COMILLAS
UNIVERSIDAD PONTIFICIA

Facultad de Ciencias Económicas y Empresariales

Análisis de perfiles de riesgo crediticio utilizando segmentación supervisada con valores SHAP en solicitantes de crédito

Autor: Sofia González-Montagut Tanure
Director: Jenny Alexandra Cifuentes Quintero

MADRID| Abril 2026

Resumen

La gestión del riesgo crediticio constituye una de las funciones fundamentales del sistema financiero, ya que una estimación rigurosa de la probabilidad de incumplimiento contribuye a preservar la solvencia de las entidades financieras, optimizar el capital regulatorio y favorecer una asignación más eficiente del crédito. La adopción de técnicas de aprendizaje automático ha mejorado sustancialmente la capacidad predictiva de los modelos de *credit scoring*, pero también ha introducido una disyuntiva entre precisión e interpretabilidad que limita su aceptación en un entorno regulatorio que exige transparencia y trazabilidad en la toma de decisiones automatizadas.

En este contexto, este Trabajo de Fin de Grado propone una metodología de segmentación supervisada basada en valores SHAP para analizar perfiles de riesgo crediticio. El enfoque combina diversas técnicas en una metodología secuencial. En primer lugar, se entrena un modelo predictivo LightGBM sobre el conjunto de datos *German Credit Data* (1.000 solicitantes). A continuación, se calculan los valores SHAP para construir un espacio explicativo en el que cada observación queda descrita por la contribución de sus variables a la predicción. Posteriormente, se aplica UMAP para reducir la dimensionalidad y DBSCAN para identificar clusters densos de solicitantes con patrones explicativos similares. Finalmente, los grupos resultantes se caracterizan mediante reglas interpretables extraídas con *SkopeRules*.

El modelo predictivo alcanzó un ROC-AUC de 0,686 y una exactitud del 74 % , suficiente para generar explicaciones SHAP estables. La segmentación en el espacio explicativo identificó tres clusters densos y bien separados (coeficiente de Silhouette de 0,802; índice de Davies-Bouldin de 0,307), caracterizados por la combinación de la duración del crédito y el nivel de ahorro: (i) solicitantes de corto plazo con bajo ahorro (20,3 % de incumplimiento), (ii) solicitantes vulnerables de largo plazo (46,4 %) y (iii) solicitantes con mayor estabilidad de ahorro (22,8 %). Los resultados muestran que la metodología propuesta permite trasladar las predicciones individuales de riesgo a perfiles homogéneos descritos mediante reglas observables, integrando capacidad predictiva, interpretabilidad y segmentación en un marco analítico coherente con las exigencias de transparencia del sector financiero.

Palabras clave: *riesgo crediticio*, *clustering supervisado*, *valores SHAP*, *Inteligencia Artificial Explicable (XAI)*, *LightGBM*.

Abstract

Credit risk management is one of the core functions of the financial system, as a rigorous estimation of the probability of default helps preserve the solvency of financial institutions, optimise regulatory capital and promote a more efficient allocation of credit. The adoption of machine learning techniques has substantially improved the predictive capacity of credit scoring models, but it has also introduced a trade-off between accuracy and interpretability that limits their acceptance in a regulatory environment that demands transparency and traceability in automated decision-making.

Against this background, this Bachelor's Thesis proposes a supervised segmentation methodology based on SHAP values to analyse credit risk profiles. The approach combines several techniques in a sequential methodology. First, a LightGBM predictive model is trained on the German Credit Data set (1,000 applicants). Next, SHAP values are computed to build an explanatory space in which each observation is described by the contribution of its variables to the prediction. UMAP is then applied to reduce dimensionality, and DBSCAN is used to identify dense clusters of applicants with similar explanatory patterns. Finally, the resulting groups are characterised through interpretable rules extracted with SkopeRules.

The predictive model achieved a ROC-AUC of 0.686 and an accuracy of 74 % sufficient to generate stable SHAP explanations. The segmentation in the explanatory space identified three dense and well-separated clusters (Silhouette coefficient of 0.802; Davies-Bouldin index of 0.307), characterised by the combination of loan duration and savings level: (i) short-term applicants with low savings (20.3 % default rate), (ii) vulnerable long-term applicants (46.4 %) and (iii) applicants with greater savings stability (22.8 %). The results show that the proposed methodology successfully translates individual risk predictions into homogeneous profiles described by observable rules, integrating predictive performance, interpretability and segmentation within a coherent analytical framework aligned with the transparency requirements of the financial sector.

Keywords: *credit risk*, *supervised clustering*, *SHAP values*, *Explainable Artificial Intelligence (XAI)*, *LightGBM*.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

AUC	Area Under the ROC Curve
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNN	Deep Neural Network
EAD	Exposición al Momento del Incumplimiento
EFB	Exclusive Feature Bundling
GBDT	Gradient Boosting Decision Tree
GOSS	Gradient-based One-Side Sampling
IRB	Internal Rating-Based
LGD	Pérdida Dado el Incumplimiento
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-Agnostic Explanations
ML	Machine Learning
P2P	Peer-to-Peer
PCA	Análisis de Componentes Principales
PD	Probabilidad de Incumplimiento
ROC-AUC	Area Under the Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
TFG	Trabajo de Fin de Grado
UMAP	Uniform Manifold Approximation and Projection
XAI	Inteligencia Artificial Explicable
XGBoost	eXtreme Gradient Boosting

Capítulo 1

Introducción

En las últimas décadas, la expansión del crédito ha adquirido un papel destacado en la dinámica del crecimiento económico mundial (García-Escribano y Han, 2015). El acceso a financiación permite a los hogares adquirir bienes duraderos, a las empresas acometer inversiones productivas y a los gobiernos sostener políticas orientadas al desarrollo económico y social. En este marco, la gestión del riesgo crediticio ocupa una posición central en el funcionamiento del sistema financiero. Una evaluación adecuada del riesgo protege la solidez de las entidades frente a pérdidas potenciales y, al mismo tiempo, incide sobre la economía real al condicionar el acceso al crédito y el ritmo de la actividad económica. Cuando los mecanismos de evaluación son rigurosos y transparentes, se fortalece la confianza de inversores y depositantes, se favorece la inclusión financiera y se facilita una asignación más eficiente de recursos hacia proyectos productivos (Alonso, Durán, García-Olmedo, y Quesada, 2024).

En España, la importancia de esta gestión se observa con claridad en la evolución reciente de la morosidad bancaria, sintetizada en la Figura 1.1. Tras el inicio de la crisis financiera de 2008, la tasa de morosidad experimentó un incremento pronunciado que alcanzó su máximo en 2013, con un valor del 13,6 %. Este comportamiento reflejó el deterioro simultáneo del mercado laboral, del sector inmobiliario y de la actividad empresarial, y generó tensiones significativas en los balances bancarios, además de procesos de reestructuración y recapitalización en varias entidades. A partir de 2014, la morosidad inició una senda descendente, en línea con la recuperación económica, hasta estabilizarse en torno al 3 %-4 % durante la segunda mitad de la década. No obstante, en 2020 se produjo un repunte hasta el 7,4 %, asociado al impacto económico de la pandemia de la COVID-19, seguido de una nueva reducción que situó la tasa en torno al 3,3 % en 2024 (Banco de España, 2024; Funcas, 2020).

Esta evolución evidencia la estrecha relación entre el ciclo económico y la calidad del crédito, así como la necesidad de contar con sistemas de evaluación del riesgo capaces de anticipar el incumplimiento. Una estimación imprecisa de la probabilidad de impago puede intensificar los efectos de cada fase del ciclo. La sobrestimación del riesgo tiende a restringir el acceso a la financiación y a limitar la inversión, mientras que su subestimación incre-

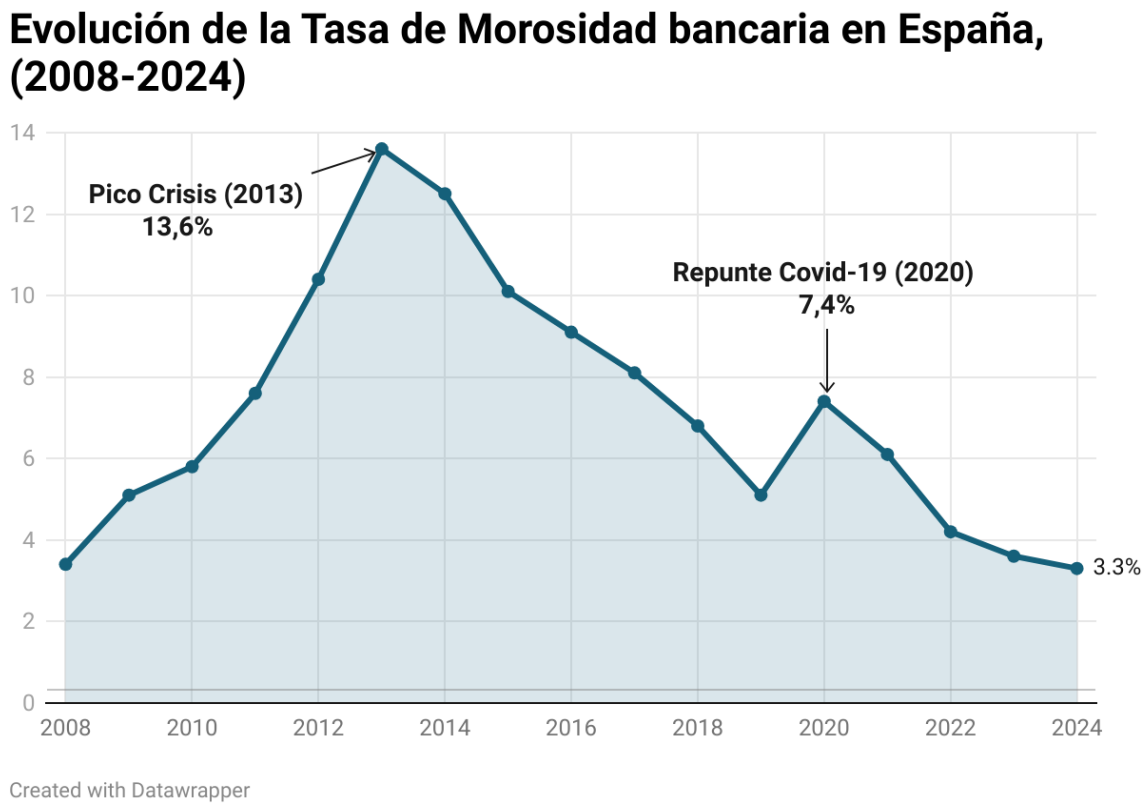

Figura 1.1: Evolución de la Tasa de Morosidad Bancaria en España (2008-2024). Elaboración Propia con datos de (Banco de España, 2024; Funcas, 2020)

menta la exposición a impagos y deteriora la solvencia del sistema bancario (Borio, Furfine, Lowe, et al., 2001, pp. 3–5). En este marco, una estimación rigurosa de la Probabilidad de Incumplimiento (PD, *Probability of Default*) adquiere relevancia estratégica para las entidades financieras, al permitirles preservar su solvencia, optimizar el uso del capital regulatorio y conceder crédito de manera más responsable. Según el Banco de España (Banco de España, 2025), las variaciones en los indicadores de riesgo crediticio explican una parte sustancial de las oscilaciones en la rentabilidad del sector bancario, influyen en la percepción de estabilidad por parte de los mercados y condicionan el flujo de financiación hacia la economía real. Desde una perspectiva macroeconómica, una gestión eficaz del riesgo contribuye a reducir la probabilidad de crisis financieras, fortalecer la estabilidad del sistema bancario y favorecer el acceso de hogares y empresas a fuentes de financiación en condiciones más seguras y transparentes (Banco de España, 2024).

En este contexto regulatorio y operativo, los acuerdos de Basilea II y III consolidaron el uso de los modelos de calificación interna (*Internal Ratings-Based*, IRB) para que las entidades financieras estimen internamente los principales parámetros asociados al riesgo de crédito, entre ellos la PD, la Pérdida Dado el Incumplimiento (LGD) y la Exposición en el Momento del Incumplimiento (EAD) (on Banking Supervision, 2004). Este marco regulato-

rio busca garantizar que las entidades mantengan niveles de capital acordes con los riesgos asumidos y reforzar la resiliencia del sistema financiero internacional (on Banking Supervision, 2011). Una mayor precisión en la estimación de estos parámetros, tal como señala Alonso y Carbó (2021), permite optimizar los requerimientos de capital y mejorar la estabilidad financiera.

En los últimos años, la incorporación de técnicas avanzadas de *machine learning* (ML) ha modificado el desarrollo de estos modelos internos de riesgo. Frente a las limitaciones de los modelos tradicionales de *credit scoring* basados en regresiones lineales, los algoritmos de ML permiten utilizar grandes volúmenes de datos estructurados y no estructurados, detectar relaciones no lineales y capturar interacciones complejas entre variables asociadas a la probabilidad de impago (Munafo, 2019). Estas técnicas pueden mejorar la capacidad predictiva y la eficiencia operativa, al permitir evaluaciones más rápidas y precisas de los solicitantes (Menasalvas, Ángel Guzmán, y Ruíz Bonilla, 2022). En esta línea, Alonso y Carbó (2021) señalan que la aplicación de modelos de ensamble, como *eXtreme Gradient Boosting*(XGBoost), podría reducir los requerimientos de capital entre un 12,4 % y un 17 % en comparación con modelos logísticos tradicionales, lo que evidencia su potencial para optimizar recursos y mejorar la rentabilidad del sistema financiero.

No obstante, la adopción de modelos avanzados introduce un reto asociado a la pérdida de interpretabilidad. Como se observa en la Figura 1.2, los modelos con mayor capacidad predictiva, entre ellos los métodos de ensamble (*XGBoost* o *Random Forest*) y las redes neuronales, suelen presentar una estructura interna de difícil interpretación. Esta característica dificulta identificar los factores que explican cada predicción y limita la posibilidad de auditar decisiones automatizadas. Dicha opacidad supone un obstáculo técnico y regulatorio, especialmente en un contexto en el que normativas como el Reglamento General de Protección de Datos (Parlamento Europeo y Consejo de la Unión Europea, 2016) y las directrices de la Autoridad Bancaria Europea (Authority, 2021) exigen transparencia, claridad y rendición de cuentas en los procesos automatizados. En el ámbito bancario, donde las decisiones crediticias pueden incidir directamente en el bienestar financiero de los hogares y en la viabilidad de las empresas, la explicabilidad y auditabilidad de los modelos adquieren una relevancia particular.

En contraste con estos modelos de mayor complejidad, los algoritmos tradicionalmente empleados en la industria, como la regresión logística o los árboles de decisión, se ubican en la zona de mayor interpretabilidad representada en la Figura 1.2. Su estructura permite comprender con mayor facilidad la relación entre las variables explicativas y el riesgo de impago, aunque esta ventaja suele ir acompañada de una menor capacidad para capturar patrones no lineales e interacciones complejas. Esta relación inversa entre precisión e interpretabilidad genera una tensión significativa en la adopción de modelos de ML dentro del sector financiero. Las entidades deben equilibrar el uso de herramientas avanzadas que mejoren la exactitud de las estimaciones con la necesidad de ofrecer decisiones comprensibles y

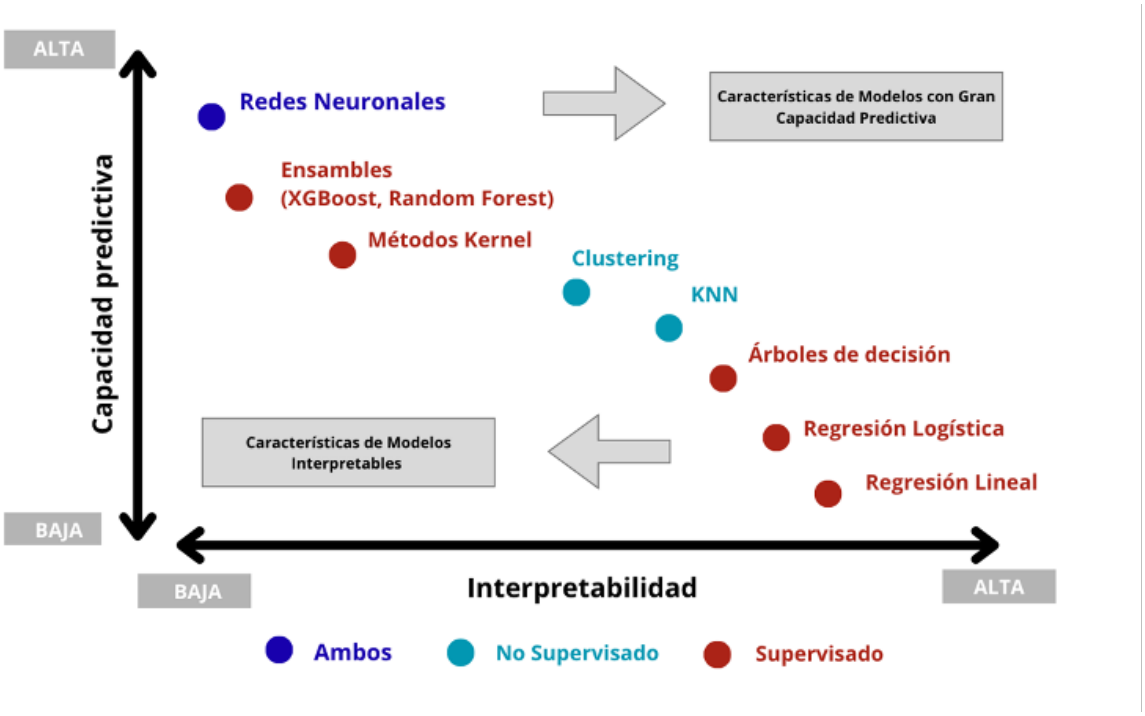

Figura 1.2: Compromiso entre interpretabilidad y capacidad predictiva de los modelos de ML. Elaboración Propia, adaptada de (Dhamangaonkar, 2020)

auditables. Al mismo tiempo, la presión competitiva incentiva la protección de los modelos más sofisticados como activos estratégicos, dado que una divulgación excesiva podría facilitar su replicación y reducir la ventaja competitiva de la entidad (Menasalvas et al., 2022). En este sentido, los modelos más transparentes pueden resultar más vulnerables a la ingeniería inversa, lo que limita su atractivo en entornos donde la diferenciación tecnológica constituye un factor estratégico.

Ante este panorama, resulta necesario avanzar hacia metodologías que permitan conciliar la capacidad predictiva de los modelos avanzados con las exigencias de transparencia propias del entorno regulatorio y financiero. En este contexto, la Inteligencia Artificial Explicable (XAI, *Explainable Artificial Intelligence*) ha adquirido una presencia creciente, al proporcionar herramientas orientadas a interpretar algoritmos de aprendizaje automático sin comprometer su rendimiento. Su finalidad es ofrecer explicaciones comprensibles y cuantificables sobre el proceso de decisión del modelo, facilitando su validación técnica y su aceptación por parte de supervisores y usuarios finales. Entre las aproximaciones más utilizadas destacan las técnicas *post-hoc*, como SHAP (*SHapley Additive exPlanations*) y LIME (*Local Interpretable Model-Agnostic Explanations*), que permiten analizar la contribución de cada variable a las predicciones de modelos complejos (Dhamangaonkar, 2020; Menasalvas et al., 2022). Su aplicación en la gestión del riesgo crediticio favorece la trazabilidad de las decisiones automatizadas y refuerza la confianza en el uso de sistemas basados en ML.

En este sentido, la estimación del riesgo crediticio mediante modelos avanzados no solo

permite predecir la probabilidad de incumplimiento, sino que también puede emplearse para estructurar la toma de decisiones financieras. La clasificación de los solicitantes según su solvencia esperada permite definir condiciones contractuales adecuadas, orientar las políticas de concesión y seguimiento del crédito, y diseñar estrategias de mitigación ajustadas a distintos perfiles de riesgo (on Banking Supervision, 2004; Pozo Fonseca, 2024). No obstante, para que esta clasificación tenga utilidad operativa y regulatoria, los perfiles identificados deben ser homogéneos, comprensibles y justificables. En este marco, la segmentación de solicitantes permite identificar grupos de individuos con patrones explicativos similares. Desde una perspectiva de gestión, estos segmentos favorecen decisiones más coherentes, optimizan la asignación de recursos y facilitan el diseño de políticas diferenciadas de concesión, precios y seguimiento. Desde una perspectiva regulatoria, una segmentación basada en criterios explicables contribuye a justificar la lógica de las decisiones automatizadas y a garantizar un tratamiento consistente entre solicitantes con características comparables (EBA, 2021). De este modo, explicar simultáneamente el nivel de riesgo y los factores que lo determinan contribuye a mejorar la transparencia, la eficiencia y la responsabilidad en la toma de decisiones financieras.

En este marco de referencia, el presente Trabajo de Fin de Grado propone el desarrollo de un modelo de segmentación supervisada basado en valores SHAP para la evaluación del riesgo crediticio. Esta aproximación combina la capacidad predictiva de modelos avanzados de ML con una representación interpretable de los factores que explican el riesgo de incumplimiento. A partir de dicha representación, se construyen segmentos homogéneos en el espacio explicativo generado por SHAP, con el propósito de identificar perfiles diferenciados de solicitantes y proporcionar una herramienta que favorezca decisiones crediticias más coherentes, auditables y alineadas con las exigencias regulatorias. Con ello, el trabajo busca integrar precisión predictiva, interpretabilidad y segmentación en un mismo marco analítico, contribuyendo a una gestión del riesgo más transparente, responsable y orientada a la mejora del proceso de concesión de crédito.

1.1. Objetivos

1.1.1. Objetivo General

Desarrollar un modelo de segmentación supervisada de solicitantes de crédito basado en valores SHAP, con el propósito de integrar capacidad predictiva e interpretabilidad en la evaluación del riesgo crediticio, permitiendo identificar perfiles homogéneos de riesgo y aportar mayor transparencia, coherencia y trazabilidad al proceso de toma de decisiones financieras.

1.1.2. Objetivos Específicos

- Analizar la importancia de la gestión del riesgo crediticio en el sistema financiero, considerando el papel de la estimación de la Probabilidad de Incumplimiento, el marco regulatorio asociado a los modelos internos de riesgo y la necesidad de decisiones crediticias transparentes, auditables y eficientes.

- Identificar, a partir de la revisión de la literatura, las principales técnicas automáticas que permiten abordar la predicción del riesgo crediticio y la interpretabilidad de los modelos, con especial atención a los métodos de ML y a las aproximaciones de XAI.

- Diseñar una metodología de segmentación supervisada basada en valores SHAP que permita transformar las predicciones del modelo de riesgo en una representación explicativa adecuada para identificar perfiles homogéneos de solicitantes de crédito.

- Implementar y evaluar un caso de estudio empírico mediante un modelo predictivo de riesgo crediticio, el cálculo de valores SHAP, la reducción de dimensionalidad y la aplicación de técnicas de *clustering*, con el fin de caracterizar segmentos diferenciados de solicitantes.

- Valorar la utilidad de los segmentos obtenidos en términos de interpretabilidad, coherencia interna y aplicabilidad para la toma de decisiones financieras, considerando su contribución a una gestión del riesgo más transparente y consistente.

1.2. Organización de la Memoria

El documento se organiza en cinco capítulos que desarrollan de forma progresiva los fundamentos teóricos, la revisión metodológica y la aplicación empírica del estudio. En el capítulo 1 se presenta el contexto general del trabajo y la motivación que lo sustenta, atendiendo a la gestión del riesgo crediticio y al papel de la explicabilidad en los modelos de *machine learning*. Asimismo, se formulan el objetivo general, los objetivos específicos y la justificación del enfoque adoptado. El capítulo 2 aborda la revisión del estado del arte. En él se examinan los principales modelos predictivos empleados en el ámbito del *credit scoring*, así como las metodologías de explicabilidad asociadas a la Inteligencia Artificial Explicable, con especial atención a los valores SHAP. A continuación, el capítulo 3 desarrolla un análisis crítico de las técnicas identificadas, valorando su desempeño, limitaciones y aplicabilidad en el contexto del riesgo de crédito.

