de encontrar divisiones que reduzcan la impureza en los nodos del árbol, lo que puede inflar artificialmente la importancia de estas variables.
- Capacidad de división precisa: Los árboles de decisión pueden crear divisiones muy precisas en variables numéricas, ajustándose más exactamente a los datos de entrenamiento. Esto puede resultar en una mayor reducción de la impureza y, por ende, en una mayor importancia asignada a estas variables.
- Sesgo hacia características informativas: Las variables numéricas, al proporcionar más opciones de partición, a menudo se consideran más informativas durante el entrenamiento del modelo, aunque esto no siempre refleje su verdadera capacidad predictiva. Este sesgo puede causar que se les otorgue mayor importancia.
- Uso de estadísticas del conjunto de entrenamiento: La importancia basada en la impureza se calcula utilizando las estadísticas del conjunto de entrenamiento. Esto puede llevar a que se sobrevaloren las variables numéricas, especialmente si el modelo tiene la capacidad de sobreajustar los datos de entrenamiento.

B.1.5. Análisis de sesgo en función del género utilizando SHAP

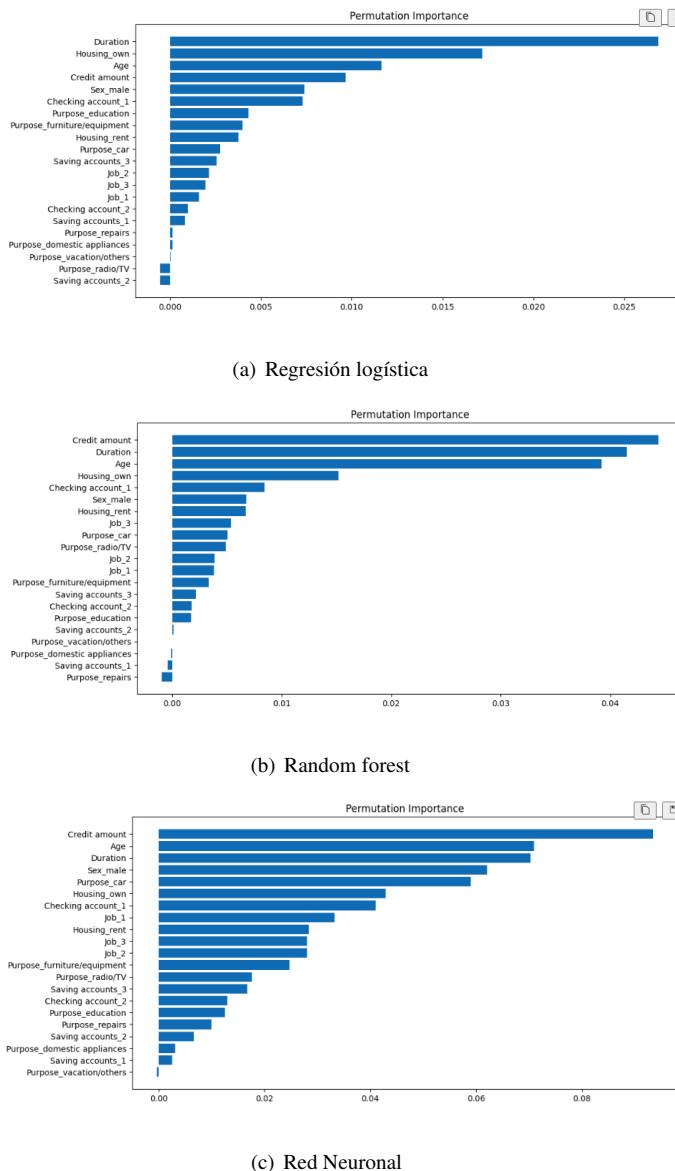


Figura B.7: Permutación de la Importancia de las Características (PFI) para diferentes modelos utilizando la variable del género.

En las figura B.7 se muestran los valores obtenidos con PFI para los tres modelos utilizados usando la variable del género en la base de datos. Como se puede observar todos los modelos utilizan esta variable para sesgar y clasificar, siendo al menos la quinta variable más importante del modelo.

A continuación se muestran en las figuras B.8 y B.9 la descomposición de la diferencia en la paridad geográfica completa para los modelos comentados en el experimento.

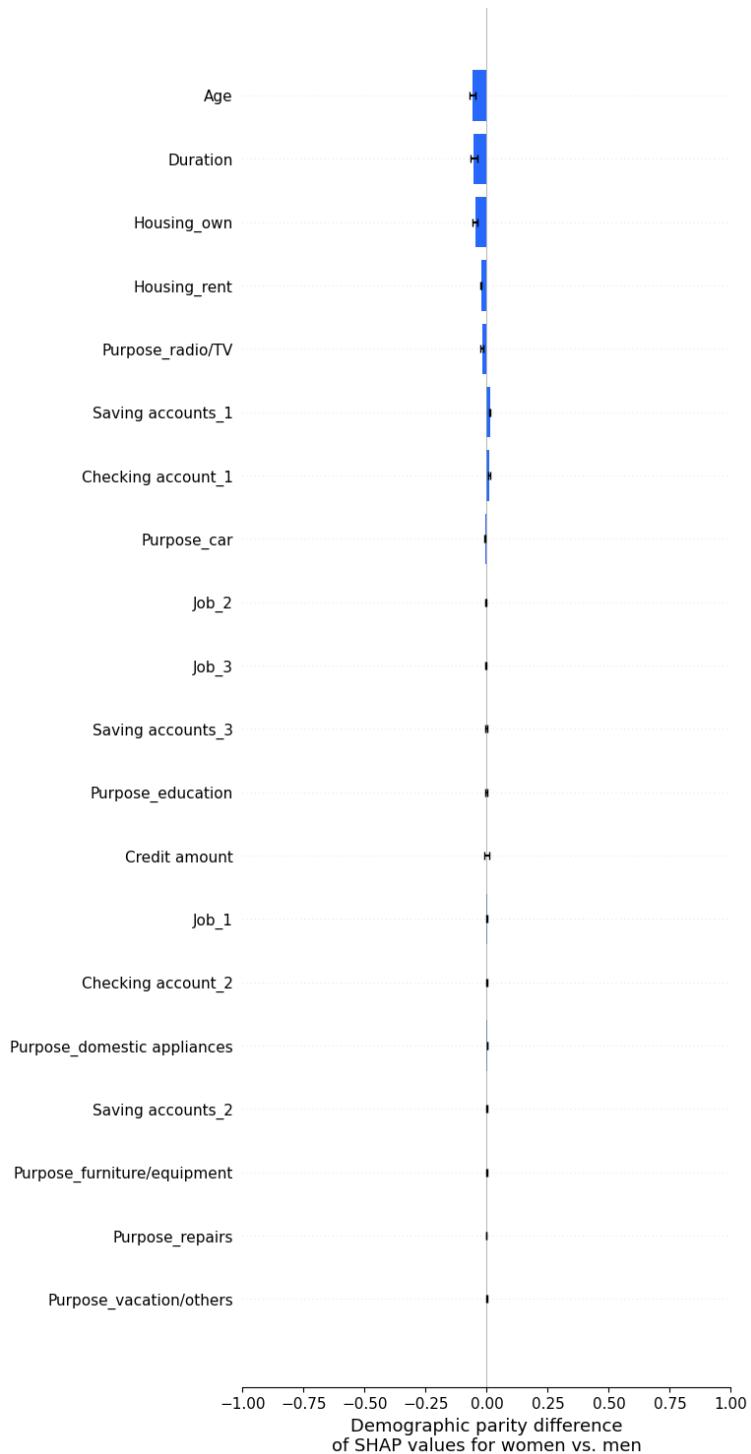


Figura B.8: Diferencia de paridad demográfica descompuesta por características para random forest.