El capítulo 4 presenta el caso de estudio. Se describe el conjunto de datos seleccionado, los procedimientos de preprocesamiento, el modelo predictivo implementado, el cálculo de valores SHAP y la transformación del espacio de representación. También se detalla el

algoritmo de *clustering* empleado para la segmentación de solicitantes y se analizan los resultados obtenidos. Finalmente, el capítulo 5 recoge las conclusiones del estudio y plantea posibles líneas futuras de investigación orientadas al desarrollo de modelos explicables y a su aplicación en entornos financieros reales.

Capítulo 2

Revisión de la literatura sobre modelos predictivos e interpretabilidad en riesgo crediticio

La evaluación del riesgo crediticio ha sido históricamente uno de los ámbitos de aplicación más relevantes de los métodos cuantitativos en finanzas. La necesidad de distinguir entre solicitantes con distintos niveles de solvencia llevó al desarrollo de técnicas de *credit scoring* orientadas a transformar información financiera, demográfica y comportamental en una medida sintética del riesgo de incumplimiento. Estos modelos permiten apoyar decisiones relacionadas con la concesión de crédito, la fijación de condiciones contractuales y la gestión del capital regulatorio. No obstante, la evolución de estas metodologías ha estado condicionada por una tensión persistente entre precisión predictiva, interpretabilidad y aplicabilidad operativa. Los primeros enfoques priorizaron modelos estadísticos relativamente simples, adecuados para entornos regulados y fácilmente auditables, mientras que el aumento de la disponibilidad de datos y la complejidad de los mercados financieros impulsó progresivamente la incorporación de técnicas más flexibles capaces de capturar patrones no lineales y relaciones de interacción entre variables.

En esta trayectoria, los primeros enfoques de *credit scoring*, desarrollados a partir de la década de 1940, se apoyaron en modelos estadísticos lineales orientados a estimar el riesgo de impago a partir de características observables del solicitante, como la edad, el historial financiero o los ingresos (Durand, 1941). Estos modelos aportaron una primera aproximación sistemática a la evaluación crediticia, al permitir transformar información individual en una medida cuantitativa de riesgo. Sin embargo, la regresión lineal presentaba limitaciones estructurales para este tipo de problemas, ya que presupone una relación lineal entre las variables explicativas y la probabilidad de incumplimiento, no garantiza predicciones acotadas en el intervalo $[0, 1]$ y asume errores con distribución normal y varianza constante. Estas condiciones rara vez se satisfacen en problemas de clasificación binaria, lo que dificulta interpretar

los valores estimados como probabilidades y puede afectar a la calibración y estabilidad del modelo (Hand y Henley, 1997). En respuesta a estas limitaciones, la regresión logística se consolidó progresivamente como una alternativa más adecuada para la evaluación crediticia. El estudio de Bolton (2009), basado en datos bancarios sudafricanos, comparó modelos lineales y logísticos, mostrando que la regresión logística ofrecía un mejor equilibrio entre precisión, estabilidad e interpretabilidad. Al modelar directamente la probabilidad de incumplimiento dentro del intervalo $[0, 1]$, este enfoque permite estimaciones más consistentes de la PD y facilita la construcción de *scorecards* auditables por analistas y supervisores. Por esta razón, el modelo logístico se convirtió en una referencia habitual en la industria financiera y en los procesos internos de evaluación del riesgo. No obstante, el aumento de la complejidad de los mercados, la disponibilidad de mayores volúmenes de datos y la necesidad de capturar relaciones no lineales impulsaron el desarrollo de técnicas más flexibles, capaces de representar interacciones entre variables y patrones de riesgo más heterogéneos.

En esta línea, Cardona Hernández (2004) aplicó árboles de decisión al análisis del riesgo crediticio en entidades financieras dentro del marco regulatorio colombiano. Su trabajo mostró que estos modelos permiten estimar probabilidades de incumplimiento e identificar perfiles de solicitantes con distintos niveles de riesgo mediante una estructura jerárquica de reglas fácilmente representable e interpretable. Esta característica favorece su uso en contextos financieros, donde no solo interesa clasificar a los solicitantes, sino también justificar los criterios que sustentan cada decisión. Sin embargo, la autora señaló que su utilidad depende de la calidad, estabilidad y actualización de la información disponible, dado que los modelos deben conservar su capacidad de discriminación a lo largo del tiempo para apoyar adecuadamente la gestión del riesgo crediticio (Cardona Hernández, 2004). Aunque los árboles de decisión aportaron una alternativa más interpretable frente a otros modelos estadísticos, su desempeño puede verse limitado cuando los datos presentan alta complejidad, ruido o relaciones no lineales difíciles de capturar mediante una única estructura de decisión. Esta restricción impulsó la adopción de métodos basados en la combinación de múltiples árboles, orientados a mejorar la estabilidad y la capacidad predictiva del modelo. En este contexto, Sharma (2011) comparó la regresión logística tradicional con modelos de *Random Forest* aplicados al análisis del riesgo crediticio utilizando datos hipotecarios. Los resultados indicaron que los *Random Forests* ofrecían mayor precisión y estabilidad predictiva, especialmente en presencia de multicolinealidad, aunque con una reducción en la interpretabilidad del modelo (Sharma, 2011).

Posteriormente, el interés por modelos con mayor flexibilidad predictiva favoreció la incorporación de redes neuronales en la evaluación crediticia. En esta línea, Angelini, di Tollo, y Roli (2008) desarrollaron dos sistemas neuronales para clasificar el riesgo de crédito en pequeñas empresas italianas, utilizando datos reales de 76 empresas y variables financieras observadas durante el periodo 2001-2003. Los autores compararon una red *feedforward* clásica con una arquitectura específica que organizaba la información histórica por grupos de varia-

bles correspondientes a tres años consecutivos. Los resultados mostraron un buen desempeño predictivo, aunque también evidenciaron que el tipo de error cometido resulta especialmente relevante en riesgo crediticio, dado que clasificar erróneamente a un prestatario incumplidor como solvente puede generar pérdidas financieras significativas. De forma complementaria, Khashman (2010) analizó redes neuronales supervisadas entrenadas mediante retropropagación sobre el conjunto de datos *German Credit*, comparando tres modelos y nueve esquemas de aprendizaje. Sus resultados mostraron que estas arquitecturas pueden capturar relaciones complejas entre variables y apoyar la automatización del proceso de evaluación crediticia. Sin embargo, tanto en este estudio como en el de Angelini et al. (2008), la mejora en flexibilidad predictiva se acompaña de una menor transparencia interna, ya que las redes neuronales no ofrecen reglas de decisión directamente interpretables. Esta limitación refuerza la necesidad de técnicas de explicabilidad aplicables a modelos de alta capacidad predictiva.

Ante la menor transparencia de los modelos de aprendizaje automático más flexibles, la literatura comenzó a desarrollar métodos orientados a explicar el comportamiento de algoritmos cuya estructura interna no resulta directamente interpretable. Este campo dio lugar a la XAI, cuyo propósito es proporcionar mecanismos de interpretación para modelos con alto rendimiento predictivo. Entre las aproximaciones más utilizadas se encuentran SHAP y LIME, técnicas *post-hoc* que permiten analizar la contribución de las variables a las predicciones individuales de modelos complejos (Hadji Misheva, Hirsa, Osterrieder, Kulkarni, y Lin, 2021). En particular, los valores SHAP cuantifican la aportación marginal de cada característica a una predicción concreta, permitiendo descomponer la probabilidad estimada de incumplimiento en contribuciones atribuibles a los distintos factores considerados por el modelo (Lundberg y Lee, 2017). La aplicación de estas técnicas en el ámbito financiero ha permitido avanzar hacia modelos predictivos más trazables y coherentes con las exigencias de supervisión. En el contexto de plataformas de préstamos entre particulares, Hadji Misheva et al. (2021) entrenaron distintos modelos de *machine learning* para la predicción del impago y aplicaron SHAP y LIME para interpretar sus resultados. Sus análisis identificaron como variables explicativas destacadas el importe total pagado, el importe del préstamo, el último pago realizado, los intereses recibidos y las recuperaciones. Desde una perspectiva financiera, las explicaciones indicaron que mayores pagos acumulados reducían la probabilidad estimada de impago, mientras que mayores importes de préstamo y recuperaciones se asociaban con un incremento del riesgo. Además, las explicaciones obtenidas mediante distintos explicadores resultaron consistentes, reforzando la confianza en la interpretación del modelo.

En una línea próxima, Ji (2021) analizaron SHAP y LIME en la detección de fraude en tarjetas de crédito. Tras comparar una red neuronal profunda y un modelo de *Random Forest*, los autores aplicaron técnicas de XAI sobre la DNN (*Deep Neural Network*), que obtuvo un desempeño ligeramente superior en términos de F1-score. Los resultados mostraron que variables como el uso de chip, el importe de la transacción, la hora de transferencia y el historial reciente de transacciones contribuían a identificar operaciones potencialmente fraudulentas.

El estudio incorporó además una evaluación con usuarios del sector bancario, en la que las explicaciones generadas mediante LIME y SHAP mejoraron la comprensión y la confianza percibida frente a un escenario sin XAI. Estos resultados muestran que la explicabilidad no solo aporta trazabilidad técnica, sino que también favorece la aceptación de los modelos por parte de profesionales financieros.

A partir de estos avances, la literatura ha comenzado a explorar el uso de explicaciones locales no solo como mecanismo de interpretación individual, sino también como representación alternativa para la segmentación. En lugar de agrupar a los solicitantes a partir de sus características originales, estos enfoques representan cada observación mediante los factores que explican su predicción, lo que permite identificar grupos con mecanismos de riesgo similares. En esta línea, Cooper (2022) introdujo de forma preliminar el concepto de *supervised clustering*, mostrando mediante datos simulados que los valores SHAP pueden actuar como una representación intermedia capaz de mejorar la separación visual de los subgrupos frente al *clustering* tradicional. Su contribución tiene un carácter principalmente conceptual, al ilustrar el potencial de emplear explicaciones del modelo como base para el agrupamiento.Un desarrollo empírico más cercano al riesgo crediticio es el de Gramegna y Giudici (2021), quienes comparan la capacidad discriminativa de SHAP y LIME cuando sus explicaciones locales se utilizan como variables de entrada para técnicas no supervisadas y predictivas. A partir de un modelo XGBoost entrenado sobre datos de pymes italianas, los autores aplican K-means y *spectral clustering* sobre los valores explicativos, además de modelos *Random Forest* para evaluar su capacidad predictiva posterior. Sus resultados muestran que SHAP genera una representación con mayor cohesión y separación de grupos que LIME, con mejores métricas de silueta y *Davies-Bouldin*, así como un AUC (*Area Under the ROC Curve*) superior en el análisis predictivo. Estos trabajos abren la posibilidad de utilizar explicaciones del modelo como base para construir segmentos interpretables de riesgo, línea en la que se sitúa el presente trabajo.

La Tabla 2.1 sintetiza los estudios revisados y permite observar la evolución metodológica del riesgo crediticio desde los modelos estadísticos tradicionales hasta los enfoques recientes basados en explicabilidad y segmentación supervisada. En conjunto, la literatura muestra un desplazamiento desde modelos transparentes, como la regresión logística y los árboles de decisión, hacia algoritmos con mayor capacidad predictiva, como los *Random Forests* y las redes neuronales. Aunque estos métodos han mejorado la estimación del riesgo de incumplimiento, también han incrementado la dificultad para justificar sus decisiones en entornos financieros regulados. En respuesta a este compromiso entre rendimiento e interpretabilidad, los estudios más recientes han incorporado técnicas de XAI, especialmente SHAP y LIME, para aportar trazabilidad a las predicciones de modelos complejos. Estas herramientas permiten identificar los factores que explican cada predicción y evaluar su coherencia financiera. Además, la literatura reciente muestra que los valores SHAP pueden utilizarse como una representación explicativa para segmentar solicitantes según los factores

que impulsan su riesgo estimado. En esta línea, el presente trabajo propone una metodología de segmentación supervisada basada en valores SHAP, incorporando UMAP (*Uniform Manifold Approximation and Projection*) para analizar la estructura del espacio explicativo y DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) para identificar perfiles homogéneos de riesgo sin imponer previamente un número fijo de grupos.

Tabla 2.1: Resumen de los estudios sobre modelos de riesgo crediticio, explicabilidad y segmentación

Referencia	Dataset	Objetivo	Algoritmo / Técnica	Resultados
(Bolton, 2009)	Datos bancarios sudafricanos	Comparar regresión lineal y regresión logística para la predicción de incumplimiento	Regresión lineal y regresión logística	Demuestra que la regresión logística ofrece mejores probabilidades de impago que la regresión lineal, manteniendo transparencia e interpretabilidad. Establece la regresión logística como estándar en scoring.
(Cardona Hernández, 2004)	Datos internos de entidades financieras colombianas	Analizar árboles de decisión como herramienta de estimación de probabilidad de incumplimiento	Árboles de decisión (CART/CHAID, etc.)	Modelo interpretable y útil dentro del marco regulatorio. Muestra ejemplos de potencia (ROC, Kolmogorov-Smirnov Statistic) y estabilidad, pero no incluye análisis cuantitativo completo. Señala limitaciones por estabilidad y calidad de datos.
(Sharma, 2011)	Dataset hipotecario (home equity)	Evaluar el uso de Random Forests como alternativa a scorecards clásicos	Random Forests vs. regresión logística	Random Forest logra mayor precisión y estabilidad en presencia de multicolinealidad; aumenta capacidad predictiva pero reduce interpretabilidad. Propuesta híbrida: usar importancia de variables Random Forest para mejorar modelos logísticos.
(Angelini et al., 2008)	Datos reales de 76 pequeñas empresas italianas	Evaluar la aplicación de redes neuronales en la clasificación del riesgo de crédito	Redes neuronales *feedforward* y arquitectura neuronal específica	Muestra que las redes neuronales pueden aprender patrones asociados a la solvencia e incumplimiento de los prestatarios y alcanzar buen desempeño predictivo. También evidencia la necesidad de analizar el tipo de error cometido y la menor interpretabilidad de estos modelos.
(Khashman, 2010)	German Credit Dataset	Analizar la capacidad de redes neuronales supervisadas para evaluar solicitudes de crédito	Redes neuronales con algoritmo de *backpropagation*	Muestra que las redes neuronales pueden capturar relaciones complejas y mejorar el desempeño en la clasificación de solicitudes, aunque su estructura interna dificulta la interpretación directa de las decisiones.
(Hadji Misheva et al., 2021)	Dataset de préstamos Peer-to-Peer (P2P)	Evaluar utilidad de SHAP y LIME en predicción e interpretabilidad en P2P lending	Modelos de ML + SHAP + LIME	SHAP ofrece explicaciones robustas pero costosas computacionalmente; LIME es rápido pero limitado. Muestran viabilidad y dificultades prácticas de XAI en crédito.
(Ji, 2021)	Dataset de fraude en tarjetas	Aplicar SHAP y LIME para detectar fraude y mejorar trazabilidad	Modelos ML + SHAP + LIME	Permiten identificar patrones anómalos e incrementan transparencia en detección de fraude. Mejoran confianza del usuario y diferenciación entre transacciones normales y sospechosas.
(Cooper, 2022)	Un Dataset simulado con 1000 observaciones y 50 variables	Proponer supervised clustering basado en SHAP como alternativa al clustering tradicional	SHAP + PCA + clustering	Demuestra que usar SHAP como representación supervisada crea clusters más separados y coherentes con los patrones del modelo que el clustering sobre datos originales. Aporte conceptual.
(Gramegna y Giudici, 2021)	Dataset de pymes italianas	Comparar capacidad discriminativa de SHAP y LIME para agrupar observaciones en riesgo	XGBoost + SHAP + LIME + K-Means + Spectral clustering	SHAP genera clusters más homogéneos, estables y coherentes. Silhouette y *Davies-Bouldin Index* confirman superioridad de SHAP. AUC superior en modelo supervisado (0.864 vs. 0.839). Validan el uso de XAI como base para segmentación en riesgo crediticio.

Capítulo 3

Metodología de Análisis de Datos

El presente capítulo describe la metodología diseñada para desarrollar un modelo de segmentación supervisada de solicitantes de crédito basado en valores SHAP. El propósito del procedimiento es identificar perfiles homogéneos de riesgo crediticio a partir de una representación interpretable de los factores que explican las predicciones del modelo. Como se muestra en la Figura 3.1, el proceso metodológico se estructura en cuatro etapas. La primera corresponde a la identificación y recolección de datos, donde se selecciona el conjunto de información utilizado para el análisis del riesgo crediticio. Esta fase incluye la definición de la variable objetivo, representada por la clasificación del solicitante como *good* o *bad*, así como la revisión de las variables sociodemográficas, económicas y financieras disponibles. Asimismo, se evalúa la calidad general del conjunto de datos y se identifican las características con potencial explicativo para el modelo.

La segunda etapa comprende el preprocesamiento y el análisis descriptivo de los datos. En esta fase se realizan las tareas necesarias para garantizar la consistencia de la información, incluyendo la revisión de duplicados, el tratamiento de valores faltantes, la detección de posibles inconsistencias, la gestión de valores atípicos y la transformación de variables. Posteriormente, se desarrolla un análisis exploratorio de datos con el fin de examinar la distribución de las variables, la relación entre las características disponibles y la variable objetivo, así como posibles patrones asociados al riesgo crediticio. Este análisis permite comprender la estructura inicial del conjunto de datos y proporciona una base empírica para interpretar los resultados posteriores.

La tercera etapa corresponde a la aplicación del modelo supervisado, el cálculo de valores SHAP y la construcción de los segmentos. En primer lugar, se entrena un modelo predictivo orientado a estimar la probabilidad de riesgo o incumplimiento de cada solicitante. A partir de dicho modelo, se calculan los valores SHAP, que permiten descomponer cada predicción en contribuciones atribuibles a las variables explicativas. Esta representación transforma el conjunto de datos original en un espacio explicativo, donde cada observación queda descrita por los factores que impulsan su predicción de riesgo. Sobre este espacio se aplican técnicas

de reducción de dimensionalidad y algoritmos de *clustering*, con el objetivo de identificar grupos homogéneos de solicitantes definidos por mecanismos explicativos similares. La calidad de los grupos obtenidos se evalúa mediante métricas de validación apropiadas para el análisis de segmentación.

Finalmente, la cuarta etapa se centra en el análisis e interpretación de los resultados. En esta fase se caracterizan los segmentos identificados a partir de sus valores SHAP, sus niveles de riesgo estimado y las variables que explican su comportamiento. Este análisis permite construir perfiles diferenciados de solicitantes y valorar su utilidad para la gestión del riesgo crediticio. Asimismo, se examina la aplicabilidad del enfoque propuesto en procesos de decisión financiera, considerando su contribución a la transparencia, la coherencia y la eficiencia en la concesión de crédito.

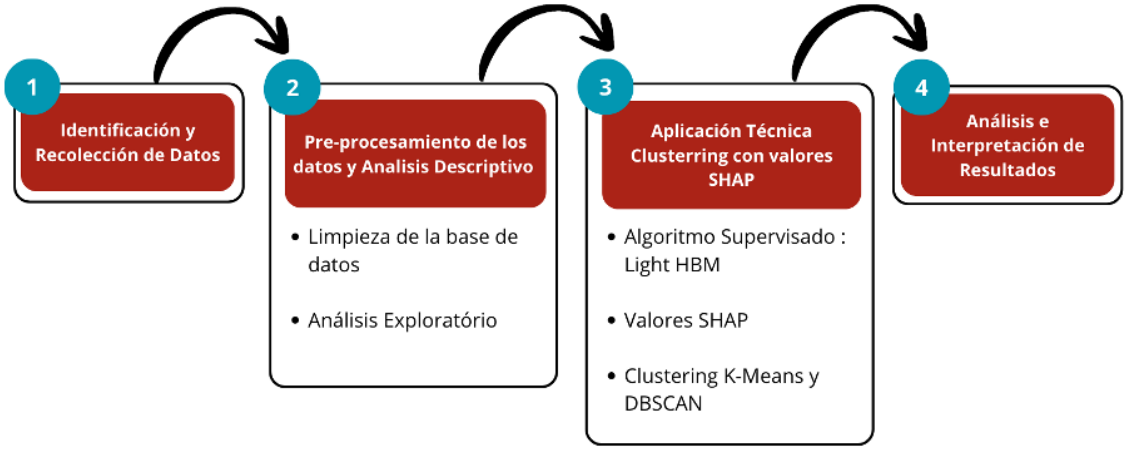

Figura 3.1: Etapas de la metodología seguida en el estudio. Elaboración propia.

3.1. Identificación y recolección de datos

La primera etapa del estudio se centra en la identificación y selección de un conjunto de datos adecuado para analizar perfiles de riesgo crediticio mediante una metodología de segmentación supervisada. Para ello, se consideran bases de datos disponibles en repositorios reconocidos, como Kaggle, el *UCI Machine Learning Repository* y otras fuentes especializadas en riesgo financiero. Los conjuntos de datos candidatos se evalúan atendiendo a su calidad, relevancia y grado de adecuación a los objetivos del trabajo, priorizando aquellos que incluyan información sociodemográfica, económica y crediticia de los solicitantes, así como variables habitualmente empleadas en modelos de *credit scoring*. La evaluación inicial de los datos considera tanto el volumen de observaciones como la diversidad y naturaleza de las variables disponibles. Un número suficiente de registros favorece la estabilidad del modelo predictivo y reduce el riesgo de obtener resultados excesivamente dependientes de patrones muestrales específicos. Por su parte, la presencia de variables relacionadas con la situación

financiera, el historial crediticio, la finalidad del préstamo y las características personales del solicitante permite representar de forma más completa los factores asociados al riesgo de incumplimiento. Esta diversidad resulta especialmente importante en un enfoque basado en valores SHAP, dado que la posterior interpretación del modelo dependerá de la capacidad de las variables para capturar mecanismos explicativos del riesgo.

Dado que la metodología propuesta se fundamenta en un enfoque supervisado, el conjunto de datos debe disponer de una variable objetivo que permita distinguir entre distintos niveles de riesgo crediticio. En este estudio, dicha función la desempeña la variable *risk*, clasificada en las categorías *good* y *bad*. Esta variable permite entrenar el modelo predictivo, estimar la probabilidad de incumplimiento y calcular posteriormente los valores SHAP asociados a cada observación. De este modo, la etiqueta supervisada no solo cumple una función predictiva, sino que también actúa como punto de partida para construir el espacio explicativo sobre el que se realizará la segmentación. Finalmente, se considera la estructura y organización del conjunto de datos seleccionado. Se priorizan bases de datos con un formato tabular claro, variables correctamente documentadas y compatibilidad con las herramientas de análisis empleadas. Una estructura bien definida facilita las tareas de limpieza, codificación y transformación de variables, y contribuye a garantizar la reproducibilidad del procedimiento metodológico.

3.2. Pre-procesamiento y Análisis Descriptivo de los Datos

El preprocesamiento de los datos constituye una etapa orientada a asegurar que la información utilizada sea consistente, comparable y adecuada para el entrenamiento del modelo supervisado y para la posterior segmentación basada en valores SHAP. En esta fase se depura la base de datos, se transforman las variables y se preparan las estructuras necesarias para que los algoritmos puedan procesar correctamente la información disponible. Este procedimiento se complementa con un análisis descriptivo inicial, cuyo propósito es examinar la composición del conjunto de datos, estudiar la distribución de las variables e identificar patrones preliminares asociados al riesgo crediticio.

En primer lugar, se realiza una revisión de la calidad de los datos con el fin de identificar registros incompletos, duplicados, inconsistentes o potencialmente atípicos. Los valores faltantes se analizan atendiendo tanto a su frecuencia como a su distribución entre variables. Cuando su presencia sea reducida y no comprometa la representatividad del conjunto de datos, se optará por la eliminación de las observaciones afectadas. En caso contrario, se valorarán estrategias de imputación acordes con la naturaleza de la variable y con el impacto potencial sobre el modelo. Las observaciones duplicadas serán eliminadas para evitar que determinados perfiles queden sobrerrepresentados y distorsionen el aprendizaje del algoritmo. Los valores atípicos se evaluarán mediante criterios estadísticos y visuales, evaluando si

corresponden a errores de registro, situaciones extremas pero plausibles o patrones de riesgo potencialmente informativos. Esta distinción resulta importante en riesgo crediticio, donde ciertos valores extremos pueden reflejar perfiles financieros de interés y no necesariamente ruido.