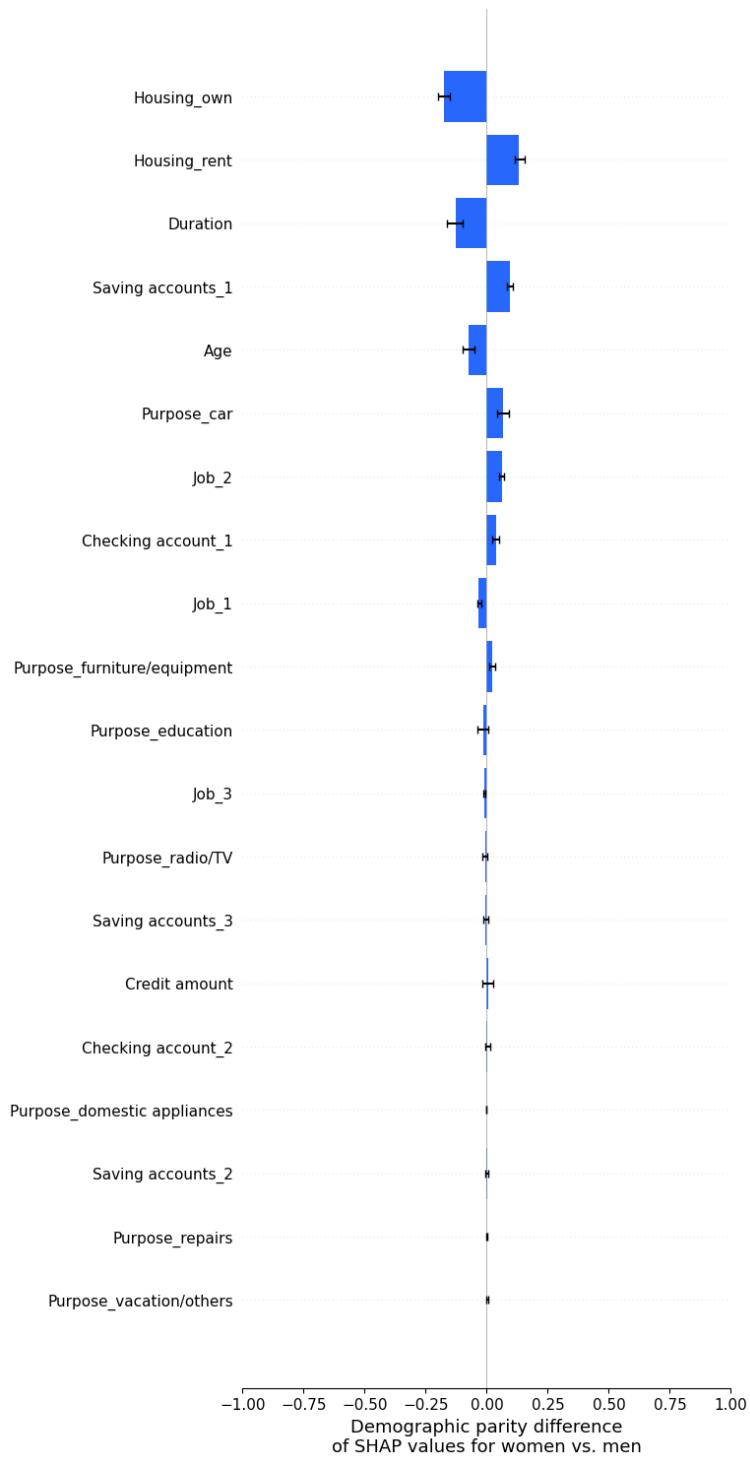


Figura B.9: Diferencia de paridad demográfica descompuesta por características para red neuronal.

B.1.6. Satisfacción de clientes del banco Santander

Este anexo complementa la sección principal proporcionando detalles sobre el análisis de la importancia de las características. A continuación, se muestran las importancias de todas las características para los tres modelos: regresión logística B.10, random forest B.11, y red neuronal B.12.

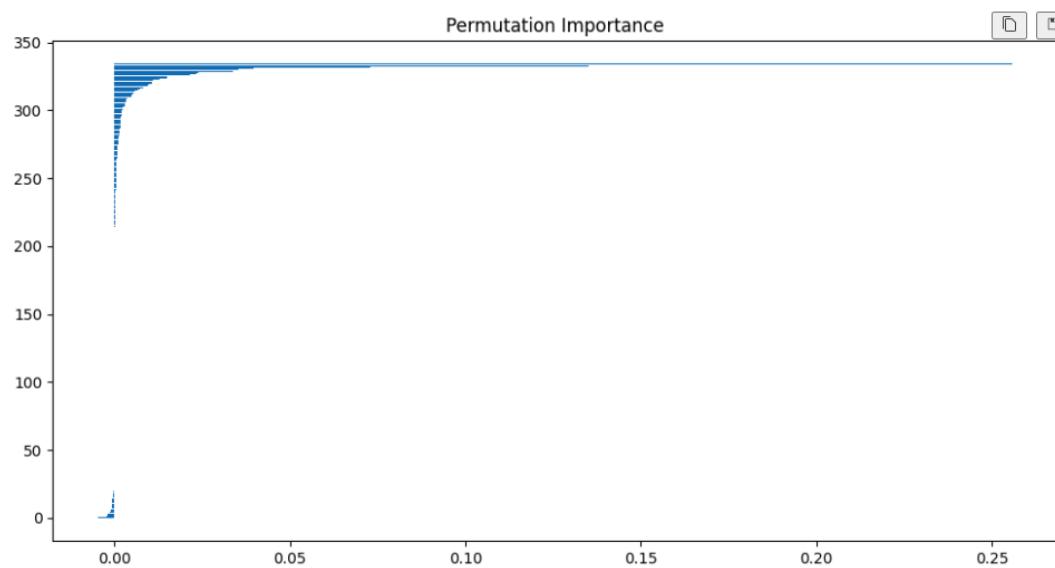


Figura B.10: Importancia de las características por permutación para el modelo de regresión logística.

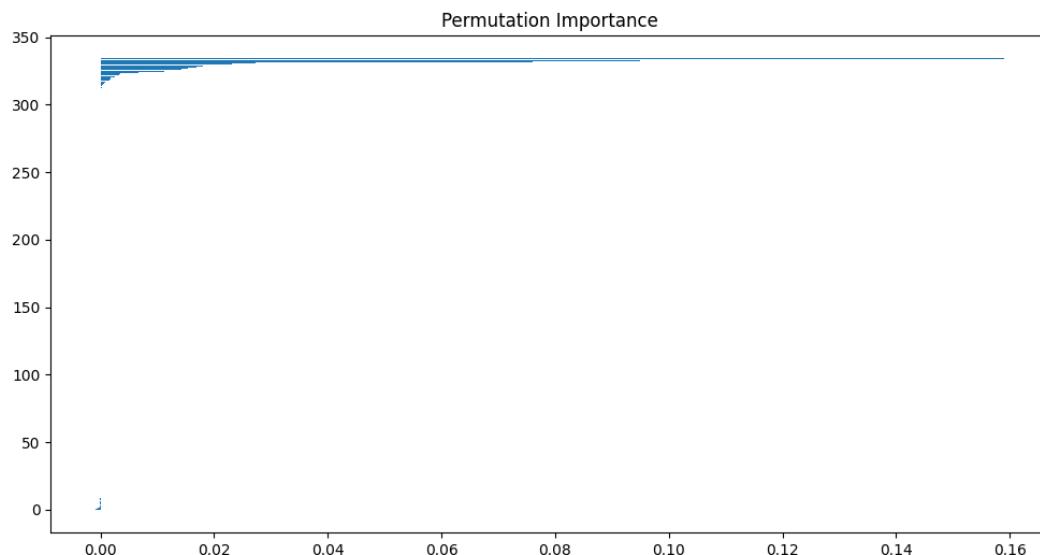


Figura B.11: Importancia de las características por permutación para el modelo de random forest.