En segundo lugar, se procede a la codificación y transformación de las variables. Las variables categóricas ordinales se codificarán respetando el orden natural de sus categorías, mientras que las variables nominales se transformarán mediante *one-hot encoding*, evitando introducir relaciones jerárquicas inexistentes entre categorías. Las variables numéricas podrán ser escaladas cuando el algoritmo o la técnica posterior lo requiera, especialmente en etapas sensibles a la escala de las variables, como el análisis de distancias, la reducción de dimensionalidad o el *clustering*. En este sentido, aunque modelos basados en árboles como LightGBM (*Light Gradient Boosting Machine*) no requieren necesariamente estandarización para su entrenamiento, las fases posteriores de representación, proyección y agrupamiento sí pueden beneficiarse de una preparación cuidadosa del espacio de entrada.

De manera complementaria, se llevará a cabo un análisis descriptivo y exploratorio de los datos. Este análisis incluirá estadísticas de tendencia central y dispersión, así como representaciones gráficas que permitan analizar la distribución de las variables, la presencia de asimetrías, la concentración de categorías y posibles diferencias entre solicitantes clasificados como *good* y *bad*. También se analizarán asociaciones entre variables predictoras y la variable objetivo *risk*, mediante tablas de contingencia, comparaciones de distribuciones, medidas de correlación o gráficos exploratorios, según la naturaleza de cada variable. Este análisis no tiene como finalidad seleccionar variables únicamente por su correlación lineal con el riesgo, sino comprender la estructura del conjunto de datos y orientar la interpretación posterior del modelo.

3.3. Clustering Supervisado con valores SHAP

Para implementar el enfoque de *clustering* supervisado basado en valores SHAP, se plantea una secuencia metodológica en la que cada técnica cumple una función específica dentro del proceso de modelado, explicación y segmentación. En primer lugar, se emplea LightGBM como modelo supervisado para estimar la probabilidad de riesgo crediticio de cada solicitante. La elección de este algoritmo responde a su capacidad para modelar relaciones no lineales, manejar interacciones entre variables y ofrecer un buen desempeño en problemas de clasificación con datos tabulares. Una vez entrenado el modelo, se calculan los valores SHAP con el fin de descomponer cada predicción en contribuciones atribuibles a las variables explicativas. Esta etapa permite transformar el espacio original de los datos en una representación explicativa, donde cada observación queda descrita por los factores que impulsan su riesgo estimado.

Posteriormente, se aplica UMAP para proyectar el espacio SHAP en una dimensión reducida que facilite el análisis de la estructura de los datos y la visualización de posibles agrupamientos. Esta técnica permite conservar relaciones locales entre observaciones, lo que resulta adecuado para identificar patrones de proximidad en función de mecanismos explicativos similares. Finalmente, sobre la representación generada se aplica DBSCAN, un algoritmo de *clustering* basado en densidad que permite identificar grupos sin fijar previamente el número de segmentos y detectar observaciones atípicas o de difícil asignación. En los apartados siguientes se describen con mayor detalle estas técnicas y su función dentro de la metodología propuesta.

3.3.1. LightGBM: Modelo Supervisado para la Predicción del Riesgo Crediticio

Los modelos basados en *Gradient Boosting Decision Trees* (GBDT) han adquirido una amplia presencia en problemas de clasificación y regresión debido a su capacidad para combinar múltiples árboles de decisión débiles en un predictor agregado con alto rendimiento. En el ámbito del riesgo crediticio, este tipo de modelos resulta adecuado porque permite capturar relaciones no lineales, interacciones entre variables y patrones heterogéneos asociados a la probabilidad de incumplimiento. No obstante, las implementaciones tradicionales de GBDT pueden presentar limitaciones computacionales cuando se aplican a conjuntos de datos con un elevado número de observaciones o variables. Ke et al. (2017) señalan que una de las principales fuentes de coste proviene del cálculo de la ganancia de información, que exige evaluar múltiples puntos de partición para cada característica y sobre un número amplio de instancias. LightGBM surge como una implementación optimizada del *gradient boosting* diseñada para mejorar la eficiencia computacional sin renunciar a la capacidad predictiva de los modelos basados en árboles (Ke et al., 2017).

Una de sus características distintivas es su estrategia de crecimiento *leaf-wise*, que difiere del esquema *level-wise* empleado en otros algoritmos de árboles. En el crecimiento *level-wise*, ilustrado en la parte izquierda de la Figura 3.2, el árbol se expande de forma equilibrada por niveles, dividiendo simultáneamente los nodos situados a la misma profundidad. Este procedimiento genera estructuras más simétricas, pero puede dedicar particiones a regiones del espacio donde la reducción del error es limitada. En cambio, el crecimiento *leaf-wise*, representado en la parte derecha de la Figura 3.2, selecciona en cada iteración la hoja cuya división produce la mayor reducción de la función de pérdida. Como resultado, el árbol tiende a desarrollarse de forma asimétrica, profundizando en aquellas ramas donde el modelo identifica mayor ganancia predictiva. Esta estrategia permite concentrar la complejidad del modelo en las zonas del espacio de datos que presentan mayor dificultad de clasificación, lo que puede traducirse en una mejora del rendimiento frente a esquemas más uniformes de expansión. En el contexto del riesgo crediticio, esta propiedad resulta útil para capturar perfiles

de solicitantes con patrones de riesgo no lineales o con interacciones específicas entre variables financieras y sociodemográficas. No obstante, la flexibilidad del crecimiento *leaf-wise* requiere un control cuidadoso de la complejidad del modelo. Al permitir que determinadas ramas alcancen mayor profundidad, LightGBM puede ajustarse en exceso a patrones particulares del conjunto de entrenamiento si no se regulan adecuadamente sus hiperparámetros. Por ello, en este trabajo se considera necesario prestar atención a parámetros como *max_depth*, que limita la profundidad máxima de los árboles; *num_leaves*, que controla el número máximo de hojas generadas y, por tanto, la complejidad estructural del modelo; *min_child_samples*, que establece el número mínimo de observaciones requeridas en cada hoja para evitar divisiones excesivamente específicas; *learning_rate*, que regula la contribución de cada árbol al modelo global y permite controlar la velocidad de aprendizaje; y el número de estimadores (*n_estimators*), que determina cuántos árboles se incorporan al proceso de *boosting*. El ajuste conjunto de estos parámetros permite equilibrar capacidad predictiva y generalización, reduciendo el riesgo de sobreajuste y favoreciendo un comportamiento más estable del modelo.

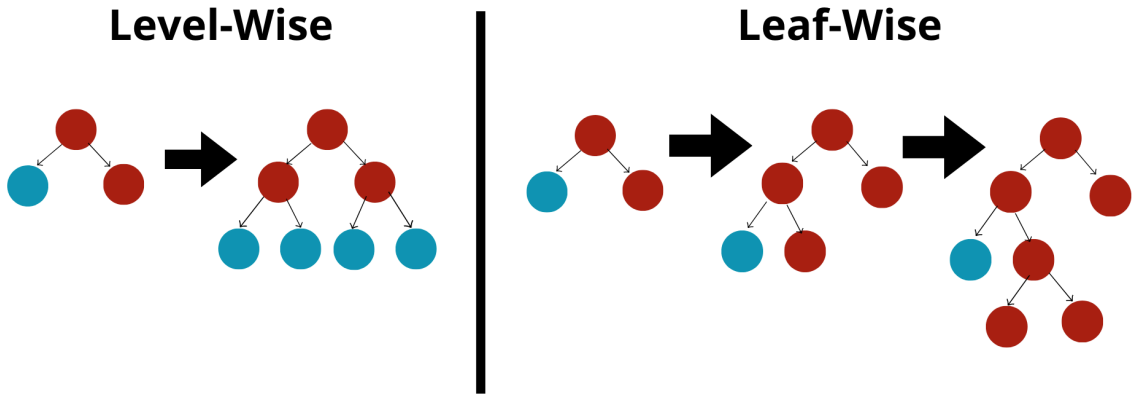

Figura 3.2: Ilustración de las divisiones Leaf-Wise y Level-Wise. Elaboración propia, adaptada de: Technology (2023)

Además de la estrategia de crecimiento *leaf-wise*, LightGBM incorpora mecanismos orientados a reducir el coste computacional del entrenamiento sin alterar la lógica del *gradient boosting*. Entre ellos destacan *Gradient-based One-Side Sampling* (GOSS) y *Exclusive Feature Bundling* (EFB), dos procedimientos diseñados para optimizar, respectivamente, la selección de observaciones y el tratamiento de variables de alta dimensionalidad (Ke et al., 2017). GOSS parte de la idea de que no todas las observaciones contribuyen del mismo modo a la actualización del modelo. En cada iteración, las instancias con gradientes más altos son aquellas sobre las que el modelo presenta un mayor error y, por tanto, contienen más información para la siguiente partición del árbol. Por ello, GOSS conserva estas observaciones y realiza un muestreo aleatorio sobre las instancias con gradientes más bajos, ajustando posteriormente su peso para mantener una estimación adecuada de la ganancia de información.

De esta forma, el algoritmo reduce el número de observaciones utilizadas en el cálculo de las divisiones, pero preserva aquellas que aportan mayor información al proceso de aprendizaje (Ke et al., 2017). Por su parte, EFB se orienta a reducir el número efectivo de variables mediante la agrupación o empaquetamiento (*bundling*) de características mutuamente excluyentes, es decir, variables que rara vez toman valores distintos de cero de manera simultánea. Esta situación es frecuente en conjuntos de datos con variables categóricas codificadas mediante indicadores binarios o en matrices altamente dispersas. Al combinar estas características en un mismo paquete, EFB disminuye la dimensionalidad del espacio de entrada y reduce el coste de búsqueda de particiones, manteniendo una pérdida de información limitada (Ke et al., 2017). Esta propiedad resulta especialmente útil cuando el conjunto de datos contiene muchas variables derivadas de procesos de codificación categórica.

La Figura 3.3 resume de forma esquemática esta lógica de funcionamiento. En primer lugar, el conjunto de datos se somete a los procedimientos de GOSS y EFB, que permiten reducir el número de observaciones y características procesadas en cada iteración. A continuación, se entrena un árbol de decisión sobre la información seleccionada y se evalúa el error del modelo. Las observaciones con mayor error, representadas en la figura como puntos destacados, adquieren mayor peso en la siguiente iteración, de modo que los árboles posteriores se concentran en corregir los errores cometidos por los anteriores. Este proceso iterativo permite construir un modelo agregado en el que cada nuevo árbol contribuye a mejorar la predicción final. Estos mecanismos explican la eficiencia computacional de LightGBM frente a implementaciones más tradicionales de GBDT.

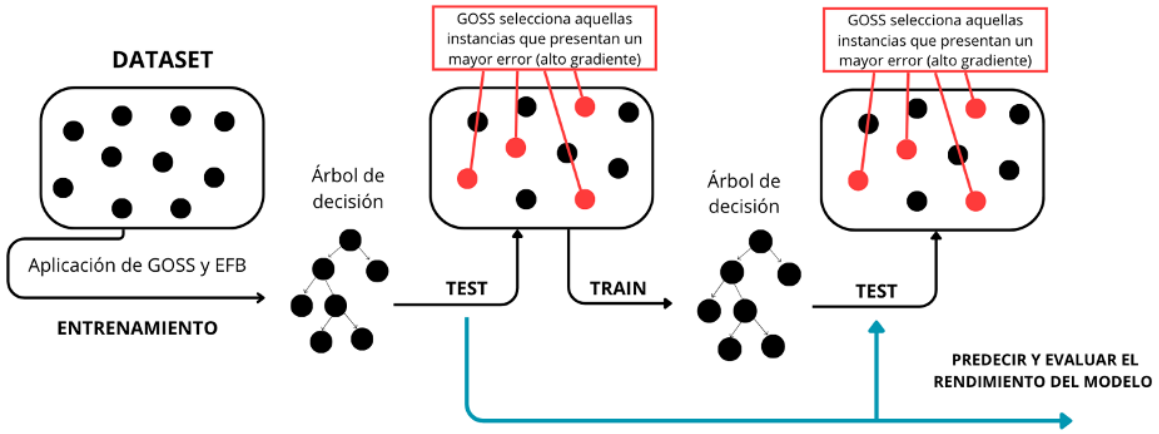

Figura 3.3: Proceso ilustrativo del algoritmo LightGBM. Elaboración propia.

3.3.2. Valores SHAP para la construcción del espacio explicativo

Una vez entrenado el modelo predictivo mediante LightGBM, el siguiente paso consiste en interpretar sus predicciones para identificar la contribución de cada variable al riesgo

estimado de cada solicitante. Esta interpretación se realiza mediante SHAP, una técnica de explicabilidad *post-hoc* que permite descomponer la salida del modelo en efectos atribuibles a las variables explicativas. De este modo, SHAP no solo permite conocer la importancia global de las variables, sino también analizar, para cada observación individual, qué factores incrementan o reducen la probabilidad estimada de incumplimiento. Los valores SHAP se fundamentan en los valores de Shapley, procedentes de la teoría de juegos cooperativos. En este marco, se plantea el problema de distribuir de forma equitativa la ganancia obtenida por una coalición entre los distintos participantes que la conforman. Trasladado al contexto del riesgo crediticio, la predicción del modelo puede entenderse como el resultado de la cooperación entre varias características del solicitante, tales como la edad, el historial crediticio o los ingresos. Cada variable actúa como un participante que aporta información al resultado final, y el valor SHAP mide la contribución media de dicha variable a la predicción, considerando todas las combinaciones posibles de variables (Lundberg y Lee, 2017).

La Figura 3.4 ilustra esta lógica mediante un ejemplo simplificado. En la parte izquierda se muestra la comparación entre dos predicciones del modelo. Una obtenida con un conjunto de variables que incluye la característica analizada y otra calculada sin dicha característica. La diferencia entre ambas predicciones representa la contribución marginal de esa variable en una determinada combinación de información. Sin embargo, esta contribución puede variar dependiendo de las demás variables presentes en el modelo. Por ello, como se observa en la parte derecha de la figura, SHAP calcula dicha diferencia en distintas coaliciones de variables y obtiene un promedio ponderado de todas ellas. En el ejemplo, el valor SHAP asociado a la variable edad se obtiene a partir de las contribuciones marginales calculadas en las diferentes combinaciones posibles con el historial crediticio y los ingresos.

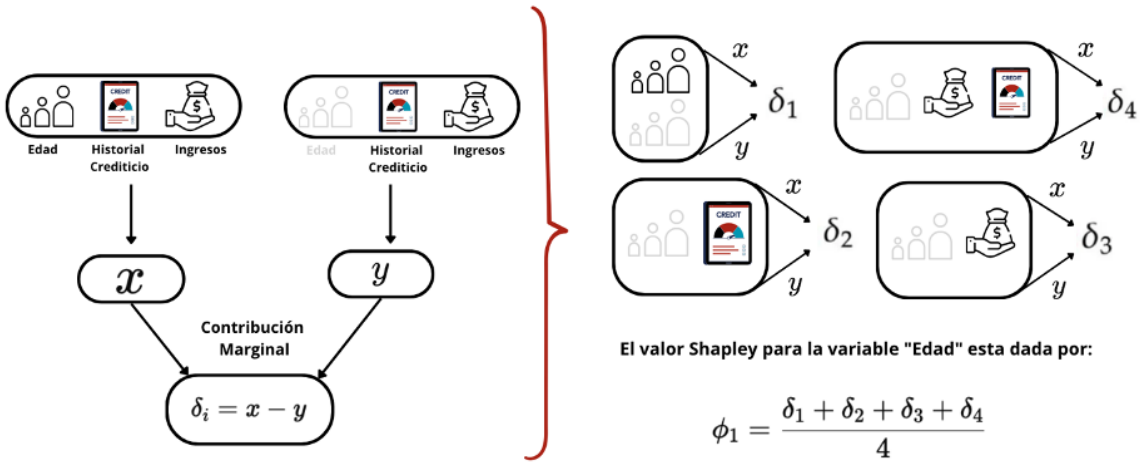

Figura 3.4: Ilustración del funcionamiento de SHAP. Elaboración Propia.

Desde el punto de vista interpretativo, un valor SHAP positivo indica que la variable contribuye a aumentar la salida del modelo respecto al valor base, mientras que un valor negativo indica que reduce dicha salida. En un problema de riesgo crediticio, si la salida del modelo se

interpreta como probabilidad de incumplimiento, una contribución positiva desplaza la predicción hacia un mayor riesgo estimado, mientras que una contribución negativa la desplaza hacia un menor riesgo. Esta propiedad permite analizar tanto la dirección como la magnitud del efecto de cada característica sobre la predicción individual. Formalmente, el valor SHAP de una variable i se define como el promedio ponderado de sus contribuciones marginales sobre todos los subconjuntos posibles de variables. Su expresión matemática es la siguiente:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right], \tag{3.1}$$

donde ϕ_i representa el valor SHAP asociado a la variable i, N denota el conjunto completo de variables del modelo, S corresponde a un subconjunto de variables que no incluye a i, $v(S)$ representa la predicción obtenida al considerar únicamente las variables incluidas en S, y el término $v(S \cup \{i\}) - v(S)$ mide la contribución marginal de la variable i al incorporarse a dicho subconjunto. El factor $\frac{|S|!\,(|N|-|S|-1)!}{|N|!}$ representa la ponderación asignada a cada coalición y refleja la proporción de órdenes posibles en los que la variable i puede incorporarse a la coalición S. Esta formulación permite asignar a cada característica una contribución coherente con su efecto medio sobre la predicción del modelo.

Para ilustrar esta descomposición de manera más intuitiva, la Figura 3.5 muestra la agregación de las contribuciones individuales de las variables hasta obtener la salida final del modelo. En el ejemplo representado, el modelo parte de una tasa base de $0{,}1$, entendida como la predicción promedio antes de incorporar la información específica del solicitante. A partir de ese valor inicial, cada característica desplaza la predicción en una dirección determinada, aumentando o reduciendo la probabilidad estimada de riesgo. En concreto, la edad del solicitante, con un valor de 65 años, incrementa la predicción en $+0{,}4$, mientras que la variable género, correspondiente a la categoría femenina, la reduce en $-0{,}3$. Por su parte, unos ingresos de 5.000 euros y un historial crediticio de dos años aportan contribuciones positivas de $+0{,}1$ cada una. La predicción final se obtiene mediante la suma de la tasa base y de todas las contribuciones SHAP asociadas a las variables consideradas:

$$\text{Predicción final} = \text{Tasa base} + \sum_i \phi_i = 0{,}1 + (0{,}4 - 0{,}3 + 0{,}1 + 0{,}1) = 0{,}4. \tag{3.2}$$

Esta representación permite identificar las variables que aumentan o reducen la predicción y observar la propiedad aditiva de SHAP, según la cual la suma de las contribuciones individuales coincide con la diferencia entre el valor base y la predicción final del modelo. Por lo tanto, cada predicción puede interpretarse como una composición de efectos positivos y negativos asociados a las características del solicitante. Además de facilitar la interpretación individual, los valores SHAP ofrecen una ventaja metodológica para el *clustering* supervi-

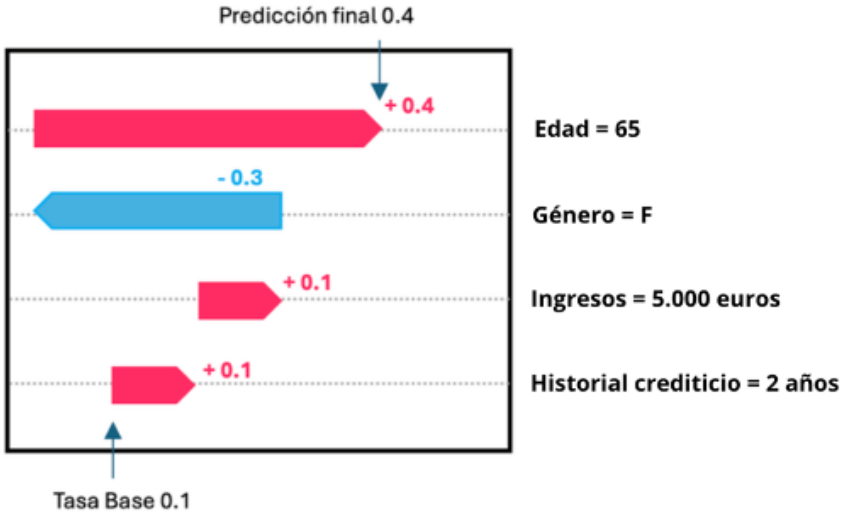

Figura 3.5: Decomposición de la predicción mediante valores SHAP. Elaboración propia, adaptada de (Echanove Puig, 2025)

sado. En los enfoques tradicionales de segmentación, las observaciones suelen agruparse a partir de sus variables originales, después de aplicar tareas de limpieza, transformación y, en algunos casos, técnicas de reducción de dimensionalidad. Sin embargo, cuando el agrupamiento se realiza directamente sobre el espacio original de variables, los segmentos pueden verse afectados por diferencias de escala, variables poco informativas o relaciones complejas que no se reflejan de forma explícita en las distancias entre observaciones. Como resultado, los *clusters* pueden presentar solapamientos, baja separación interna o dificultades de interpretación.

El enfoque de *clustering* supervisado basado en SHAP reformula este proceso al introducir una etapa intermedia de representación explicativa. En lugar de agrupar directamente a los solicitantes a partir de sus características observadas, primero se entrena un modelo supervisado y, posteriormente, se calculan los valores SHAP para cada observación. De este modo, cada solicitante queda representado por la contribución de sus variables a la predicción de riesgo, y no únicamente por los valores originales de dichas variables. Esta transformación genera un espacio en el que la proximidad entre observaciones refleja similitud en los mecanismos que impulsan la decisión del modelo. La Figura 3.6 ilustra esta diferencia entre el *clustering* tradicional y el *clustering* supervisado. En la parte superior, los datos originales se proyectan en un espacio de menor dimensión, pero la representación obtenida muestra una nube de puntos con alto solapamiento, donde no se identifican grupos claramente diferenciados. Esto ocurre porque la proyección se construye a partir de variables que no han sido ponderadas por su influencia en la predicción. En cambio, en la parte inferior, el mismo proceso se realiza sobre los valores SHAP. Al proyectar este espacio explicativo en menor dimensión, los puntos aparecen organizados en grupos más separados, lo que indica que las observaciones comparten patrones similares en términos de los factores que explican su ries-

go estimado.

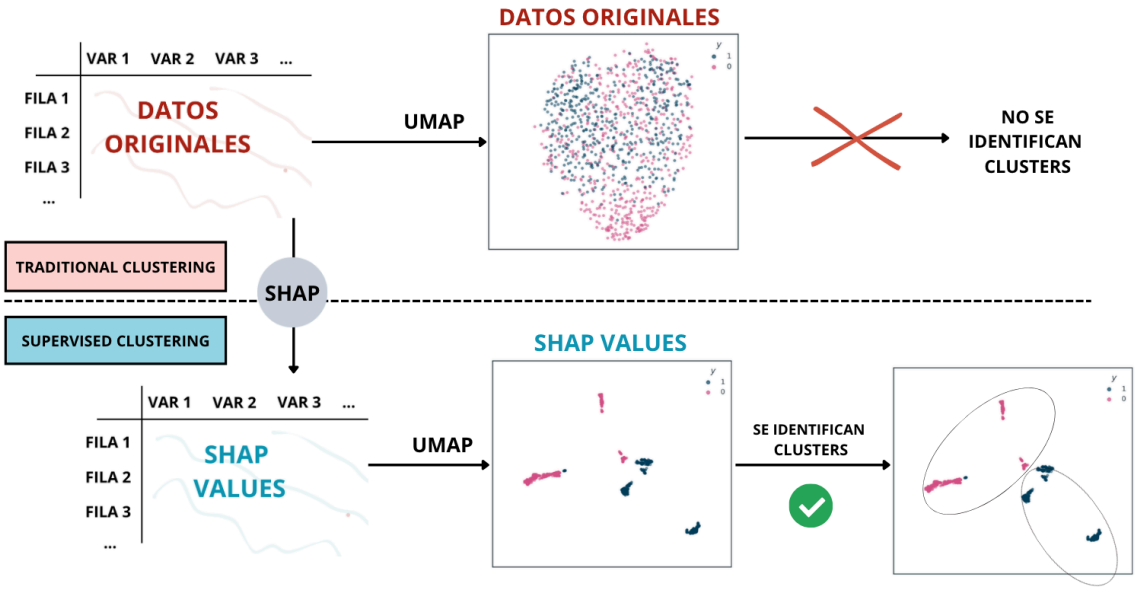

Figura 3.6: Diagrama conceptual comparando *clustering* tradicional y *clustering* supervisado. Elaboración propia, adaptada de (Cooper, 2022)

Esta representación no solo favorece la separación de los grupos, sino que también facilita su interpretación. Cada *cluster* puede analizarse a partir de las variables que contribuyen positiva o negativamente a la predicción del riesgo, lo que permite caracterizar perfiles diferenciados de solicitantes en términos de sus factores explicativos. En el contexto del riesgo crediticio, esta propiedad permite pasar de una segmentación basada en semejanzas descriptivas a una segmentación basada en mecanismos de decisión del modelo. En el presente trabajo, los valores SHAP cumplen, por tanto, una doble función. Por una parte, permiten interpretar las predicciones generadas por LightGBM en términos de los factores que aumentan o reducen el riesgo estimado. Por otra, constituyen una nueva representación de los solicitantes que servirá como base para aplicar posteriormente técnicas de reducción de dimensionalidad y *clustering*, con el fin de identificar perfiles homogéneos de riesgo crediticio definidos por factores explicativos.