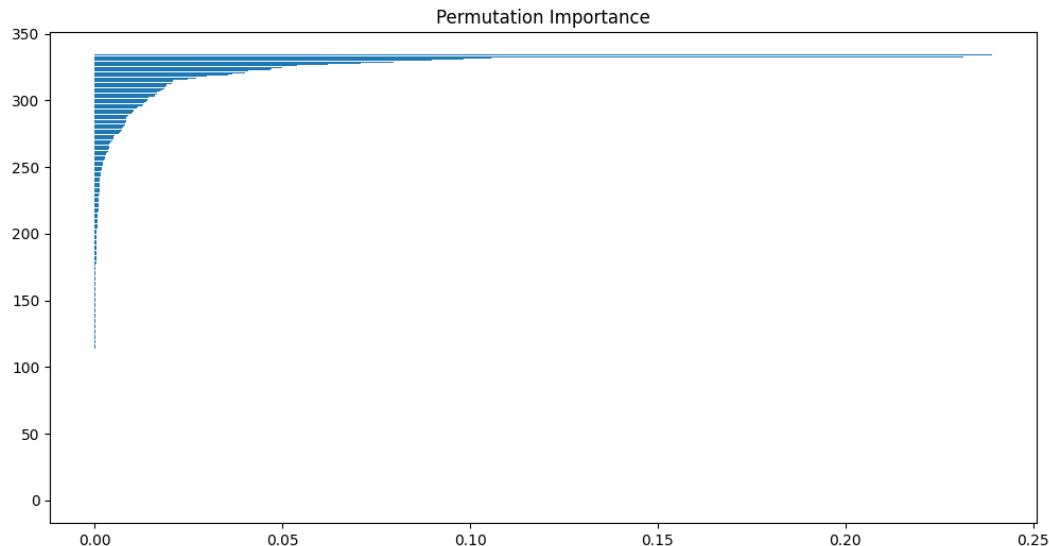


Figura B.12: Importancia de las características por permutación para el modelo de red neuronal.

Importancia acumulada de las características

La tabla B.13 muestra la suma de las importancias de las diez características más importantes y la suma del resto para cada modelo.

Modelo	Suma de las principales características	Suma del resto
Regresión logística (10)	0.6562	0.1971
Random forest (10)	0.4563	0.0226
Red neuronal (15)	1.0787	1.0068

Tabla B.13: Suma de las importancias de las principales características y del resto para cada modelo.

Importancia acumulada para la red neuronal

En el caso del modelo de red neuronal, se analiza el número de características necesarias para que la suma de las importancias sea menor a varios umbrales (0.5, 0.3, 0.2, 0.1). Se muestran los resultados en la tabla B.14.

Este análisis detalla la complejidad de interpretar modelos complejos como las redes neuronales, y como el número de variables es un elemento que añade complejidad a la interpretabilidad y que no hemos tratado específicamente en este TFG.

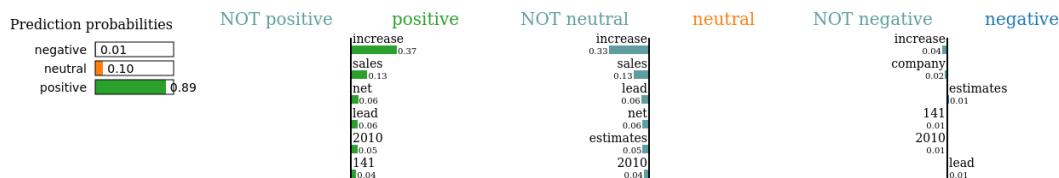
B.2. Texto

Umbral	Número de características
0.5	29
0.3	43
0.2	55
0.1	75

Tabla B.14: Número de características necesarias para que la suma de las importancias sea menor a los umbrales especificados en el modelo de red neuronal.