3.3.3. Reducción de dimensionalidad mediante UMAP

Una vez obtenida la representación basada en valores SHAP, el siguiente paso consiste en proyectar dicha información en un espacio de menor dimensión que facilite la visualización de la estructura de los datos y la posterior identificación de patrones. Dado que el espacio SHAP conserva una dimensión asociada a cada variable explicativa del modelo, su análisis directo puede resultar complejo cuando el número de características es elevado. Por ello, se emplea una técnica de reducción de dimensionalidad que permite representar las ob-

servaciones en un espacio de baja dimensión, manteniendo en la medida de lo posible las relaciones de proximidad existentes en el espacio original. En este trabajo se utiliza UMAP, un método no lineal de reducción de dimensionalidad basado en principios de aprendizaje de variedades (*manifold learning*) y teoría de grafos. El algoritmo parte del supuesto de que los datos observados en alta dimensión se organizan sobre una estructura subyacente de menor dimensión. Su objetivo es aproximar dicha estructura mediante una representación reducida que conserve las relaciones de vecindad entre observaciones (with Josh Starmer, 2022). En el contexto del presente estudio, UMAP se aplica sobre el espacio SHAP, por lo que la proximidad entre solicitantes en la proyección resultante refleja similitudes en los factores que explican su riesgo estimado.

La Figura 3.7 resume de forma esquemática el funcionamiento general del algoritmo. En primer lugar, UMAP analiza las distancias entre observaciones en el espacio de alta dimensionalidad e identifica, para cada punto, un conjunto de observaciones vecinas. Esta vecindad se define mediante el hiperparámetro *n_neighbors*, que determina el número de puntos considerados próximos a cada observación. A partir de estas relaciones, el algoritmo construye un grafo ponderado en el que los nodos representan observaciones y las aristas expresan la intensidad de la relación de proximidad entre ellas. Posteriormente, dicho grafo se proyecta en un espacio de menor dimensión mediante un proceso de optimización que busca conservar, en la representación final, la estructura de vecindad estimada en el espacio original.

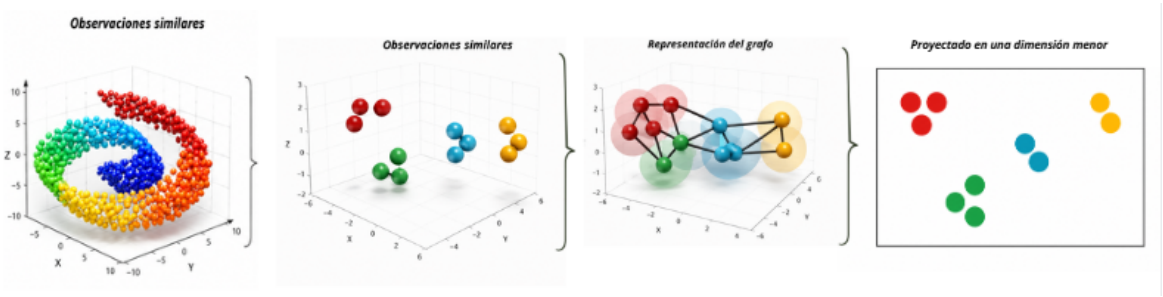

Figura 3.7: Esquema del funcionamiento del algoritmo UMAP. A partir de datos de alta dimensionalidad, se identifican relaciones de vecindad entre observaciones, se construye un grafo que captura la estructura local del conjunto de datos y, finalmente, dicha estructura se proyecta en un espacio de menor dimensión. Elaboración propia.

El parámetro *n_neighbors* regula el equilibrio entre la preservación de la estructura local y la estructura global de los datos. Valores bajos hacen que el algoritmo se concentre en relaciones muy próximas, lo que tiende a generar agrupamientos más compactos y diferenciados. En cambio, valores altos amplían la vecindad considerada y favorecen una representación más orientada a la organización global del conjunto de datos. Por su parte, el parámetro *min_dist* controla la distancia mínima permitida entre observaciones en la proyección final. Valores bajos de *min_dist* permiten que los puntos se sitúen más cerca entre sí, generando grupos densos. Valores más altos producen representaciones más dispersas y preservan mejor las

transiciones entre regiones del espacio (Twomey, 2025). La Figura 3.8 ilustra el efecto conjunto de estos parámetros. En el panel izquierdo, valores bajos de *n_neighbors* y *min_dist* producen grupos compactos, lo que permite resaltar relaciones locales entre observaciones. En el panel derecho, valores elevados de ambos parámetros generan una distribución más extendida, con mayor atención a la organización global del conjunto. Esta diferencia debe tenerse en cuenta en el análisis de perfiles de riesgo, ya que una parametrización demasiado local puede fragmentar excesivamente los grupos, mientras que una configuración demasiado global puede suavizar diferencias relevantes entre segmentos.

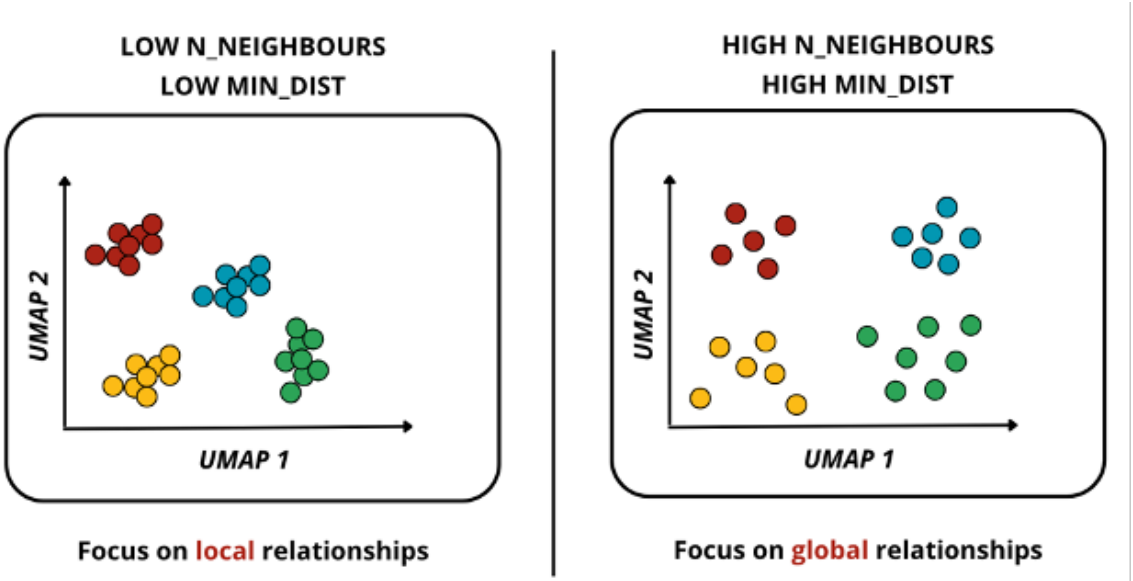

Figura 3.8: Influencia de los parámetros *n_neighbors* y *min_dist* en la proyección UMAP. Valores bajos enfatizan relaciones locales y generan grupos más compactos, mientras que valores elevados priorizan la organización global del conjunto de datos. Elaboración propia.

Como resultado de este proceso, las observaciones con patrones explicativos similares tienden a situarse próximas entre sí en el espacio reducido, mientras que aquellas con mecanismos de riesgo diferenciados se ubican en regiones más separadas. Esta proyección facilita el análisis visual de la estructura del espacio SHAP y prepara los datos para la aplicación posterior del algoritmo de *clustering*. En este sentido, UMAP permite explorar la organización de los solicitantes según las contribuciones explicativas generadas por el modelo supervisado.

A diferencia de técnicas lineales de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA, *Principal Component Analysis*), UMAP no se limita a construir combinaciones lineales de las variables originales. PCA proyecta los datos sobre ejes ortogonales que maximizan la varianza, por lo que resulta adecuado cuando la estructura del conjunto puede representarse mediante relaciones lineales. Sin embargo, en espacios explicativos como el generado por los valores SHAP, las diferencias entre observaciones pueden responder a patrones no lineales e interacciones complejas capturadas por el modelo supervisado. En este contexto, UMAP ofrece una alternativa más flexible al preservar relaciones de

vecindad y estructuras topológicas que no necesariamente quedan reflejadas en los primeros componentes principales. La Figura 3.9 ilustra esta diferencia. En la proyección obtenida mediante PCA, las observaciones aparecen distribuidas en una nube más continua, con mayor solapamiento entre grupos y menor separación visual entre patrones diferenciados. En cambio, la proyección mediante UMAP muestra agrupamientos más delimitados, lo que facilita la identificación de regiones del espacio con comportamientos explicativos similares. Esta diferencia no implica que UMAP sustituya a PCA en todos los contextos, sino que, para el objetivo de explorar subestructuras locales en el espacio SHAP, proporciona una representación más adecuada para el análisis visual y la posterior segmentación.

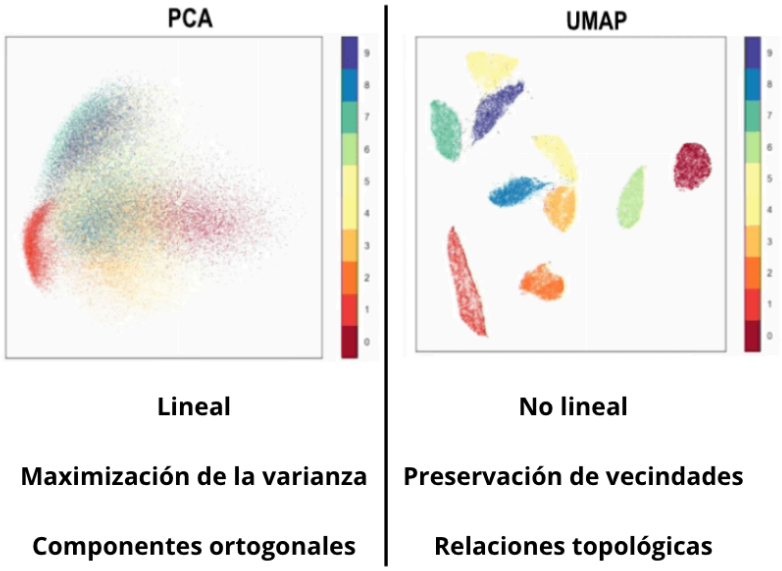

Figura 3.9: Comparación entre la proyección obtenida mediante PCA y UMAP sobre valores SHAP. UMAP muestra una separación más clara entre observaciones con patrones diferenciados, facilitando la identificación de subgrupos. Elaboración propia, adaptada de (Schröder, 2024)

En el presente trabajo, UMAP se utiliza como paso intermedio entre la representación explicativa generada por SHAP y la aplicación de DBSCAN. Su función consiste en reducir la complejidad del espacio explicativo y favorecer una disposición de las observaciones en la que los solicitantes con mecanismos de riesgo similares se sitúen en regiones próximas. Esta proyección facilita la identificación de perfiles diferenciados sin imponer una geometría estrictamente lineal al espacio de datos.

3.3.4. Algoritmo DBSCAN para la Segmentación de Solicitantes de Crédito

Una vez obtenida la representación reducida del espacio SHAP mediante UMAP, la siguiente etapa consiste en identificar grupos de observaciones que compartan patrones explicativos similares. Con este propósito se emplea DBSCAN, un algoritmo de agrupamiento basado en densidad propuesto por Ester, Kriegel, Sander, Xu, et al. (1996). A diferencia de métodos particionales como K-Means, DBSCAN no requiere especificar previamente el número de grupos presentes en los datos y permite identificar observaciones atípicas que no pertenecen a ninguna región densa. Estas características resultan adecuadas en el contexto de la segmentación supervisada, donde la distribución de los solicitantes en el espacio explicativo puede presentar formas irregulares, tamaños heterogéneos y zonas con distinta densidad. El principio de funcionamiento de DBSCAN consiste en detectar regiones del espacio donde las observaciones se encuentran concentradas y diferenciarlas de aquellas zonas con baja densidad. Para ello, el algoritmo utiliza dos hiperparámetros. El primero es ε (*epsilon*), que define el radio de búsqueda empleado para determinar la vecindad de cada observación. El segundo es *MinPts*, que establece el número mínimo de observaciones que deben encontrarse dentro de dicha vecindad para considerar que existe una región suficientemente densa. Formalmente, la vecindad ε asociada a una observación p se define como:

$$N_\varepsilon(p) = \{q \in D \mid d(p,q) \leq \varepsilon\} \tag{3.3}$$

donde D representa el conjunto de datos y $d(p,q)$ corresponde a la distancia entre las observaciones p y q. En este trabajo, dicha distancia se calcula sobre la representación reducida obtenida mediante UMAP a partir del espacio SHAP. Una observación se considera punto núcleo cuando el número de observaciones contenidas en su vecindad satisface la siguiente condición:

$$|N_\varepsilon(p)| \geq \textit{MinPts} \tag{3.4}$$

A partir de esta definición, las observaciones pueden clasificarse en tres categorías. Los puntos núcleo corresponden a observaciones situadas en regiones de alta densidad y constituyen el punto de partida para la formación de los grupos. Los puntos frontera se encuentran dentro de la vecindad de algún punto núcleo, aunque por sí mismos no alcanzan el umbral mínimo de densidad establecido por *MinPts*. Finalmente, las observaciones que no pertenecen a la vecindad de ningún punto núcleo se etiquetan como ruido y permanecen sin asignación a un grupo. Esta distinción permite que el algoritmo no fuerce la inclusión de casos aislados dentro de segmentos que no representan adecuadamente su comportamiento.

El procedimiento comienza seleccionando una observación no visitada y calculando su vecindad ε. Si dentro de esa vecindad se encuentran al menos *MinPts* observaciones, el pun-

to se clasifica como núcleo y se inicia la formación de un *cluster*. A partir de este núcleo, el algoritmo incorpora las observaciones situadas dentro de su vecindad y evalúa cada una de ellas para determinar si también cumple la condición de densidad mínima. Cuando una de estas observaciones es también un punto núcleo, su propia vecindad se añade al grupo, expandiendo progresivamente la región densa. Este proceso continúa hasta que no existen nuevos puntos núcleo conectados a la región analizada. En ese momento, el *cluster* queda cerrado y el algoritmo selecciona una nueva observación no visitada para repetir el procedimiento.

La Figura 3.10 muestra de forma esquemática las etapas del algoritmo. En primer lugar, se selecciona una observación y se calcula su vecindad a partir del radio ε. A continuación, se verifica si el número de observaciones dentro de dicha vecindad alcanza el umbral definido por *MinPts*. Cuando esta condición se cumple, el punto se clasifica como núcleo y se inicia la expansión del grupo mediante la incorporación de puntos conectados por densidad. Si una observación no cumple la condición de núcleo, pero se encuentra dentro de la vecindad de uno, se clasifica como punto frontera. Los puntos que no se encuentran conectados a ninguna región densa se consideran ruido (Ester et al., 1996; Starmer, 2018).

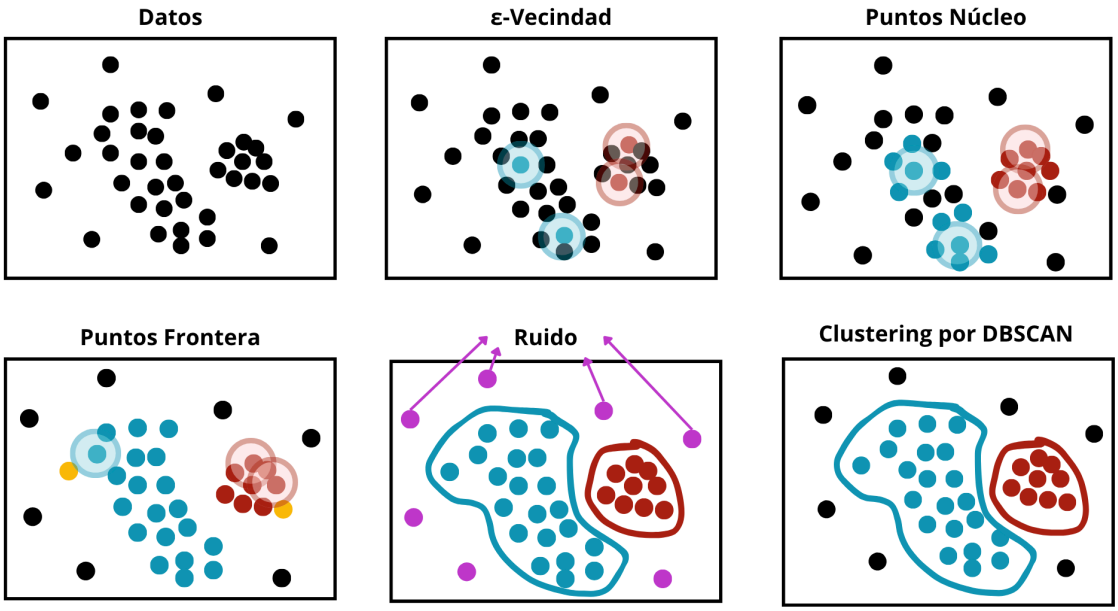

Figura 3.10: Proceso ilustrativo del funcionamiento de DBSCAN. Elaboración propia.

DBSCAN presenta ventajas metodológicas relevantes para este tipo de análisis, ya que permite detectar agrupamientos con formas no necesariamente esféricas y separar observaciones atípicas del resto de la estructura. Esta propiedad resulta adecuada cuando los perfiles de solicitantes no se distribuyen en grupos regulares o cuando existen casos con comportamientos explicativos alejados del patrón general. No obstante, su desempeño depende de la elección de los parámetros ε y *MinPts*. Un valor demasiado bajo de ε puede fragmentar regiones que deberían formar parte de un mismo grupo, mientras que un valor demasiado alto

puede unir segmentos distintos. De forma similar, un valor inadecuado de *MinPts* puede alterar la distinción entre puntos núcleo, frontera y ruido, especialmente cuando las regiones del espacio presentan densidades heterogéneas (Schubert, Sander, Ester, Kriegel, y Xu, 2017).

En este trabajo, DBSCAN se aplica sobre la representación de baja dimensión obtenida mediante UMAP a partir de los valores SHAP. Por tanto, la segmentación no se realiza en función de la semejanza directa entre las variables originales de los solicitantes, sino a partir de la similitud entre las contribuciones que dichas variables tienen sobre la predicción del modelo. Esta aproximación permite agrupar solicitantes que son evaluados de manera semejante por el modelo predictivo y que, por tanto, comparten mecanismos explicativos de riesgo. La utilización de un criterio basado en densidad permite identificar regiones diferenciadas dentro del espacio SHAP reducido y conservar como ruido aquellas observaciones que no se ajustan claramente a ningún segmento, lo que facilita una interpretación posterior más precisa de los perfiles crediticios obtenidos.

3.4. Evaluación y Validación de los clusters

Una vez aplicada la técnica de *clustering* sobre la representación generada a partir de los valores SHAP y proyectada mediante UMAP, resulta necesario evaluar la calidad de los grupos obtenidos. Para ello, se emplean métricas de validación interna, que permiten analizar la estructura del agrupamiento sin recurrir a etiquetas externas. En este trabajo se consideran el coeficiente de *Silhouette* y el índice de Davies–Bouldin, dos medidas utilizadas para examinar la relación entre la cohesión interna de los grupos y la separación existente entre ellos (Ansari, Azeem, Ahmed, y Babu, 2015; Davies y Bouldin, 1979; Rousseeuw, 1987). El coeficiente de *Silhouette* evalúa el grado de adecuación de cada observación al *cluster* al que ha sido asignada. Para una observación i, esta métrica compara la distancia media entre dicha observación y el resto de puntos de su propio grupo con la distancia media respecto al grupo alternativo más cercano. De este modo, permite valorar si una observación se encuentra bien integrada en su *cluster* o si se sitúa en una zona de frontera entre grupos. Formalmente, el coeficiente de *Silhouette* se define como:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}\{a(i),\, b(i)\}}, \tag{3.5}$$

donde $a(i)$ representa la distancia media entre la observación i y las demás observaciones pertenecientes a su mismo *cluster*, mientras que $b(i)$ corresponde a la menor distancia media entre i y las observaciones de otro *cluster*. El valor de $s(i)$ se encuentra en el intervalo $[-1, 1]$. Valores próximos a 1 indican que la observación está más cerca de los puntos de su propio grupo que de los pertenecientes a otros grupos. Valores cercanos a 0 sugieren una posición intermedia o fronteriza, mientras que valores negativos indican que la observación podría estar más próxima a un grupo distinto del asignado.

El índice de Davies-Bouldin ofrece una evaluación complementaria al considerar simultáneamente la dispersión interna de los grupos y la separación entre ellos. Esta medida compara, para cada *cluster*, su grado de compactación con la distancia que lo separa de los demás grupos. La expresión general del índice se define como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right), \tag{3.6}$$

donde k denota el número total de *clusters*, σ_i y σ_j representan la dispersión interna de los grupos i y j, respectivamente, y $d(c_i, c_j)$ corresponde a la distancia entre sus centroides. El término dentro del máximo compara, para cada par de grupos, la suma de sus dispersiones internas con la distancia que los separa. Por tanto, valores elevados del índice indican que existen grupos con alta dispersión interna o baja separación entre sí, mientras que valores más reducidos reflejan agrupamientos más compactos y mejor diferenciados. El uso conjunto de ambas métricas permite obtener una evaluación más completa de la estructura generada por DBSCAN. El coeficiente de *Silhouette* aporta información a nivel de observación y permite identificar puntos situados en zonas ambiguas entre grupos, mientras que el índice de Davies–Bouldin resume la relación global entre compactación y separación. En el contexto de este trabajo, estas medidas se emplean para valorar si los segmentos identificados en el espacio SHAP reducido presentan una estructura suficientemente coherente para ser interpretada posteriormente como perfiles diferenciados de riesgo crediticio.

Además de la evaluación cuantitativa, se realiza un análisis descriptivo de los *clusters* con el propósito de caracterizar los perfiles obtenidos. Para ello, se construye una tabla de resumen con los valores medios de las variables originales y de las contribuciones SHAP para cada grupo. Esta comparación permite identificar las características que diferencian a los segmentos y analizar si dichas diferencias son coherentes con la probabilidad de riesgo estimada por el modelo. Las variables con mayores contrastes entre grupos aportan información especialmente útil para describir los perfiles de solicitantes, mientras que los valores SHAP permiten interpretar la dirección y magnitud de su contribución al riesgo crediticio.Esta fase descriptiva complementa las métricas de validación interna. Mientras que el coeficiente de *Silhouette* y el índice de Davies-Bouldin permiten evaluar la estructura geométrica de los grupos, el análisis de medias y contribuciones explicativas permite valorar su significado desde una perspectiva financiera. De este modo, la calidad de la segmentación no se interpreta únicamente en términos de compactación y separación, sino también en función de la coherencia de los perfiles resultantes y de su capacidad para representar mecanismos diferenciados de riesgo.

En el contexto del presente trabajo, la combinación de DBSCAN con una representación basada en valores SHAP permite obtener segmentos definidos por la similitud en los factores que explican la predicción de incumplimiento. Al analizar los grupos en un espacio donde

cada observación está descrita por sus contribuciones explicativas, la segmentación resultante puede vincularse directamente con el comportamiento del modelo predictivo. Esto facilita la identificación de perfiles homogéneos de solicitantes y proporciona una base interpretable para apoyar la toma de decisiones en procesos de evaluación y gestión del riesgo crediticio.

3.5. Interpretación de los *clusters* mediante SkopeRules

Una vez identificados y evaluados los *clusters*, la última etapa metodológica consiste en caracterizar cada segmento mediante reglas interpretables formuladas sobre las variables originales del problema. Aunque los valores SHAP permiten explicar las predicciones del modelo supervisado y DBSCAN permite identificar grupos en el espacio explicativo, las coordenadas UMAP o las contribuciones SHAP no siempre ofrecen una descripción directamente operativa de los perfiles obtenidos. Por ello, resulta necesario traducir los segmentos a condiciones observables que permitan comprender las diferencias entre grupos en términos financieros y crediticios. Para este propósito se emplea *SkopeRules*, una herramienta orientada a la extracción de reglas lógicas a partir de datos etiquetados. Su funcionamiento se basa en la generación de múltiples reglas candidatas mediante modelos basados en árboles, normalmente entrenados bajo un esquema de *bagging*. A partir de estos árboles se extraen condiciones del tipo si-entonces, que posteriormente se filtran de acuerdo con criterios de calidad como precisión y cobertura. De este modo, el método conserva reglas con capacidad descriptiva suficiente y descarta condiciones redundantes o poco informativas (Game of Bits, 2019; Kleinbort, 2021).

En el presente trabajo, *SkopeRules* no se utiliza para explicar directamente el modelo LightGBM, ya interpretado mediante valores SHAP, sino para describir los *clusters* obtenidos en la etapa anterior. Para ello, se adopta una estrategia uno contra todos (*one-vs-all*), en la que cada *cluster* se define sucesivamente como clase positiva frente al resto de observaciones. Sobre esta variable objetivo binaria se entrenan los modelos internos de *SkopeRules*, utilizando como entrada las variables originales del conjunto de datos. Esta decisión permite que las reglas finales se expresen en términos directamente interpretables, como edad, ingresos, deuda o historial crediticio. La Figura 3.11 representa de forma esquemática este procedimiento. En primer lugar, a partir de las variables originales y de la etiqueta binaria de pertenencia al *cluster*, el algoritmo entrena múltiples árboles mediante *bagging*. De estos árboles se derivan numerosas reglas candidatas que describen condiciones asociadas al grupo analizado. Posteriormente, cada regla se evalúa mediante métricas de calidad, especialmente precisión y cobertura, lo que permite descartar aquellas que no caracterizan adecuadamente el segmento. En una fase final, las reglas redundantes o demasiado similares se eliminan, conservando un conjunto reducido de condiciones con mayor capacidad descriptiva.

El resultado de este proceso es un conjunto de reglas interpretables para cada *cluster*, ex-

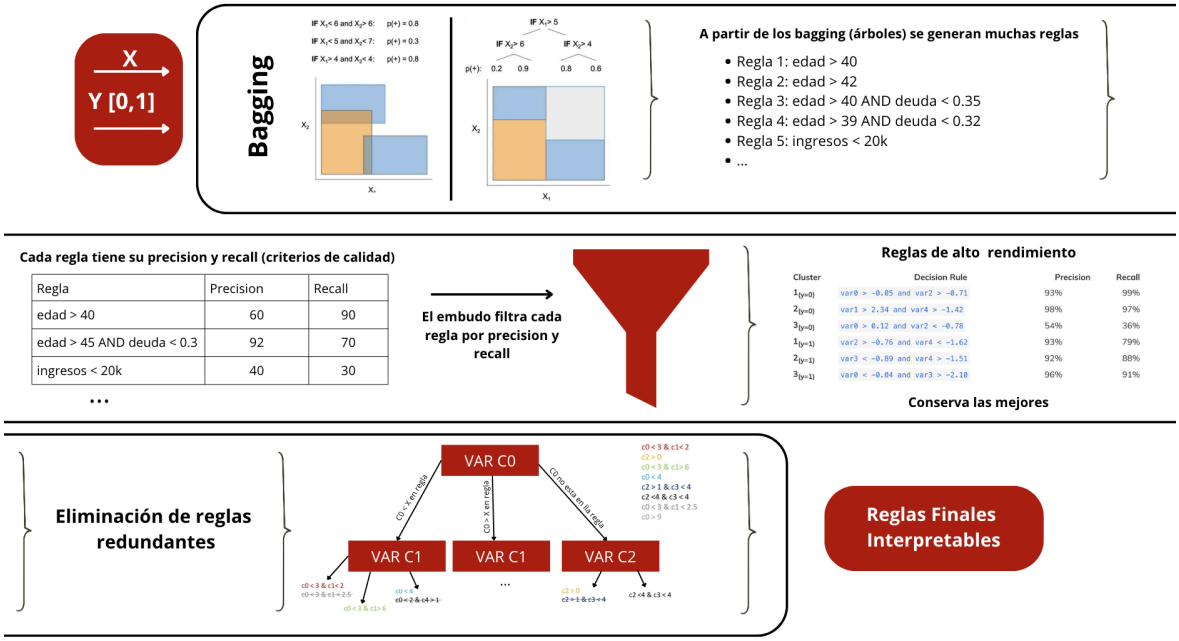

Figura 3.11: Esquema del proceso de generación, filtrado y selección de reglas interpretables mediante *SkopeRules*, desde la construcción de reglas a partir de modelos basados en árboles hasta la obtención de un conjunto final de reglas de alto rendimiento y sin redundancias. Elaboración propria, con adaptaciones de Kleinbort (2021).

presadas mediante combinaciones de umbrales y condiciones sobre las variables originales. Por ejemplo, una regla podría indicar que los solicitantes con una determinada combinación de edad, nivel de ingresos y ratio de deuda presentan una alta probabilidad de pertenecer a un segmento concreto. Estas reglas permiten describir los grupos obtenidos de forma operativa, facilitando la comprensión de los perfiles identificados y su posible uso en procesos de análisis, seguimiento o toma de decisiones en riesgo crediticio.

Este planteamiento permite caracterizar los *clusters* sin utilizar directamente los valores SHAP empleados en su identificación inicial (Cooper, 2022). De este modo, los grupos descubiertos en el espacio explicativo se traducen posteriormente en reglas formuladas sobre variables observables, lo que favorece una interpretación más cercana al contexto financiero y reduce el riesgo de circularidad analítica. Esta separación entre la etapa de descubrimiento de segmentos y la etapa de caracterización permite que los perfiles obtenidos no solo sean coherentes con el comportamiento del modelo, sino también comprensibles desde la perspectiva de las variables originales del solicitante. Con el fin de preservar la claridad de las reglas, se limita su complejidad mediante la restricción del número máximo de condiciones que pueden componer cada una (Cooper, 2022). Asimismo, se depuran reglas duplicadas o redundantes, de manera que el conjunto final conserve condiciones diferenciadas y operativamente interpretables (Game of Bits, 2019). Esta fase evita que la caracterización de los segmentos dependa de un número excesivo de reglas o de descripciones demasiado específicas, favoreciendo una lectura más sintética de los perfiles de riesgo.

Con esta etapa se completa el flujo metodológico del estudio. Las predicciones generadas por LightGBM se interpretan mediante valores SHAP, el espacio explicativo resultante se proyecta mediante UMAP, los segmentos se identifican mediante DBSCAN y, finalmente, las reglas extraídas con *SkopeRules* permiten describir dichos segmentos en términos de variables observables. En el Capítulo 4 se presentan los resultados derivados de la aplicación de esta metodología, incluyendo el análisis exploratorio de los datos, el desempeño del modelo predictivo, la segmentación obtenida y la interpretación de los perfiles de riesgo identificados.

Capítulo 4

Resultados

En el presente capítulo se presentan los resultados obtenidos tras la aplicación de la metodología descrita en el Capítulo 3. El análisis comprende la selección y el preprocesamiento del conjunto de datos de solicitantes de crédito, la estimación del riesgo mediante un modelo supervisado, la construcción del espacio explicativo a partir de valores SHAP, la segmentación mediante técnicas de *clustering* y la interpretación de los perfiles obtenidos mediante *SkopeRules*. El propósito del capítulo es identificar y analizar perfiles diferenciados de riesgo crediticio de forma interpretable y orientada a la toma de decisiones en el ámbito financiero. Todos los experimentos se han desarrollado con herramientas de código abierto. El código empleado, junto con los procedimientos necesarios para reproducir el análisis, se encuentra disponible en el repositorio de *GitHub* asociado al proyecto[1]. Esta disponibilidad contribuye a la trazabilidad del estudio, facilita la verificación de los resultados y permite futuras extensiones de la metodología propuesta.

4.1. Identificación y recolección de los datos

Para el desarrollo del presente estudio se seleccionó el conjunto de datos *German Credit Data with Risk*, disponible públicamente en la plataforma Kaggle (Kaggle, 2024). Este conjunto de datos contiene información correspondiente a 1.000 solicitantes de crédito y 10 variables que describen distintas características sociodemográficas y financieras de los individuos, junto con una variable objetivo denominada *risk*, que clasifica cada solicitud como de bajo o alto riesgo crediticio. La descripción de las variables incluidas se presenta en la Figura 4.1. Las variables disponibles abarcan aspectos relevantes para la evaluación crediticia, incluyendo información relacionada con la edad del solicitante, el importe y la duración del préstamo, la situación laboral y residencial, los niveles de ahorro y la finalidad del crédito solicitado. Esta combinación de características permite representar diferentes dimensiones

[1]Enlace al repositorio: `https://github.com/ssfa02/TFG---Credit-Risk-Profiling-via-Explainable-AI-and-Clustering/tree/main`

Variable ▲	Tipo	Descripción
Age	Cuantitativa discreta	Edad del solicitante en años.
Credit amount	Cuantitativa discreta	Cantidad de crédito solicitada (en marcos alemanes).
Duration	Cuantitativa discreta	Duración del préstamo en meses.
Housing	Cualitativa nominal	Situación de vivienda del solicitante (que indica si la vivienda es en propiedad, alquilada o cedida de forma gratuita).
Job	Cualitativa ordinal	Nivel de experiencia laboral del solicitante (0: no cualificado y no residente, 1: no cualificado y residente, 2: cualificado, 3: altamente cualificado).
Purpose	Cualitativa nominal	Finalidad del crédito (vehículo, mobiliario o equipamiento, aparatos electrónicos, electrodomésticos, reparaciones, educación, actividades empresariales u otros fines).
Risk	Cualitativa binaria	Evaluación del riesgo crediticio del préstamo (good, bad).
Saving accounts	Cualitativa nominal	Nivel de ahorro del solicitante (bajo, moderado, relativamente alto y alto).
Sex	Cualitativa binaria	Sexo del solicitante.

Table: Sofia G-M Tanure • Source: Kaggle • Created with Datawrapper

Figura 4.1: Descripción de las variables del conjunto de datos. Elaboración propia.

del perfil financiero de los individuos y analizar su relación con el riesgo de incumplimiento.

La selección de este conjunto de datos resulta adecuada para los objetivos del trabajo por varias razones. En primer lugar, incorpora una variable objetivo explícita que permite entrenar un modelo supervisado de clasificación y calcular posteriormente los valores SHAP utilizados en el proceso de segmentación. En segundo lugar, reúne variables frecuentemente empleadas en estudios de *credit scoring*, proporcionando un contexto apropiado para analizar los factores asociados al riesgo crediticio. Asimismo, las 1.000 observaciones disponibles proporcionan un volumen suficiente para entrenar el modelo predictivo, generar representaciones explicativas mediante SHAP y aplicar técnicas de agrupamiento, manteniendo al mismo tiempo una escala que facilita el análisis detallado de los resultados obtenidos. Finalmente, al tratarse de un conjunto de datos ampliamente utilizado en la literatura académica sobre riesgo crediticio, los resultados obtenidos pueden interpretarse dentro de un contexto de referencia conocido, facilitando la comparación con investigaciones previas y futuras aplicaciones de la metodología propuesta.

4.2. Preprocesamiento y Análisis Descriptivo de los Datos

Tras la selección del conjunto de datos *German Credit Data with Risk*, se realizó una revisión inicial de su estructura con el fin de preparar la información para las etapas posteriores de modelización, explicación y segmentación. Esta revisión permitió identificar la naturaleza de las variables disponibles, detectar elementos sin valor analítico y definir el tratamiento de los valores faltantes. El conjunto original contenía variables numéricas y categóricas asociadas al perfil sociodemográfico y financiero de los solicitantes. Dentro de la revisión inicial, se eliminó la variable *Unnamed: 0*, dado que correspondía a un identificador de registro y no aportaba información relevante para la predicción del riesgo crediticio.

Posteriormente, se evaluó la presencia de valores faltantes en las variables del conjunto de datos. Este análisis mostró que la variable *Checking account* presentaba un 39,4 % de observaciones ausentes, mientras que la variable *Saving accounts* contenía un 18,3 % de valores no informados. Para definir el tratamiento de estas variables, se adoptó un umbral del 30 % de valores faltantes como criterio de exclusión (Sangacha Ávila, 2024). Bajo este criterio, *Checking account* fue eliminada del análisis al superar dicho límite, ya que su conservación podía introducir un nivel elevado de incertidumbre en la interpretación y en el entrenamiento del modelo. En el caso de *Saving accounts*, el porcentaje de valores faltantes se situaba por debajo del umbral definido. Por esta razón, se optó por conservar la variable y crear una categoría adicional denominada *Missing* para las observaciones sin información disponible. Esta decisión permite mantener el tamaño muestral y, al mismo tiempo, preservar la posible información asociada a la ausencia del dato. En el contexto crediticio, la falta de información sobre una característica financiera puede reflejar un patrón específico del solicitante y, por tanto, no necesariamente debe tratarse como un error aleatorio. De este modo, la creación de una categoría explícita evita eliminar observaciones y permite que el modelo evalúe si la ausencia de dicha información contribuye a explicar el riesgo estimado.

A continuación, se analizó el comportamiento de las variables cuantitativas con el propósito de valorar la presencia de observaciones extremas y decidir si debían conservarse en el análisis. Para ello, se combinaron estadísticos descriptivos, percentiles superiores e histogramas. Este enfoque permite diferenciar entre posibles errores de registro y valores elevados que, aunque poco frecuentes, pueden ser coherentes con la naturaleza de las operaciones crediticias (Tiwari, Mehta, Jain, Tiwari, y Kanda, 2007). La Tabla 4.1 resume los principales estadísticos de las variables *Age*, *Credit amount* y *Duration*, mientras que la Figura 4.2 muestra sus distribuciones y la posición relativa del percentil 99 y del valor máximo.

En la variable *Age*, la mayor parte de las observaciones se concentra en edades jóvenes e intermedias, con una disminución progresiva de la frecuencia hacia edades superiores. El percentil 99 se sitúa en 67,01 años y el valor máximo alcanza los 75 años, lo que indica que las observaciones de mayor edad se encuentran en la cola superior de la distribución, pero no constituyen rupturas abruptas respecto al resto de la muestra. Por esta razón, se conside-

Tabla 4.1: Estadísticos descriptivos de las variables numéricas

Variable	Media	Mínimo	P25	P50	P75	P95	P99	Máximo
Age	35.55	19.00	27.00	33.00	42.00	60.00	67.01	75.00
Credit amount	3,271.26	250.00	1,365.50	2,319.50	3,972.25	9,162.70	14,180.39	18,424.00
Duration	20.90	4.00	12.00	18.00	24.00	48.00	60.00	72.00

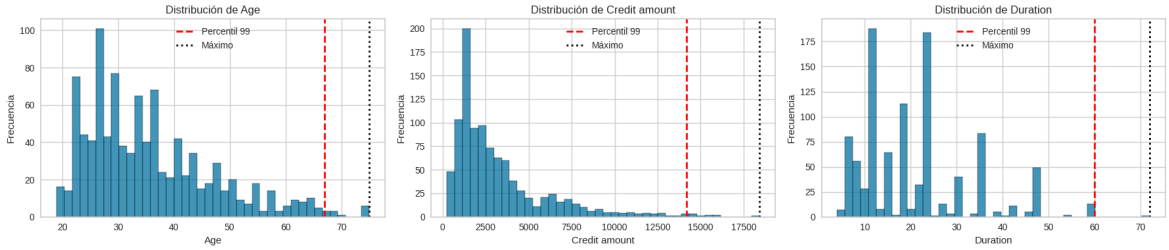

Figura 4.2: Distribución de las variables cuantitativas y análisis de valores extremos. Elaboración propia.

ran valores plausibles dentro del perfil de solicitantes analizado. La variable *Credit amount* presenta una asimetría positiva más pronunciada. La mediana se sitúa en 2.319,50 marcos alemanes, mientras que el percentil 99 alcanza los 14.180,39 y el valor máximo los 18.424. Esta diferencia refleja una dispersión elevada hacia importes superiores, propia de variables financieras en las que un número reducido de operaciones puede concentrar cuantías más altas. La progresión entre los percentiles superiores y el máximo no sugiere errores evidentes de registro, sino solicitudes de mayor importe que forman parte de la heterogeneidad crediticia del conjunto de datos. En el caso de *Duration*, la distribución responde al carácter discreto de los plazos de amortización. La mediana se sitúa en 18 meses, el percentil 99 en 60 meses y el máximo en 72 meses. Aunque existen préstamos de duración extensa, estos valores son compatibles con productos crediticios de largo plazo y no muestran un comportamiento anómalo que justifique su eliminación. Teniendo en cuenta estos análisis, se conservaron las observaciones extremas de las tres variables, dado que representan valores compatibles con la naturaleza de las operaciones crediticias analizadas y potencialmente informativos para el análisis del riesgo crediticio.

Tras completar las etapas de limpieza y tratamiento de valores faltantes, se procedió a la transformación de las variables categóricas para garantizar su correcta utilización en los modelos de aprendizaje automático empleados posteriormente. La variable *Job* representa distintos niveles de cualificación laboral y se encuentra codificada mediante valores comprendidos entre 0 y 3. Aunque estos valores se expresan numéricamente, su significado responde a una escala ordinal y no a una magnitud cuantitativa. Por esta razón, se conservó su codificación original, preservando la información asociada al orden de las categorías y evitando transformaciones adicionales que pudieran alterar su interpretación. Para el resto de las variables categóricas se aplicó la técnica de *One-Hot Encoding*, mediante la cual cada

categoría se transforma en una variable binaria independiente que toma valor 1 cuando la observación pertenece a dicha categoría y 0 en caso contrario. Con el fin de evitar problemas de dependencia lineal entre las variables generadas, se eliminó una categoría de referencia para cada variable original. De este modo, una variable con N categorías queda representada mediante $N-1$ variables binarias, manteniendo toda la información relevante y garantizando una representación adecuada para el entrenamiento del modelo (James, 2013).

Una vez completado el proceso de transformación de variables, se analizó la distribución de la variable objetivo *Risk*, encargada de identificar el nivel de riesgo crediticio asociado a cada solicitante. La categoría *good* agrupa aquellas observaciones consideradas de bajo riesgo, mientras que la categoría *bad* corresponde a solicitantes con una mayor probabilidad de incumplimiento. La Figura 4.3 muestra la distribución observada en el conjunto de datos. Del total de 1.000 registros disponibles, 700 observaciones pertenecen a la categoría *good*, representando el 70 % de la muestra, mientras que las 300 restantes corresponden a la categoría *bad*, equivalentes al 30 %. Esta distribución muestra la existencia de un desbalance moderado entre ambas clases. Aunque la diferencia de proporciones no alcanza niveles extremos, la presencia de una clase mayoritaria implica que el modelo dispondrá de un mayor número de ejemplos asociados a solicitantes de bajo riesgo durante el entrenamiento. Por ello, la evaluación posterior del rendimiento predictivo requerirá considerar métricas que permitan valorar adecuadamente la capacidad del modelo para identificar observaciones pertenecientes a la clase minoritaria, dado que estos casos son precisamente los de mayor interés desde la perspectiva de la gestión del riesgo crediticio.

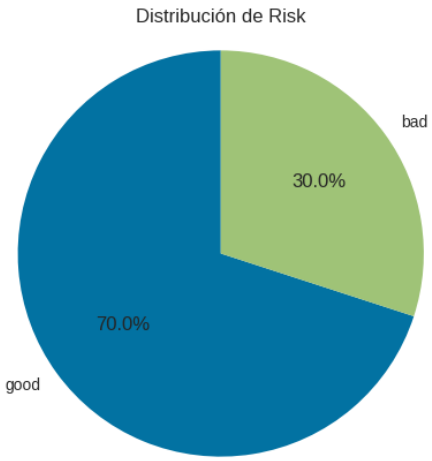

Figura 4.3: Distribución de la variable objetivo Risk. Elaboración propia.

Con el propósito de complementar el análisis de las distribuciones, se analizaron las variables numéricas mediante diagramas de caja, los cuales permiten resumir la posición, dispersión y variabilidad de los datos a través de la mediana, los cuartiles y la amplitud intercuartílica. Además, esta representación facilita la identificación de observaciones alejadas

del comportamiento predominante de cada variable. La Figura 4.4 muestra que la variable *Age* presenta una dispersión moderada, con una mediana situada en torno a los 33 años y un rango intercuartílico relativamente concentrado. Aunque se observan algunas observaciones por encima del límite superior del diagrama, estas corresponden a solicitantes de mayor edad que continúan dentro de valores plausibles para la población analizada.

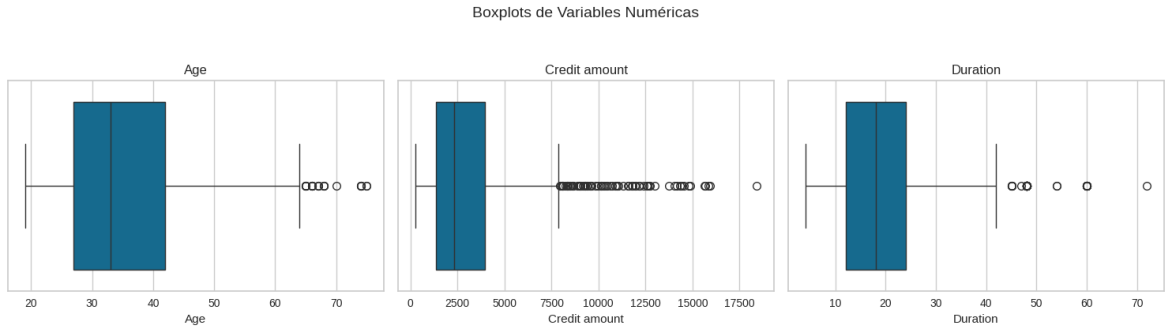

Figura 4.4: Diagramas de caja correspondientes a las variables numéricas *Age, Credit amount y Duration*.Elaboración propia.

En el caso de *Credit amount*, el rango intercuartílico se concentra en importes relativamente bajos en comparación con el máximo observado, lo que refleja una elevada dispersión hacia los valores superiores. El número de observaciones situadas por encima del límite superior es considerablemente mayor que en las restantes variables, lo que resulta coherente con la presencia de solicitudes de crédito de cuantía elevada. No obstante, estas observaciones mantienen una interpretación financiera razonable y no muestran evidencias de errores de registro. Por su parte, *Duration* presenta una dispersión intermedia y una distribución concentrada en los plazos más habituales de concesión de préstamos. Aunque se identifican algunas observaciones en la parte superior de la distribución, estas se corresponden con préstamos de duración más extensa y permanecen dentro de intervalos compatibles con productos crediticios de largo plazo. Los diagramas de caja confirman que las observaciones identificadas como potencialmente extremas responden principalmente a la propia heterogeneidad de las características crediticias analizadas y no a anomalías evidentes en los datos.