B.2.1. Clasificador de análisis de sentimientos

A continuación se presentan las interpretaciones de LIME y SHAP aplicadas a diversos ejemplos de textos clasificados por el modelo de análisis de sentimientos. Estos análisis han permitido detectar y comprender mejor el funcionamiento y las limitaciones del modelo.



Text with highlighted words

The company also estimates the already carried out investments to **lead** to an **increase** in its **net sales** for **2010** from 2009 when they reached EUR 141.7 million

Figura B.13: Interpretación LIME previa a los cambios en el preprocesamiento.

En la figura B.13, se muestra la interpretación de LIME antes de realizar cambios en el preprocesamiento. Inicialmente, se observó que números y referencias monetarias influían indebidamente en la clasificación de sentimientos, lo cual se corrigió posteriormente en el preprocesamiento.



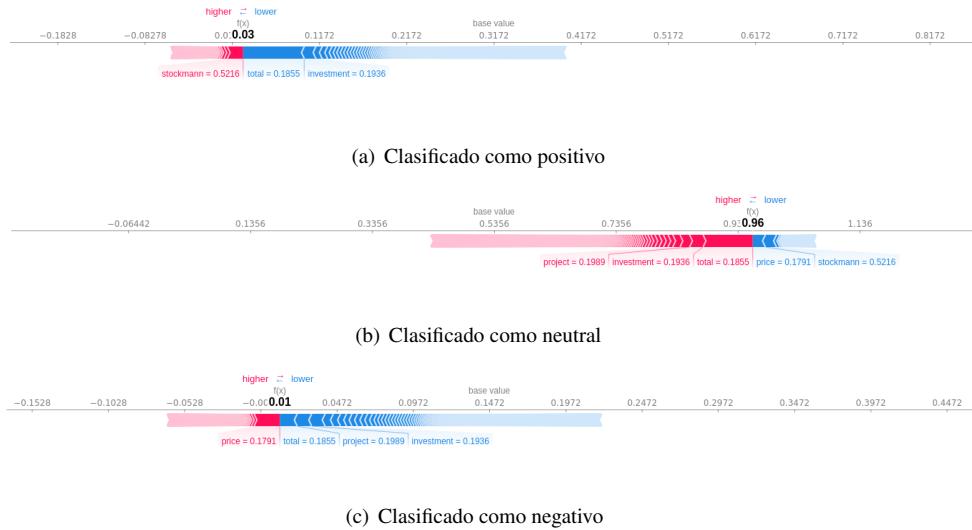
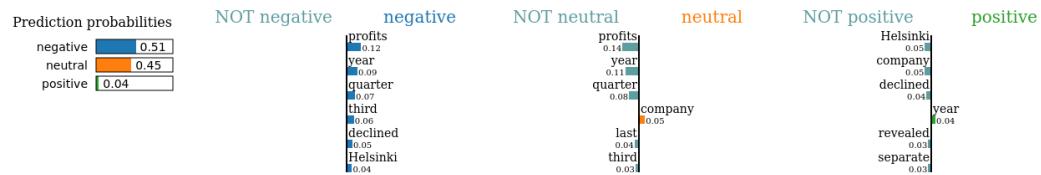
Text with highlighted words

The Stockmann **department** store will have a **total** floor space of over 8,000 square metres and Stockmann's investment in the **project** will have a **price** tag of about EUR 12 million