Por otro lado, las variables categóricas se analizaron mediante gráficos de proporciones que representan la distribución relativa de las categorías *good* y *bad* dentro de cada nivel de las variables consideradas. Este tipo de representación permite comparar el comportamiento del riesgo entre categorías independientemente de su frecuencia absoluta, facilitando la identificación de posibles asociaciones entre las características de los solicitantes y la variable objetivo. La Figura 4.5 muestra que la variable *Sex* presenta diferencias moderadas en la distribución del riesgo. En ambos grupos predomina la categoría *good*. Sin embargo, la proporción de observaciones clasificadas como *bad* es ligeramente superior entre las mujeres. Aunque la diferencia observada no es especialmente pronunciada, sugiere la existencia de patrones diferenciados que podrían ser considerados por el modelo durante el proceso

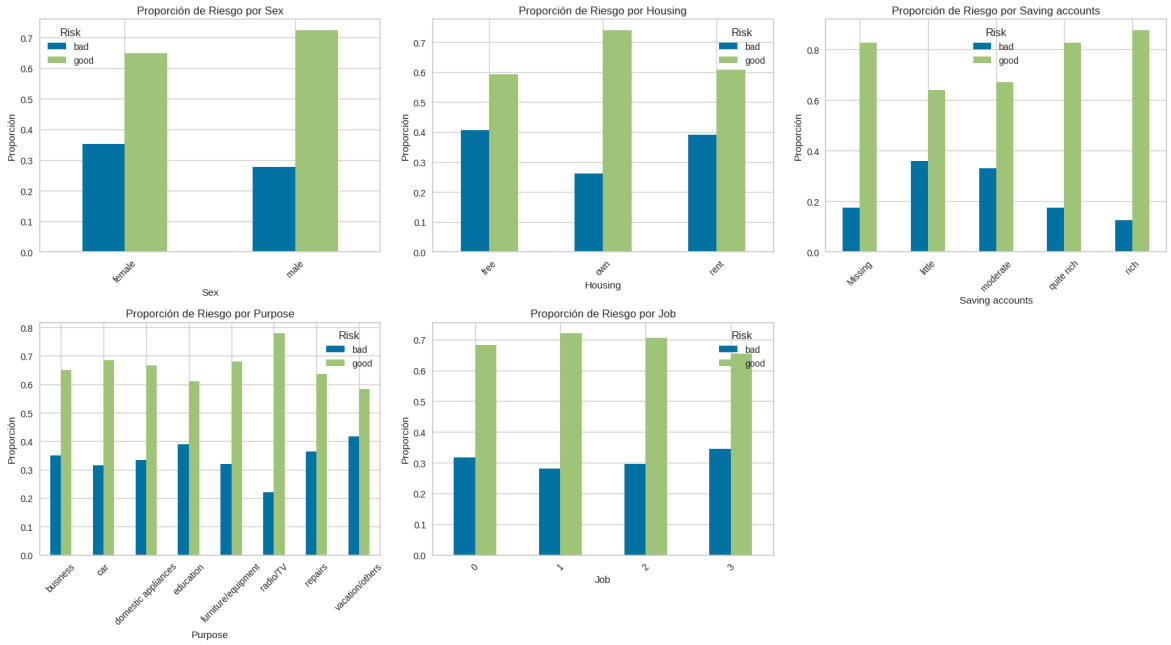

Figura 4.5: Proporción del riesgo según variables categóricas *Sex, Housing, Saving accounts, Jobs y Purpose*. Elaboración propia.

de predicción. En el caso de la variable *Job*, las proporciones de riesgo permanecen relativamente estables entre los distintos niveles de cualificación laboral. La categoría *good* es predominante en todos los grupos y las variaciones observadas en la proporción de riesgo son reducidas. Este comportamiento indica que la capacidad discriminativa de esta variable, considerada de forma aislada, es limitada, aunque ello no excluye posibles interacciones con otras características del solicitante.

Las diferencias más visibles se observan en las variables *Housing* y *Saving accounts*. Los solicitantes que disponen de vivienda en propiedad presentan una menor proporción de riesgo que aquellos que residen en viviendas alquiladas o cedidas, lo que sugiere una posible relación entre estabilidad patrimonial y capacidad de cumplimiento financiero. Por su parte, la variable *Saving accounts* muestra un patrón más consistente. A medida que aumenta el nivel de ahorro declarado, la proporción de observaciones clasificadas como *bad* disminuye progresivamente. Las categorías asociadas a mayores niveles de ahorro concentran los porcentajes más elevados de solicitantes clasificados como *good*, lo que sugiere que la capacidad de ahorro constituye una característica estrechamente vinculada al riesgo crediticio. La variable *Purpose* presenta una mayor heterogeneidad entre categorías. Algunas finalidades del crédito, como la adquisición de radio o televisión, muestran una proporción relativamente reducida de solicitantes clasificados como *bad*, mientras que otras categorías, como las asociadas a vacaciones u otros fines diversos, registran porcentajes de riesgo más elevados. Esta variabilidad sugiere que el destino de los fondos puede estar relacionado con diferentes perfiles financieros y niveles de solvencia, aportando información potencialmente relevan-

te para la caracterización posterior de los segmentos identificados mediante el proceso de agrupamiento.

Por último, se analizó la relación lineal entre las variables mediante una matriz de correlación, representada en la Figura 4.6. Esta matriz permite evaluar posibles asociaciones entre las variables numéricas y las variables categóricas transformadas mediante *One-Hot Encoding*, así como detectar posibles relaciones de redundancia que pudieran afectar a la interpretación del modelo. En términos generales, no se observan correlaciones lineales de elevada magnitud entre la mayoría de las variables explicativas, lo que indica que el conjunto de datos no presenta una estructura marcada de multicolinealidad. La relación más destacada se observa entre *Credit amount* y *Duration*, con una correlación positiva moderada. Este resultado resulta coherente con la lógica financiera del crédito, ya que préstamos de mayor cuantía suelen asociarse a plazos de amortización más extensos. También se identifican correlaciones negativas entre algunas categorías derivadas de una misma variable original, como ocurre en *Housing* o *Saving accounts*, lo que responde al propio proceso de codificación binaria y no debe interpretarse como una relación sustantiva entre variables independientes.

Una vez completado el análisis exploratorio, se realizó la partición del conjunto de datos en subconjuntos de entrenamiento y prueba. El 80 % de las observaciones, equivalente a 800 registros, se destinó al entrenamiento del modelo, mientras que el 20 % restante, correspondiente a 200 registros, se reservó para evaluar su desempeño sobre datos no utilizados durante el aprendizaje. Esta división permitió obtener los subconjuntos de variables predictoras y variable objetivo tanto para entrenamiento como para prueba, manteniendo una separación clara entre la fase de ajuste del modelo y la fase de evaluación. De este modo, el rendimiento predictivo posterior puede analizarse en términos de capacidad de generalización y no únicamente de ajuste sobre los datos empleados para entrenar el algoritmo.

4.3. *Clustering* supervisado con valores SHAP

Una vez completadas las fases de preprocesamiento y análisis exploratorio, se aplicó la metodología de segmentación supervisada basada en valores SHAP. Este procedimiento parte del entrenamiento de un modelo predictivo de riesgo crediticio y utiliza sus explicaciones individuales como representación alternativa de los solicitantes. Sobre dicha representación se desarrollan posteriormente las etapas de reducción de dimensionalidad y *clustering*, con el fin de identificar perfiles de riesgo definidos por los factores que explican las predicciones del modelo.

4.3.1. Modelo Predictivo LightGBM

La primera etapa del proceso consistió en entrenar un modelo de clasificación binaria mediante LightGBM. Este algoritmo se seleccionó por su capacidad para capturar relaciones

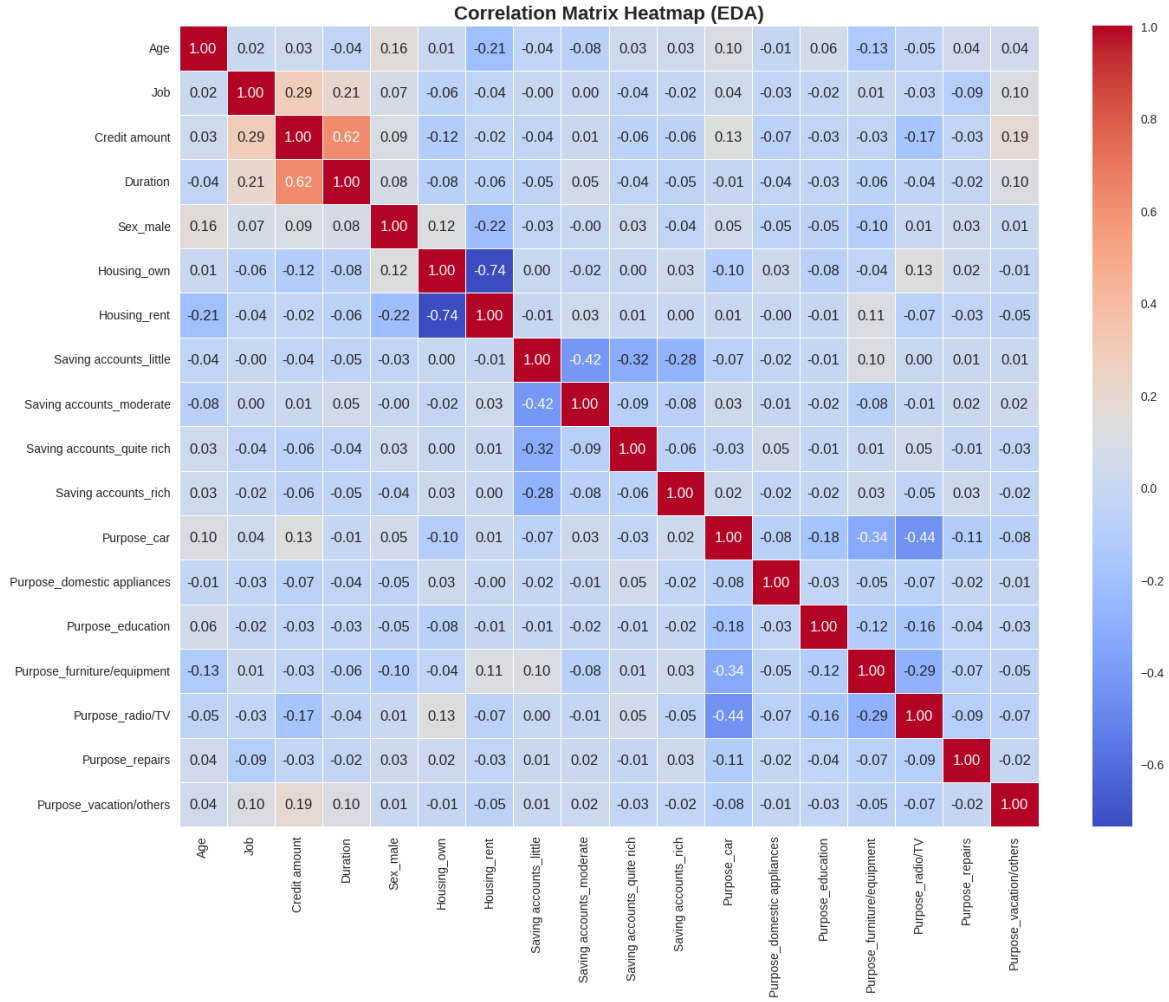

Figura 4.6: Matriz de correlación entre las variables numéricas y las variables categóricas codificadas mediante *One-Hot Encoding*.Elaboración propia.

no lineales e interacciones entre variables, así como por su compatibilidad con el cálculo posterior de valores SHAP. En el contexto de este trabajo, el modelo no se emplea únicamente como herramienta predictiva, sino como base para construir un espacio explicativo que permita segmentar a los solicitantes de acuerdo con los factores que determinan su riesgo estimado. Para obtener una configuración adecuada del modelo, se realizó una optimización de hiperparámetros mediante *Randomized Search*. Esta estrategia permite explorar distintas combinaciones de parámetros de forma eficiente, reduciendo el coste computacional frente a una búsqueda exhaustiva mediante *Grid Search*. La optimización se centró en cinco hiperparámetros de LightGBM. El parámetro *n_estimators* controla el número de árboles incorporados al proceso de *boosting*, *learning_rate* regula la contribución de cada árbol al modelo agregado, *max_depth* limita la profundidad máxima de los árboles, *num_leaves* controla la complejidad de la estructura generada, y *min_child_samples* establece el número mínimo de observaciones requerido en cada hoja. El ajuste conjunto de estos parámetros permite regular

la flexibilidad del modelo y reducir el riesgo de sobreajuste asociado al crecimiento *leaf-wise* característico de LightGBM.

La búsqueda de hiperparámetros se combinó con validación cruzada de cinco particiones. Para cada configuración evaluada se calculó el valor medio del área bajo la curva ROC (ROC-AUC, *Receiver Operating Characteristic–Area Under the Curve*) en los conjuntos de validación, seleccionándose finalmente la combinación con mejor rendimiento promedio. Esta métrica se empleó debido a su utilidad en problemas de clasificación binaria con clases moderadamente desbalanceadas, ya que permite evaluar la capacidad del modelo para discriminar entre solicitantes clasificados como *good* y *bad* sin depender de un único umbral de clasificación.

El modelo LightGBM obtuvo un desempeño moderado sobre el conjunto de prueba, con un valor de ROC-AUC de 0,686 y una exactitud global del 74 %. Como referencia, una regla basada en la clase mayoritaria, consistente en clasificar todas las observaciones como *good*, alcanzaría una exactitud del 70 %, dado que esta clase concentra 140 de las 200 observaciones del conjunto de prueba. El modelo supera esta línea base en cuatro puntos porcentuales y presenta un ROC-AUC superior a 0,5, lo que indica cierta capacidad para discriminar entre solicitantes de bajo y alto riesgo. Aunque el desempeño predictivo no es elevado, resulta suficiente para construir una base explicativa inicial mediante valores SHAP, siempre interpretando los resultados posteriores a la luz de esta limitación.

La matriz de confusión presentada en la Tabla 4.2 permite analizar con mayor detalle el comportamiento del modelo LightGBM sobre el conjunto de prueba. En ella se observa que el modelo clasifica correctamente 125 solicitantes *good* y 23 solicitantes *bad*. Asimismo, se registran 15 casos *good* clasificados erróneamente como *bad* y 37 casos *bad* clasificados como *good*.

Tabla 4.2: Matriz de confusión del modelo LightGBM

	Predicción	
Valor real	**good**	**bad**
good	125	15
bad	37	23

Para facilitar la interpretación de estos resultados, la Tabla 4.3 resume el significado de cada componente de la matriz de confusión, considerando la clase *bad* como la clase positiva.

Tabla 4.3: Interpretación de los componentes de la matriz de confusión

Métrica	**Valor**	**Interpretación**
TN	125	Solicitantes realmente *good* que el modelo predice como *good*.
FP	15	Solicitantes realmente *good* que el modelo predice como *bad*.
FN	37	Solicitantes realmente *bad* que el modelo predice como *good*.
TP	23	Solicitantes realmente *bad* que el modelo predice como *bad*.

A partir de estos valores, la sensibilidad del modelo alcanza un valor de 0,383, lo que indica que identifica correctamente el 38,3 % de los solicitantes de mayor riesgo. Por su parte, la especificidad se sitúa en 0,893, reflejando que el modelo reconoce correctamente el 89,3 % de los solicitantes de bajo riesgo. Estos resultados muestran una asimetría en el comportamiento del modelo, con mayor capacidad para identificar la clase mayoritaria *good* que para detectar la clase *bad*. Este patrón es coherente con la distribución del conjunto de datos, donde los solicitantes clasificados como *good* representan el 70 % de las observaciones. Desde la perspectiva del riesgo crediticio, la menor sensibilidad hacia la clase *bad* debe tenerse en cuenta en la interpretación de los resultados, ya que los falsos negativos corresponden a solicitantes de mayor riesgo que el modelo clasifica como de bajo riesgo. No obstante, dado que el objetivo posterior no es únicamente la clasificación, sino la construcción de una representación explicativa mediante SHAP, el modelo proporciona una base razonable para analizar los factores que influyen en las predicciones y avanzar hacia la segmentación supervisada.

La curva ROC presentada en la Figura 4.7 complementa el análisis anterior al evaluar el comportamiento del modelo para distintos umbrales de clasificación. En el eje horizontal se representa la tasa de falsos positivos, correspondiente a la proporción de solicitantes *good* clasificados erróneamente como *bad*. En el eje vertical se muestra la tasa de verdaderos positivos o sensibilidad, que indica la proporción de solicitantes *bad* correctamente identificados por el modelo. La línea diagonal representa el comportamiento esperado de un clasificador aleatorio, mientras que una curva más próxima al vértice superior izquierdo indica una mayor capacidad para discriminar entre ambas clases. El área bajo la curva ROC alcanza un valor de 0,686, lo que confirma que el modelo posee una capacidad discriminativa superior a la clasificación aleatoria, aunque de magnitud moderada. Este resultado es coherente con lo observado en la matriz de confusión, donde el modelo muestra un mejor desempeño en la identificación de solicitantes *good* que en la detección de la clase *bad*. Por tanto, el modelo permite distinguir parcialmente entre perfiles de bajo y alto riesgo, aunque mantiene limitaciones en la identificación de solicitantes con mayor probabilidad de incumplimiento.

A pesar de estas limitaciones, el rendimiento obtenido proporciona una base razonable para calcular valores SHAP y analizar los factores que influyen en las predicciones del modelo. En este sentido, la utilidad del modelo no se restringe a su capacidad clasificatoria, sino que se extiende a su función como punto de partida para construir una representación explicativa de los solicitantes, sobre la cual se desarrollará el análisis de segmentación supervisada.

4.3.2. Construcción del espacio explicativo mediante valores SHAP

Tras evaluar el desempeño del modelo LightGBM, se procedió al cálculo de los valores SHAP con el objetivo de interpretar la contribución de cada variable a las predicciones generadas. En esta fase, el propósito del análisis no consiste en volver a evaluar la capacidad predictiva del modelo, sino en construir una representación explicativa del conjunto de datos

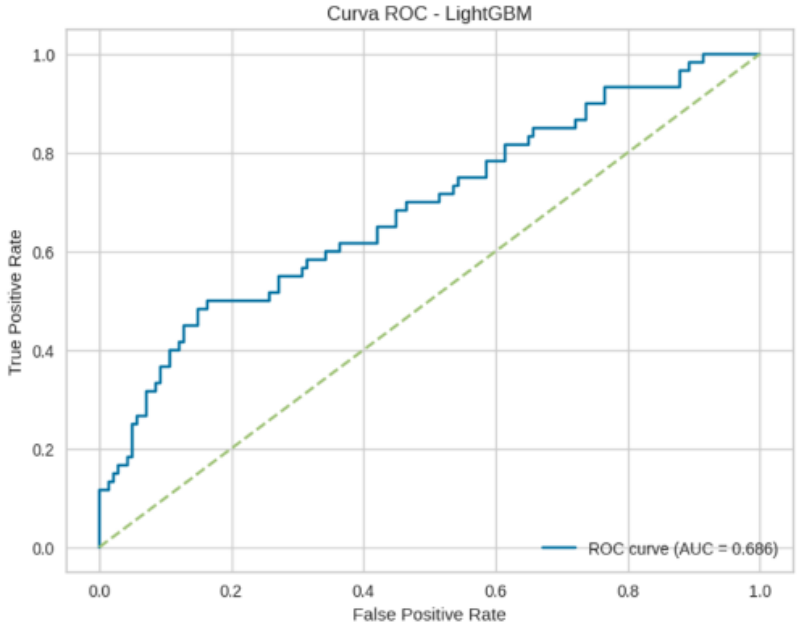

Figura 4.7: Curva ROC del modelo LightGBM. Elaboración propia.

que pueda emplearse posteriormente en la reducción de dimensionalidad y en la segmentación supervisada. Para ello, se utilizó la configuración de hiperparámetros seleccionada previamente mediante el proceso de optimización. A partir del mejor estimador obtenido, se generó una copia que conserva la estructura y los hiperparámetros del modelo, pero elimina cualquier aprendizaje previo realizado sobre los datos de entrenamiento. Posteriormente, este nuevo modelo se reentrenó utilizando la totalidad de las observaciones disponibles, es decir, el conjunto completo de variables predictoras X y la variable objetivo y. Esta decisión responde al objetivo de construir una representación explicativa global del conjunto de datos. Mientras que la partición entrenamiento-prueba resulta necesaria para evaluar la capacidad de generalización del modelo, el cálculo de valores SHAP orientado a la segmentación requiere disponer de explicaciones para todas las observaciones. Al reentrenar el modelo final sobre el conjunto completo, las contribuciones SHAP se estiman a partir de toda la información disponible, lo que permite obtener una representación más completa y estable de los factores que influyen en la predicción de riesgo.

Una vez calculados los valores SHAP asociados a la clase positiva, definida como *bad* = 1, se analizó la relación entre las variables originales y sus correspondientes contribuciones explicativas. Esta clase se considera de especial interés en el contexto del riesgo crediticio, ya que representa a los solicitantes con mayor probabilidad de incumplimiento. Por lo tanto, los valores SHAP positivos indican contribuciones que desplazan la predicción hacia la clase *bad*, mientras que los valores negativos reflejan contribuciones que reducen la probabilidad estimada de pertenecer a dicha clase. La Figura 4.8 presenta, mediante diagramas de violín, una comparación entre la distribución de las variables originales estandarizadas y la distribu-

ción de sus valores SHAP. Las distribuciones en azul corresponden a los valores observados de cada variable una vez estandarizados, mientras que las distribuciones en rosa representan la contribución de esas mismas variables a la predicción del riesgo. Esta comparación permite distinguir entre la variabilidad descriptiva de una característica y su efecto explicativo dentro del modelo. Una variable puede presentar una amplia dispersión en sus valores originales y, sin embargo, tener una contribución limitada a la predicción. Del mismo modo, una variable con menor variabilidad descriptiva puede mostrar valores SHAP más concentrados en direcciones específicas si el modelo la utiliza de forma consistente para aumentar o reducir el riesgo estimado.

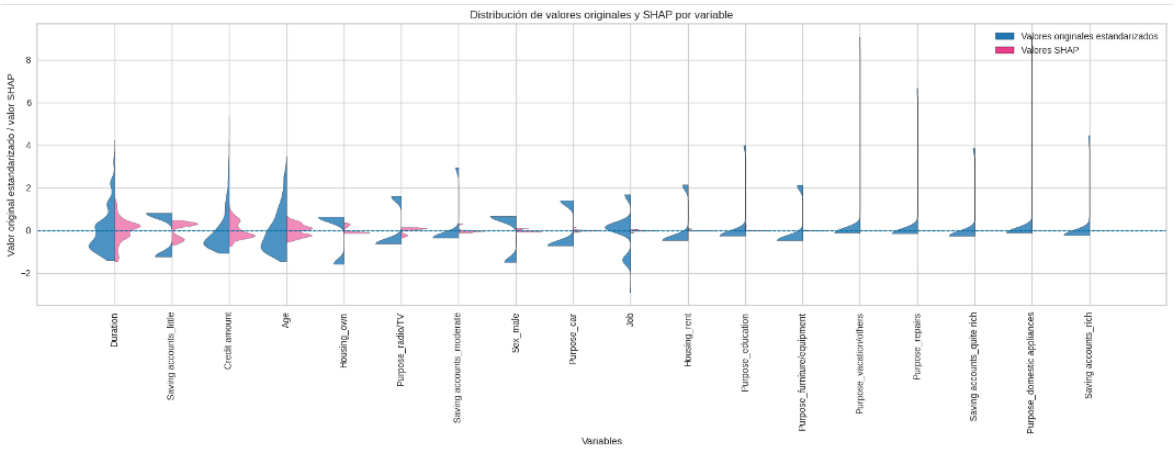

Figura 4.8: Comparación de las distribuciones de variables originales y sus correspondientes valores SHAP mediante diagramas de violín.Elaboración propia.

La comparación presentada en la Figura 4.8 muestra que las distribuciones de las variables originales estandarizadas y las de sus valores SHAP no son equivalentes. Esta diferencia indica que el modelo no utiliza las variables únicamente en función de su variabilidad observada, sino según su contribución a la predicción de la clase *bad*. En particular, variables como *Credit amount*, *Duration* y *Age* presentan contribuciones SHAP más visibles, lo que sugiere que el modelo emplea estas características para ajustar la probabilidad estimada de riesgo. En cambio, algunas variables categóricas codificadas muestran distribuciones originales muy concentradas y contribuciones SHAP más localizadas, coherentes con su naturaleza binaria. Este resultado justifica el uso del espacio SHAP como representación explicativa para las etapas posteriores del análisis. A diferencia del espacio original, donde las observaciones se describen por sus características observadas, el espacio SHAP organiza la información según el efecto que dichas características tienen sobre la predicción del riesgo.

Este análisis se complementa con la Figura 4.9, que presenta la importancia global de las variables a partir del valor medio absoluto de SHAP. Esta medida resume la magnitud promedio de la contribución de cada predictor a las predicciones del modelo, independientemente de si dicha contribución aumenta o reduce la probabilidad estimada de pertenecer a la

clase *bad*. Por tanto, el gráfico permite ordenar las variables según su influencia global en el modelo, aunque no informa por sí solo sobre la dirección del efecto.

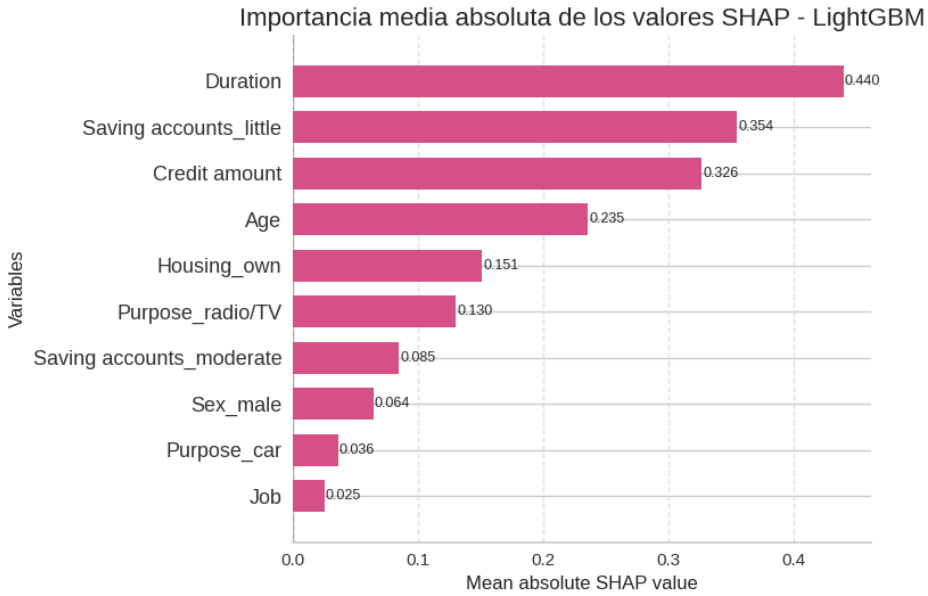

Figura 4.9: Importancia global de las diez variables principales según el valor SHAP medio absoluto. Elaboración propia.

Los resultados muestran que *Duration*, *Saving accounts_little* y *Credit amount* son las tres variables con mayor influencia global sobre la predicción del riesgo crediticio. La variable *Duration* ocupa la primera posición, con un valor SHAP medio absoluto de 0,440, lo que indica que la duración del préstamo es el predictor que genera, en promedio, las mayores variaciones en la salida del modelo. Este resultado es coherente con la lógica financiera del problema, ya que los préstamos de mayor plazo suelen incorporar un horizonte de incertidumbre más amplio. En segundo lugar, *Saving accounts_little* presenta un valor medio absoluto de 0,354, lo que indica que la presencia de esta categoría genera, en promedio, variaciones relevantes en la predicción del modelo y contribuye de manera sustancial a diferenciar entre solicitantes con distintos niveles de riesgo crediticio. Finalmente, *Credit amount*, con un valor de 0,326, se sitúa como el tercer predictor más influyente, en línea con la relevancia del importe solicitado en la evaluación de la exposición crediticia.

Después de los tres primeros predictores, *Age*, *Housing_own* y *Purpose_radio/TV* ocupan posiciones intermedias. Su presencia en este segundo nivel de importancia sugiere que el modelo no se apoya únicamente en variables estrictamente financieras, sino también en características asociadas a estabilidad personal, situación residencial y finalidad del crédito. En particular, la inclusión de *Housing_own* indica que la situación patrimonial del solicitante aporta información adicional al riesgo estimado, mientras que *Purpose_radio/TV* muestra que el destino del préstamo también introduce diferencias en la evaluación realizada por el modelo. En cambio, variables como *Saving accounts_moderate*, *Sex_male*, *Purpose_car* y

Job presentan valores SHAP medios absolutos más reducidos, lo que sugiere una menor influencia promedio en las predicciones generadas por el modelo.

La importancia global basada en valores SHAP muestra que el modelo concentra buena parte de su capacidad explicativa en un número limitado de variables, principalmente relacionadas con la duración del préstamo, el nivel de ahorro y el importe solicitado. Esta estructura resulta coherente con la lógica financiera del riesgo crediticio, ya que combina información sobre el horizonte temporal de la deuda, la capacidad financiera previa del solicitante y la magnitud de la exposición crediticia. A su vez, estos resultados respaldan el uso del espacio SHAP en la etapa posterior de segmentación supervisada, dado que dicho espacio permite representar a cada solicitante según los factores que modifican su probabilidad estimada de riesgo, y no únicamente según los valores observados de sus características originales.

4.3.3. Segmentación supervisada explicable mediante UMAP y DBSCAN

Con el propósito de identificar grupos de solicitantes caracterizados por patrones explicativos similares, se aplicó de forma secuencial la reducción de dimensionalidad mediante UMAP y, posteriormente, el algoritmo de agrupamiento DBSCAN sobre la representación obtenida. El proceso de segmentación se realizó sobre el espacio construido a partir de los valores SHAP y no sobre las variables originales del conjunto de datos. En consecuencia, la proximidad entre observaciones refleja similitudes en la forma en que las distintas variables contribuyen a la predicción del riesgo crediticio, y no únicamente semejanzas en los valores observados de dichas variables. Este planteamiento sigue la lógica del *supervised clustering* propuesta por Cooper (2022), donde la segmentación se lleva a cabo sobre una representación explicativa derivada de un modelo supervisado previamente entrenado. Dado que la representación generada por UMAP constituye la entrada directa de DBSCAN, la selección de hiperparámetros se abordó de manera conjunta. En lugar de optimizar ambos algoritmos por separado, se evaluaron simultáneamente diferentes combinaciones de los parámetros *n_neighbors* y *min_dist* de UMAP, junto con los parámetros *eps* y *min_samples* de DBSCAN. Esta estrategia permite identificar configuraciones que no solo producen proyecciones adecuadas del espacio SHAP, sino que además favorecen la obtención de agrupamientos estables y bien definidos.

La búsqueda se realizó mediante *Grid Search* sobre un conjunto acotado de combinaciones de hiperparámetros. Cada configuración fue evaluada utilizando el coeficiente de *Silhouette* y el índice de *Davies–Bouldin*, complementados con el número de *clusters* detectados y el porcentaje de observaciones clasificadas como ruido. Este último criterio resulta especialmente relevante en el contexto del presente estudio, ya que el objetivo consiste en caracterizar perfiles de riesgo para la totalidad de los solicitantes. Por esta razón, únicamente se consideraron aquellas configuraciones que generaban al menos dos grupos y no dejaban

observaciones sin asignar. La combinación seleccionada correspondió a *n_neighbors* = 50, *min_dist* = 0.01, *eps* = 2.0 y *min_samples* = 10. Con esta configuración, DBSCAN identificó tres *clusters*, sin observaciones clasificadas como ruido, alcanzando un coeficiente de *Silhouette* de 0.802 y un índice de *Davies-Bouldin* de 0.307. Estos valores muestran una elevada cohesión interna y una separación sustancial entre los grupos en la representación bidimensional generada por UMAP a partir de los valores SHAP.

Con la configuración seleccionada, la proyección UMAP se calculó sobre dos representaciones distintas del conjunto de datos. La primera se construyó a partir de las variables originales, mientras que la segunda utilizó los valores SHAP obtenidos del modelo LightGBM. En la representación basada en variables originales, las características numéricas continuas fueron previamente estandarizadas con el fin de evitar que las diferencias de escala influyeran de forma desproporcionada en el cálculo de distancias y, por tanto, en la estructura de la proyección. Los valores SHAP no requirieron esta transformación, ya que cada variable queda expresada en términos de su contribución a la predicción del modelo. El uso de la misma configuración de hiperparámetros en ambas proyecciones permite atribuir las diferencias observadas a la naturaleza de la representación utilizada y no a cambios en el procedimiento de reducción de dimensionalidad. La Figura 4.10 muestra diferencias claras entre ambas representaciones. En el panel izquierdo, correspondiente a las variables originales, las observaciones asociadas a las clases *good* y *bad* aparecen ampliamente mezcladas, formando una estructura continua en la que no se distinguen grupos claramente delimitados. Esta disposición indica que las relaciones de proximidad construidas a partir de los valores observados no reflejan de forma directa los patrones utilizados por el modelo para estimar el riesgo crediticio.

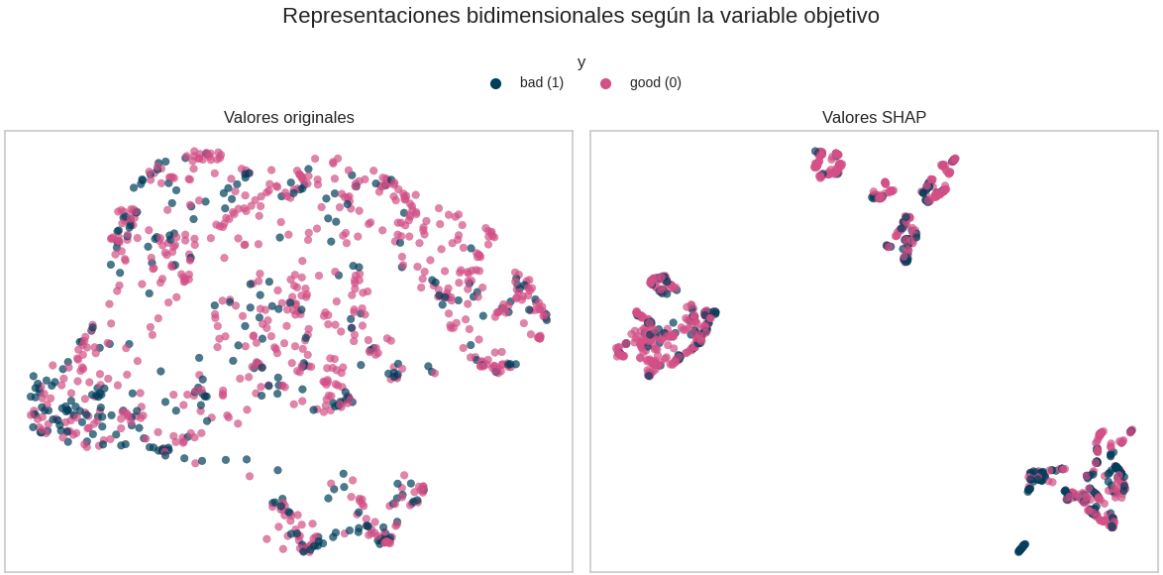

Figura 4.10: Representación UMAP del espacio original de variables y del espacio explicativo SHAP según la clase de riesgo. Elaboración propia.

En el panel derecho, correspondiente a los valores SHAP, se observan regiones más compactas y espacialmente separadas. Dado que cada dimensión del espacio SHAP representa la contribución de una variable a la predicción del modelo, la cercanía entre observaciones refleja similitudes en sus patrones explicativos y no únicamente semejanzas en sus características originales. Este resultado es coherente con el planteamiento de (Cooper, 2022), según el cual los valores SHAP proporcionan una representación adecuada para el *clustering* supervisado al expresar las variables en función de su influencia sobre la salida del modelo. Aunque las clases *good* y *bad* continúan apareciendo mezcladas dentro de algunas regiones, este comportamiento no contradice el objetivo del análisis. En este enfoque, la finalidad no es separar visualmente las clases de riesgo, sino identificar grupos de solicitantes cuyas predicciones se explican mediante patrones similares de contribución de variables. Por tanto, los grupos obtenidos deben interpretarse como perfiles explicativos de riesgo crediticio y no como clases homogéneas de riesgo observado. Esta diferencia justifica el uso del espacio SHAP como base para la segmentación posterior mediante DBSCAN.

A partir de la representación bidimensional obtenida mediante UMAP, se aplicó DBSCAN sobre el espacio proyectado de valores SHAP utilizando los hiperparámetros seleccionados previamente, con *eps* = 2.0 y *min_samples* = 10. El resultado de este procedimiento se presenta en la Figura 4.11, donde se identifican tres agrupaciones densas y diferenciadas en el espacio UMAP, sin observaciones clasificadas como ruido.

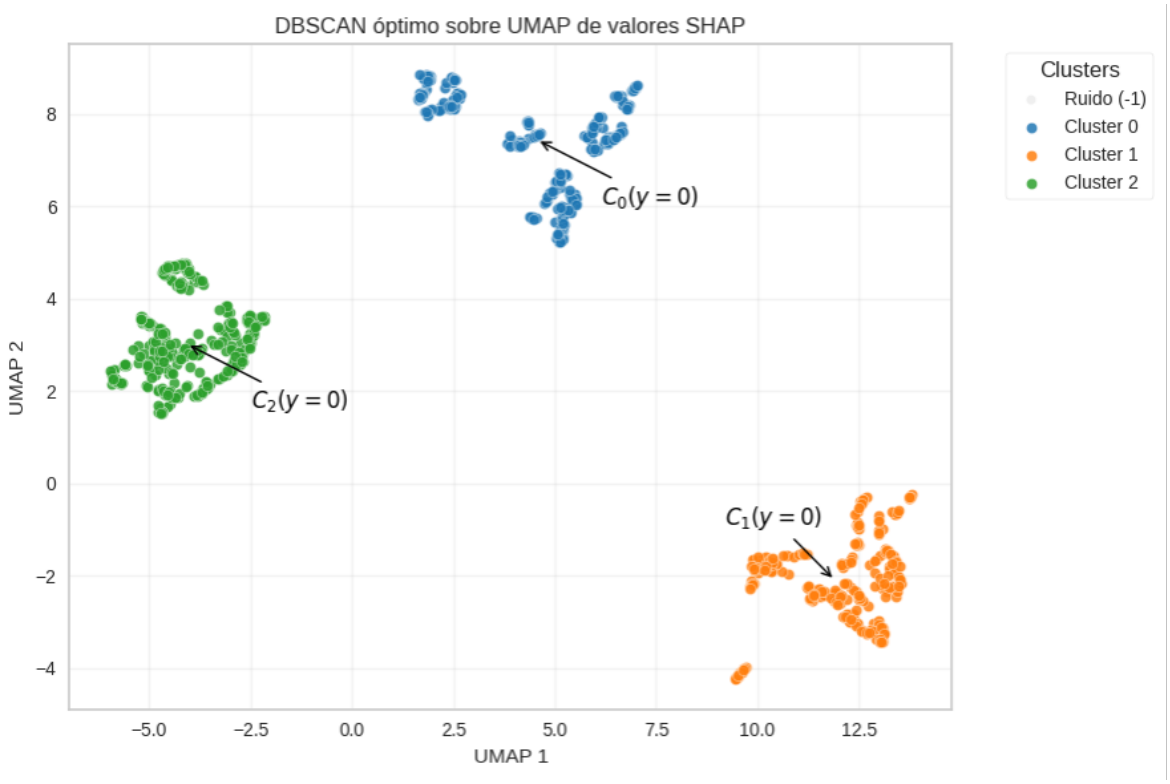

Figura 4.11: Resultado del clustering DBSCAN aplicado sobre la proyección UMAP de los valores SHAP, mostrando tres agrupaciones densas y bien diferenciadas.Elaboración propia.

La separación visual observada entre los tres grupos es coherente con los valores obtenidos en las métricas internas de validación, con un coeficiente de *Silhouette* de 0.802 y un índice de *Davies-Bouldin* de 0.307. Estos resultados indican que, en la representación bidimensional generada a partir de los valores SHAP, las observaciones dentro de cada grupo presentan una elevada proximidad entre sí y una separación considerable respecto a las observaciones de los demás grupos. Por tanto, la estructura identificada por DBSCAN no responde únicamente a una partición visual, sino también a una organización cuantitativamente consistente del espacio proyectado. En los tres *clusters* predomina la clase *good* $(y = 0)$, lo que resulta coherente con la distribución original de la variable objetivo, donde dicha clase representa la mayor parte de las observaciones. No obstante, la ausencia de un grupo claramente dominado por la clase *bad* $(y = 1)$ limita la posibilidad de interpretar alguno de los *clusters* como un segmento directo de alto riesgo observado. Esta limitación debe entenderse teniendo en cuenta el enfoque metodológico empleado. DBSCAN no agrupa a los solicitantes según su etiqueta final de riesgo, sino según la similitud de sus contribuciones SHAP en la proyección UMAP. Por tanto, los grupos obtenidos representan perfiles explicativos definidos por patrones similares de contribución de variables, y no clases homogéneas de riesgo.

Esta interpretación resulta especialmente importante para el análisis posterior. Aunque algunos *clusters* puedan presentar una mayor proporción relativa de solicitantes *bad*, su valor analítico no depende únicamente de esa composición, sino de la identificación de los factores que explican la predicción dentro de cada grupo. Por tanto, la segmentación obtenida debe evaluarse junto con la caracterización descriptiva de los *clusters* y las reglas interpretables extraídas posteriormente mediante *SkopeRules*. Con el fin de analizar si los *clusters* identificados en el espacio SHAP mantienen correspondencia con patrones observables en las variables originales, la Figura 4.12 proyecta sobre el mismo *embedding* UMAP los valores de las diez variables con mayor importancia global en el modelo. En cada subgráfico, los tonos más claros representan valores más elevados de la variable correspondiente, mientras que los tonos más oscuros indican valores más reducidos. Esta representación permite evaluar si las regiones detectadas mediante DBSCAN se asocian con diferencias interpretables en las características originales de los solicitantes. Este análisis es coherente con el planteamiento de Cooper (2022), según el cual los *clusters* generados a partir de valores SHAP pueden conservar relaciones significativas con las variables originales cuando estas variables participan de manera relevante en la predicción del modelo.

La Figura 4.12 muestra que la correspondencia entre los *clusters* y las variables originales no es homogénea para todos los predictores. La variable *Saving accounts_little* presenta uno de los patrones espaciales más definidos. El grupo situado en la región izquierda del *embedding* está compuesto mayoritariamente por observaciones que no pertenecen a esta categoría, mientras que los grupos ubicados en las regiones superior e inferior derecha concentran principalmente solicitantes asociados a un bajo nivel de ahorro. Esta distribución sugiere que la presencia de *Saving accounts_little* contribuye de forma clara a diferenciar parte de los

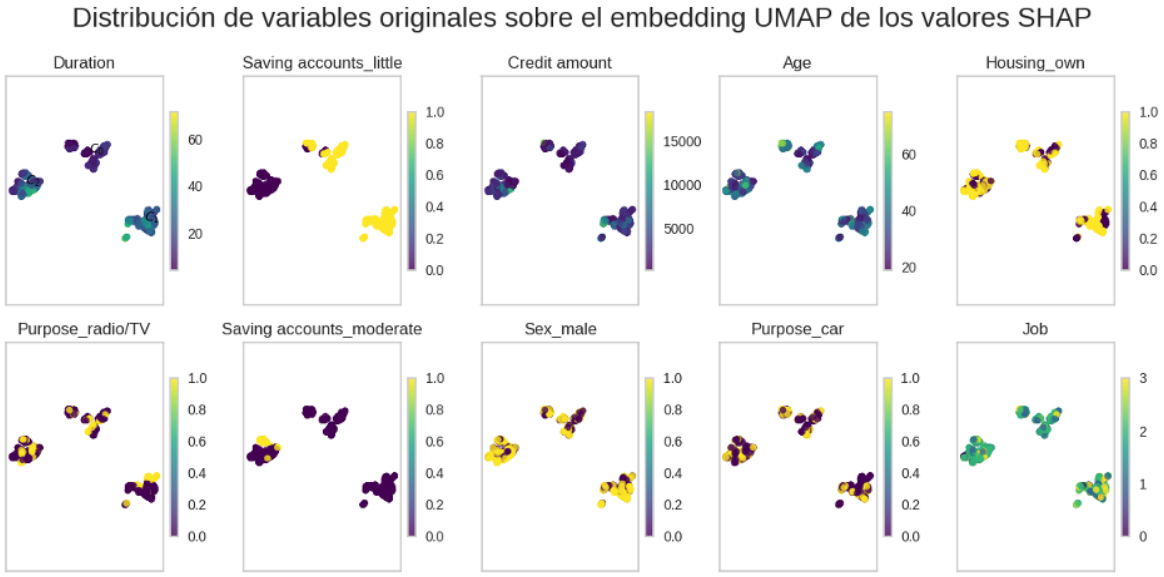

Figura 4.12: Distribución de las 10 principales variables originales del modelo sobre el embedding UMAP de los valores SHAP. Cada subgráfico representa una variable distinta, donde los tonos más claros indican valores más elevados y los tonos más oscuros valores más reducidos.Elaboración propia.

perfiles identificados, en línea con su elevada importancia global en el análisis SHAP. La variable *Duration* también muestra variaciones espaciales apreciables, aunque con un patrón menos claro. Los valores bajos y moderados aparecen distribuidos en los tres grupos, mientras que las duraciones más elevadas se concentran en determinadas zonas del *embedding*, especialmente en el grupo situado en la región inferior derecha. Este comportamiento indica que la duración del préstamo participa en la diferenciación de los perfiles, pero no define por sí sola la estructura completa de los grupos. Su efecto debe interpretarse dentro de una lógica multivariable, donde los *clusters* emergen de la combinación de distintas contribuciones SHAP.

En el caso de *Credit amount*, la distribución es más heterogénea. Los tres grupos contienen principalmente importes bajos o moderados, y los valores más elevados aparecen en un número reducido de observaciones. Por ello, aunque el importe del crédito se sitúa entre las variables con mayor importancia global, su efecto no se traduce en una separación visual clara entre *clusters* cuando se analiza de forma aislada. Algo similar ocurre con *Age*, cuyos valores aparecen distribuidos en las distintas regiones del *embedding* sin una asociación evidente con un grupo específico. Estas variables contribuyen a las predicciones del modelo, pero su influencia parece operar en combinación con otros factores más que a través de patrones espaciales simples. Así, la figura permite observar que algunos predictores relevantes conservan una relación visible con la estructura de los *clusters*, especialmente *Saving accounts_little* y, en menor medida, *Duration*. Sin embargo, no todas las variables importantes generan separaciones claras por sí mismas. Este resultado es coherente con la naturaleza del

espacio SHAP, donde cada observación queda representada por el conjunto de contribuciones que explican su predicción y no por una única característica dominante. Por ello, la segmentación obtenida debe interpretarse como una estructura explicativa multivariable. La proyección también muestra que los grupos identificados no son únicamente regiones geométricas del *embedding*, sino segmentos que guardan relación con características observables de los solicitantes. No obstante, la interpretación visual debe complementarse con análisis complementarios, ya que UMAP reduce la dimensionalidad del espacio SHAP y puede simplificar relaciones complejas.

4.3.4. Análisis e interpretación de los *clusters* mediante *SkopeRules*

Con el objetivo de formalizar la interpretación de los grupos identificados en el espacio SHAP, se aplicó *SkopeRules* sobre las variables originales del conjunto de datos. En esta etapa, el propósito no consiste en predecir nuevamente el riesgo crediticio, sino en caracterizar la pertenencia a cada *cluster* mediante reglas comprensibles formuladas a partir de características observables de los solicitantes. Por ello, las variables de entrada corresponden a las variables originales, mientras que la variable objetivo utilizada por *SkopeRules* indica la pertenencia de cada observación al *cluster* analizado. En la configuración del algoritmo se fijó *max_depth* = 2, parámetro que limita la profundidad máxima de los árboles internos y restringe la complejidad de las reglas generadas. Esta decisión permite obtener reglas breves, formadas por un máximo de dos condiciones, favoreciendo su interpretación. Como comprobación adicional, se evaluó una configuración con *max_depth* = 3 y se observó que las reglas obtenidas no variaban respecto a la configuración inicial, lo que sugiere que los *clusters* pueden describirse mediante condiciones simples sobre las variables originales.