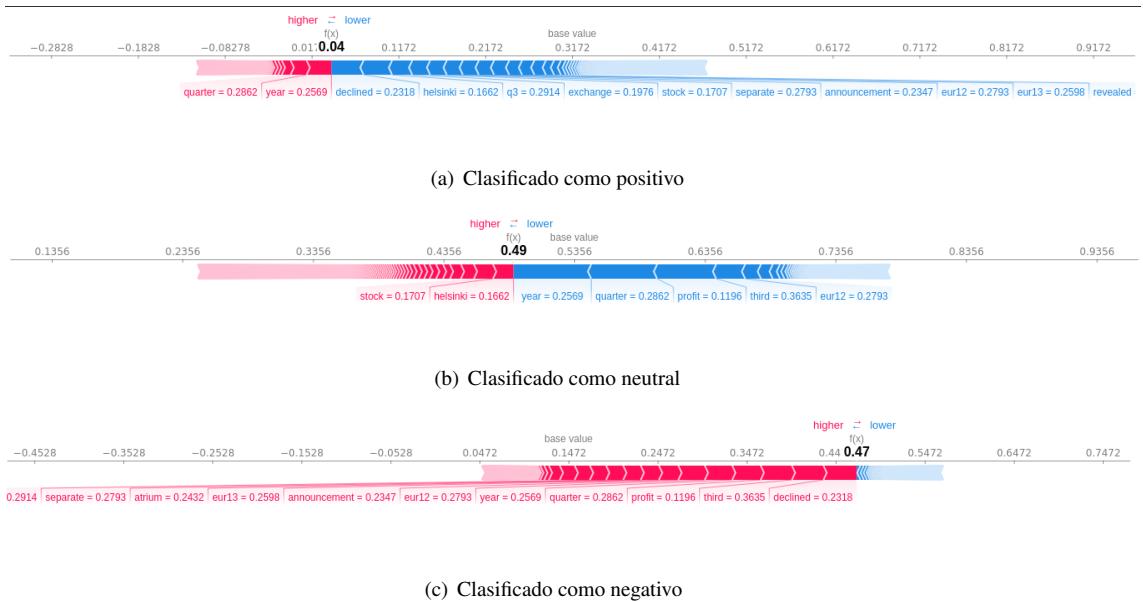
Figura B.14: Interpretación LIME para un texto clasificado como neutral.

La figura B.14 y la figura B.15 presentan interpretaciones de LIME y SHAP para un texto clasificado como neutral. Estas herramientas muestran cómo el modelo parte de una alta probabilidad inicial de neutralidad y ajusta la clasificación según las palabras presentes, incrementando o disminuyendo las probabilidades de las otras clases.

En las figuras B.16 y B.17, se analizan textos clasificados como negativos. Se observa que ciertas

**Figura B.15:** Interpretación SHAP para un texto clasificado como neutral.**Text with highlighted words**

In a separate announcement to the Helsinki stock exchange , Atria revealed that the company 's **third quarter profits declined** from EUR13 .9 m in the **third quarter** of last **year** to EUR12 .7 m in this **year** 's Q3 .

Figura B.16: Interpretación LIME para un texto clasificado como negativo.**Figura B.17:** Interpretación SHAP para un texto clasificado como negativo.

tas palabras clave, como *declined*, tienen un impacto significativo en la clasificación. Sin embargo, la interpretación también revela casos de sobreajuste, como se muestra con la palabra *Helsinki*, que incrementa la probabilidad neutral debido a su distribución en el conjunto de datos.

Sentimiento	Conteo
Neutral	107
Positivo	27
Negativo	25

Tabla B.15: Conteo de '*Helsinki*' por clase

La tabla B.15 muestra el conteo de la palabra *Helsinki* por clasificación. Este análisis evidenció un problema de sobreajuste en el modelo, ya que *Helsinki* aparece predominantemente en contextos neutrales, lo que afectó la interpretación de algunos resultados.

En resumen, aunque el modelo no es perfecto y se limita por el tamaño reducido del conjunto de datos (unas 5,000 entradas), el uso de LIME y SHAP ha sido esencial para entender su funcionamiento, identificar errores en el preprocesamiento y realizar mejoras. La interpretabilidad también permitió detectar problemas de sobreajuste, facilitando así un análisis más profundo y preciso.



Universidad Autónoma
de Madrid