La Tabla 4.4 muestra que las reglas extraídas se basan exclusivamente en dos variables, *Duration* y *Saving accounts_little*. Este resultado es coherente con el análisis SHAP previo, donde ambas variables aparecían entre los predictores con mayor importancia global. En términos interpretativos, los *clusters* quedan diferenciados principalmente por la duración del préstamo y por la presencia o ausencia de un bajo nivel de ahorro, dos factores directamente vinculados con la evaluación financiera del riesgo.

Tabla 4.4: Reglas extraídas mediante *SkopeRules* para la caracterización de los *clusters*

Cluster	**Regla**	**Precisión**	***Recall***
0	*Duration* $\leq 15{,}5$ y *Saving accounts_little* $> 0{,}5$	1.0000	0.8419
1	*Duration* $> 15{,}5$ y *Saving accounts_little* $> 0{,}5$	1.0000	1.0000
2	*Duration* $> 10{,}5$ y *Saving accounts_little* $\leq 0{,}5$	0.9926	0.9582

Los tres *clusters* presentan niveles de precisión muy elevados. La precisión indica la proporción de observaciones seleccionadas por la regla que pertenecen efectivamente al *cluster*

correspondiente. Por tanto, valores próximos a 1 reflejan que las reglas identifican cada grupo con muy pocos errores de asignación. El *recall*, en cambio, mide la proporción de observaciones del *cluster* que quedan cubiertas por la regla. Desde esta perspectiva, el *Cluster 1* presenta una cobertura completa, mientras que el *Cluster 2* alcanza una cobertura muy elevada. El *Cluster 0*, con un *recall* de 0.8419, queda representado por una regla precisa, aunque no cubre la totalidad de sus observaciones, lo que sugiere una mayor heterogeneidad interna. Estos resultados no constituyen por sí solos evidencia de sobreajuste, especialmente porque las reglas son simples, reproducen variables relevantes según SHAP y se alinean con la separación observada previamente en el espacio UMAP. No obstante, deben interpretarse como reglas descriptivas de los perfiles identificados, no como un nuevo modelo predictivo de riesgo. Su valor reside en traducir los *clusters* obtenidos en el espacio explicativo a condiciones observables y operativas sobre las variables originales.

En términos de interpretación de los perfiles identificados, la diferenciación entre los *clusters* se apoya principalmente en dos variables: la duración del crédito (*Duration*) y la presencia de un bajo nivel de ahorro (*Saving accounts_little*). La Figura 4.13 respalda visualmente esta separación. El *Cluster 0* concentra préstamos de corta duración, con una mediana cercana a los 10 meses y una dispersión reducida. En cambio, los *Clusters 1* y *2* presentan duraciones superiores, con medianas próximas a los 24 meses y una variabilidad más amplia.

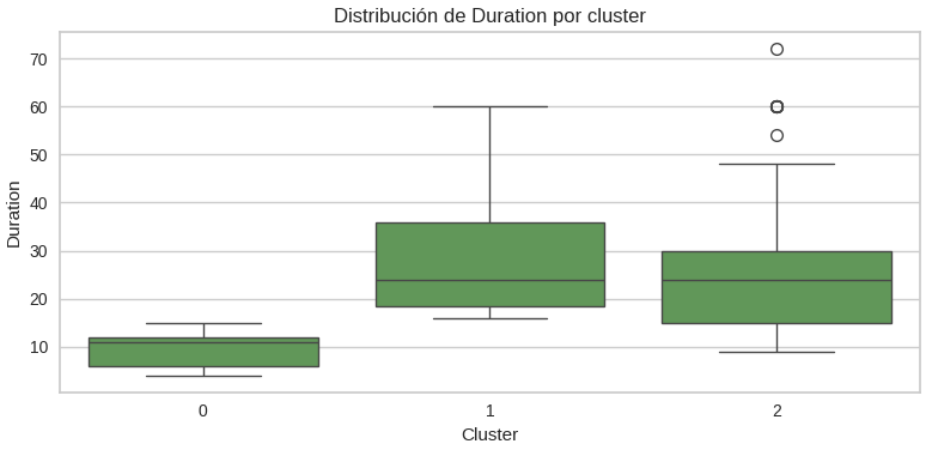

Figura 4.13: Distribución de la variable *Duration* por *cluster*. Elaboración propia.

Esta distribución es coherente con el umbral identificado por *SkopeRules*, donde *Duration* = 15.5 actúa como punto de separación entre el *Cluster 0* y los restantes grupos. Así, el *Cluster 0* queda asociado a solicitantes con préstamos de menor plazo y bajo nivel de ahorro, mientras que los *Clusters 1* y *2* reúnen perfiles con duraciones más extensas, diferenciándose entre sí por la presencia o ausencia de la categoría *Saving accounts_little*. La mayor dispersión observada en los *Clusters 1* y *2*, especialmente por la presencia de préstamos de larga duración, sugiere una mayor heterogeneidad interna en estos grupos frente al perfil más compacto del *Cluster 0*.

La variable *Saving accounts_little* funciona como segundo eje de diferenciación entre los perfiles identificados. La Figura 4.14 muestra que el *Cluster 1* está compuesto íntegramente por solicitantes con bajo nivel de ahorro. El *Cluster 0* presenta también una fuerte concentración de esta categoría, aunque incorpora una proporción reducida de solicitantes pertenecientes a otros niveles de ahorro. En contraste, el *Cluster 2* no contiene observaciones clasificadas como *Saving accounts_little* y presenta una composición más diversificada, con presencia de las categorías *moderate*, *quite rich*, *rich* y *Missing*.

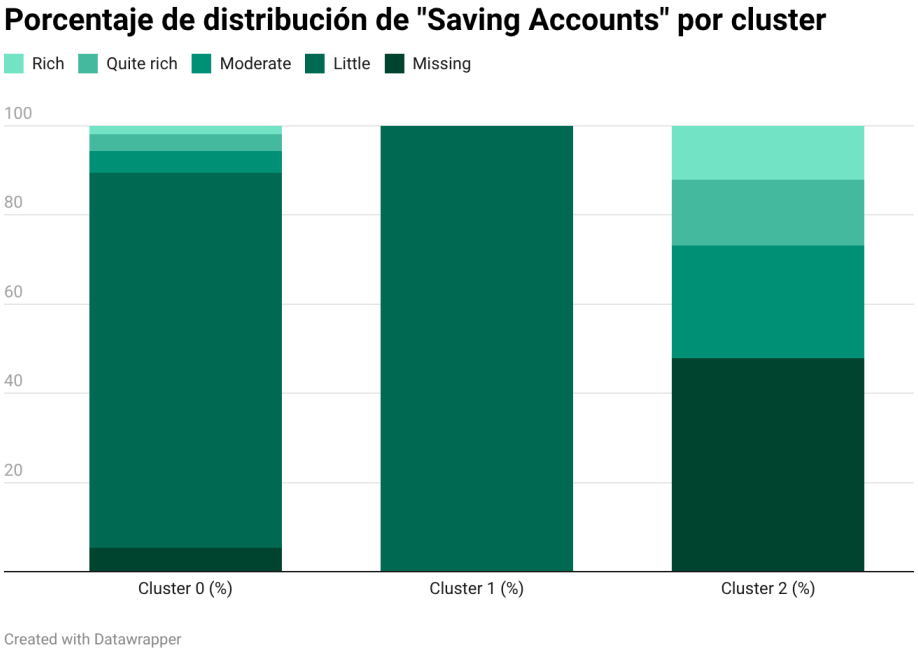

Figura 4.14: Distribución porcentual de la variable *Saving accounts* en cada *cluster*. Elaboración propia.

Este patrón permite diferenciar dos tipos de perfiles. Por un lado, los *Clusters 0* y *1* agrupan mayoritariamente solicitantes con bajo nivel de ahorro, aunque se distinguen entre sí por la duración del crédito. Por otro lado, el *Cluster 2* concentra solicitantes sin bajo nivel de ahorro, lo que sugiere una situación financiera más favorable o, al menos, menos asociada a la ausencia de ahorro disponible. Esta diferencia resulta coherente con el análisis SHAP, donde *Saving accounts_little* aparecía entre las variables con mayor importancia global. La comparación entre los *Clusters 0* y *1* resulta especialmente informativa. Ambos grupos están formados mayoritariamente por solicitantes con bajo nivel de ahorro, pero presentan tasas de incumplimiento diferentes, del 20,3 % y 46,4 %, respectivamente. Esta diferencia indica que el bajo nivel de ahorro, considerado de forma aislada, no explica completamente la separación entre perfiles de riesgo. La duración del crédito introduce una distinción adicional, ya que el *Cluster 0* se asocia a préstamos de corta duración, mientras que el *Cluster 1* agrupa préstamos de mayor plazo. Por tanto, la segmentación obtenida refleja una interacción interpretativa entre nivel de ahorro y duración del préstamo, donde la combinación de ambas

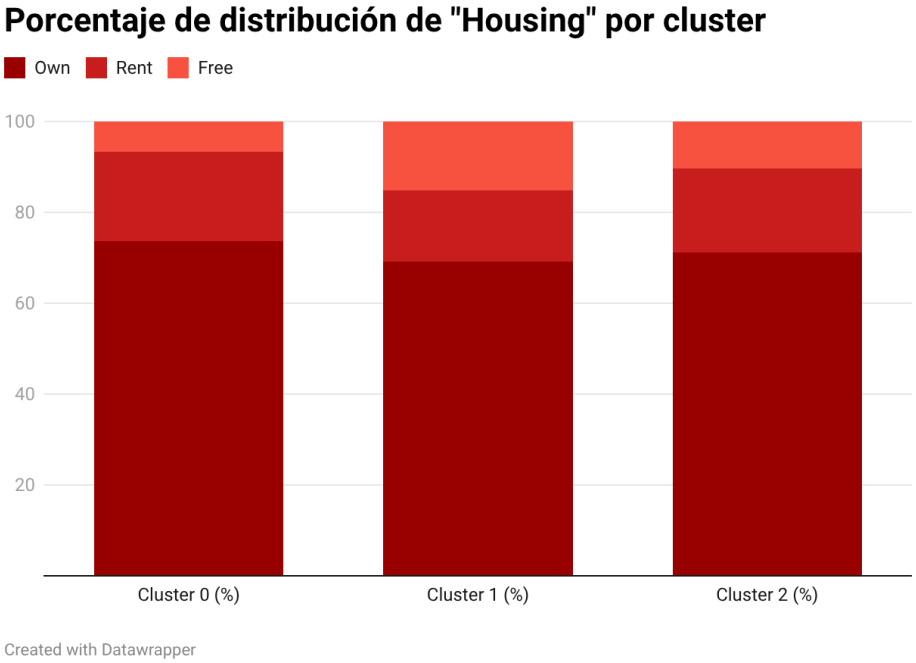

Figura 4.15: Distribución porcentual de la variable *Housing* en cada *cluster*. Elaboración propia.

condiciones permite caracterizar perfiles de riesgo más diferenciados que cada variable por separado.

Con el fin de contrastar la capacidad diferenciadora de las variables con menor peso relativo en el análisis SHAP, se examinaron las distribuciones de *Housing* y *Purpose* por *cluster*. Las Figuras 4.15 y 4.16 muestran que estas variables presentan composiciones relativamente similares entre los tres grupos, sin un patrón de separación tan marcado como el observado en *Duration* y *Saving accounts*. Este resultado es coherente con la menor importancia global que estas variables alcanzaron en el análisis SHAP, donde su contribución promedio a las predicciones del modelo fue inferior a la de los principales factores explicativos. En el caso de *Housing*, la categoría *own* predomina en los tres *clusters*, seguida por proporciones menores de *rent* y *free*. Aunque existen pequeñas diferencias entre grupos, estas no parecen definir una estructura diferenciada de segmentación. De forma similar, la variable *Purpose* muestra una distribución relativamente estable entre los *clusters*, con presencia de finalidades diversas y sin que una categoría específica se asocie de manera dominante a un único grupo. Estos resultados sugieren que la segmentación obtenida se encuentra principalmente estructurada por la duración del crédito y el nivel de ahorro, mientras que variables como *Housing* y *Purpose* aportan información secundaria para la caracterización de los perfiles.

A partir de la evidencia acumulada, es posible asignar una denominación interpretativa a cada *cluster*, con el fin de facilitar la lectura aplicada de los perfiles obtenidos. El *Cluster 0* puede caracterizarse como un perfil de solicitantes de corto plazo con bajo ahorro. Este

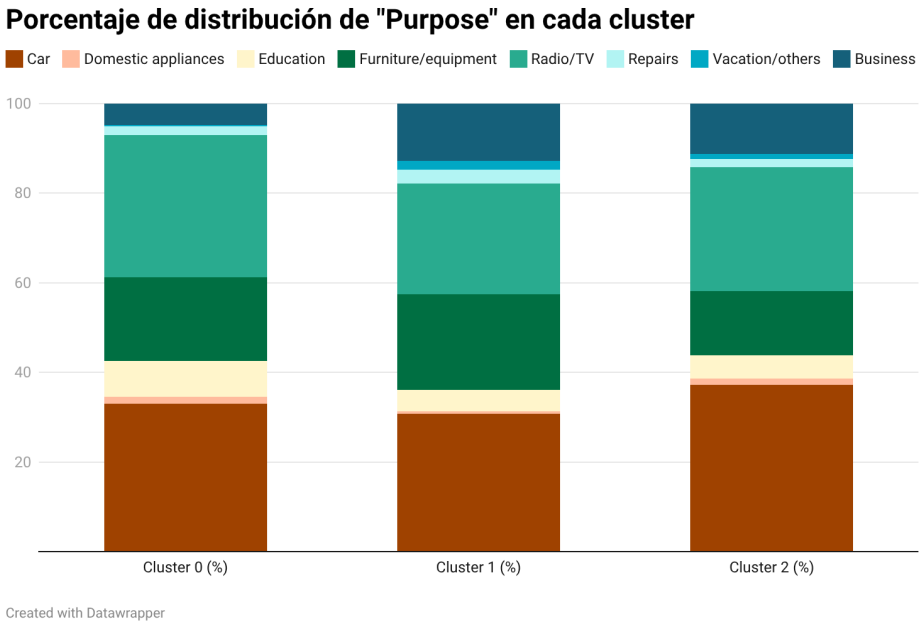

Figura 4.16: Distribución porcentual de la variable *Purpose* en cada *cluster*. Elaboración propia.

grupo combina préstamos de duración reducida con una elevada presencia de la categoría *Saving accounts_little*, lo que describe solicitantes con menor respaldo de ahorro, pero con una exposición temporal acotada. El *Cluster 1* puede interpretarse como un perfil vulnerable de largo plazo, ya que reúne solicitantes con bajo nivel de ahorro y créditos de mayor duración. Esta combinación configura el segmento de mayor fragilidad financiera dentro de la muestra analizada. Por su parte, el *Cluster 2* puede denominarse perfil con mayor estabilidad de ahorro, al no presentar observaciones clasificadas en la categoría *Saving accounts_little* y mostrar una composición más diversa en términos de ahorro, aun cuando incluye préstamos de duración superior al umbral identificado por las reglas.

La Figura 4.17 recoge la distribución de la variable objetivo en cada *cluster* y permite analizar si los perfiles explicativos identificados se asocian con diferentes niveles de riesgo observado. El *Cluster 1* concentra la mayor proporción de solicitantes clasificados como *bad*, con un 46,4 %, lo que confirma su posición como el segmento de mayor vulnerabilidad crediticia. En cambio, los *Clusters 0* y *2* presentan proporciones de riesgo más contenidas y relativamente similares, con un 20,3 % y un 22,8 %, respectivamente. Esta similitud en la tasa de incumplimiento no implica que ambos grupos respondan a la misma lógica. El *Cluster 0* se caracteriza por créditos de corta duración y bajo nivel de ahorro, mientras que el *Cluster 2* agrupa solicitantes con niveles de ahorro más favorables o no clasificados como *little*, junto con préstamos de duración superior a 10.5 meses.

Estos resultados muestran que la metodología propuesta permite pasar de una predicción individual de riesgo a una segmentación interpretable de solicitantes. Los perfiles identi-

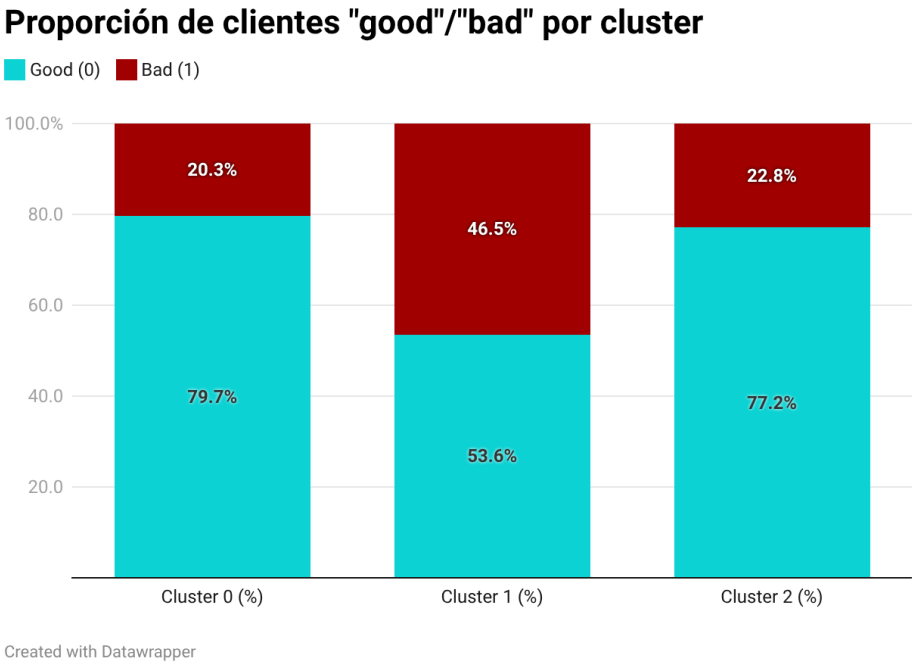

Figura 4.17: Proporción de solicitantes clasificados como *good* y *bad* en cada *cluster* identificado. Elaboración propia.

ficados no solo presentan diferencias en su composición y en la proporción de solicitudes clasificadas como *bad*, sino que además pueden describirse mediante reglas simples basadas en variables observables. La principal limitación es que los *clusters* no constituyen clases puras de riesgo, sino perfiles explicativos con distinta composición interna. Sin embargo, esta característica es coherente con el enfoque de segmentación supervisada adoptado, cuyo propósito es identificar mecanismos de riesgo diferenciados más que reproducir directamente la variable objetivo.

Capítulo 5

Conclusiones

La evaluación del riesgo crediticio constituye un ámbito de especial relevancia en la gestión financiera, ya que una estimación inadecuada de la probabilidad de incumplimiento puede afectar tanto a la rentabilidad de las entidades como al acceso responsable al crédito. Durante años, los modelos de *credit scoring* se han orientado principalmente a clasificar a los solicitantes según su nivel de riesgo. No obstante, el uso creciente de algoritmos más flexibles ha intensificado la dificultad de justificar las decisiones generadas por el modelo y de comprender los factores que influyen en la evaluación de distintos perfiles. En este contexto, la predicción del incumplimiento no resulta suficiente por sí sola. Es necesario avanzar hacia metodologías que permitan explicar la lógica del modelo, caracterizar perfiles de solicitantes a partir de los factores que condicionan su evaluación y traducir dichos patrones en información interpretable para apoyar la toma de decisiones en riesgo crediticio.

La revisión de la literatura permite ubicar esta necesidad dentro de una evolución metodológica caracterizada por el equilibrio entre capacidad predictiva e interpretabilidad. Los primeros enfoques de *credit scoring*, basados en modelos estadísticos como la regresión logística, ofrecieron un marco transparente para estimar la probabilidad de incumplimiento y justificar las decisiones en entornos financieros regulados. Sin embargo, su estructura funcional limita la representación de relaciones no lineales, interacciones entre variables y patrones heterogéneos de riesgo. La incorporación de árboles de decisión, métodos de ensamble y redes neuronales permitió ampliar la capacidad de los modelos para capturar dichas estructuras, aunque a costa de reducir la interpretación directa de sus decisiones. Frente a esta limitación, las técnicas de XAI, especialmente SHAP y LIME, han permitido descomponer las predicciones individuales en contribuciones atribuibles a las variables del modelo, ofreciendo una vía para analizar la lógica interna de algoritmos complejos. Más recientemente, trabajos como los de (Cooper, 2022) y (Gramegna y Giudici, 2021) han extendido este enfoque al utilizar las explicaciones locales como una representación alternativa para segmentar observaciones según los factores que explican su riesgo estimado. Esta línea de investigación constituye el punto de partida del presente trabajo, al desplazar el análisis desde la predicción individual

hacia la identificación de perfiles explicativos de riesgo crediticio.

En este marco, el objetivo principal de este Trabajo de Fin de Grado ha sido analizar la utilidad del *supervised clustering* basado en valores SHAP para identificar perfiles explicativos de riesgo crediticio de forma interpretable. La propuesta parte de la idea de que la segmentación de solicitantes puede enriquecerse si los grupos no se construyen directamente sobre las variables originales, sino sobre una representación derivada del modelo predictivo, en la que cada observación queda descrita por los factores que explican su riesgo estimado. Para ello, se diseñó una metodología híbrida que combina aprendizaje supervisado, explicabilidad y segmentación no supervisada. En primer lugar, se entrenó un modelo LightGBM sobre el conjunto *German Credit Data with Risk* para estimar el riesgo crediticio de los solicitantes. A partir de este modelo se calcularon los valores SHAP, utilizados como representación explicativa del conjunto de datos. Posteriormente, dicha representación se proyectó mediante UMAP con el fin de reducir su dimensionalidad y facilitar el análisis de su estructura. Sobre el espacio proyectado se aplicó DBSCAN para identificar agrupaciones densas de solicitantes con patrones explicativos similares. Finalmente, los segmentos obtenidos se caracterizaron mediante *SkopeRules*, lo que permitió traducir los perfiles identificados a reglas interpretables formuladas sobre variables originales del solicitante.

Los resultados obtenidos permiten valorar la utilidad del enfoque propuesto para construir perfiles explicativos de riesgo crediticio. El modelo LightGBM alcanzó un ROC-AUC de 0,686 y una exactitud del 74 % en el conjunto de prueba, superando en cuatro puntos porcentuales la línea base definida por la regla de la mayoría. Aunque se exploró un amplio espacio de hiperparámetros mediante *Randomized Search* con validación cruzada, el rendimiento predictivo fue moderado. Este comportamiento puede relacionarse con el desbalance entre clases, la dimensión reducida del conjunto de datos y la complejidad limitada de las variables disponibles. No obstante, en este trabajo el modelo no se emplea únicamente como clasificador final, sino como base para generar una representación explicativa mediante valores SHAP. Desde esta perspectiva, su desempeño resulta suficiente para analizar los factores que influyen en las predicciones y avanzar hacia la segmentación supervisada.

La segmentación realizada sobre el espacio SHAP permitió identificar tres *clusters* en la proyección UMAP, con un coeficiente de *Silhouette* de 0,802 y un índice de *Davies–Bouldin* de 0,307. Estas métricas indican una estructura compacta y bien separada en el espacio proyectado. La distribución de la variable objetivo dentro de los grupos mostró que el *Cluster 1* concentra la mayor proporción de solicitantes clasificados como *bad*, configurando el perfil de mayor vulnerabilidad crediticia. Los *Clusters 0* y *2* presentan tasas de incumplimiento más contenidas, aunque responden a lógicas distintas. La interpretación mediante *SkopeRules* permitió traducir esta separación a reglas simples basadas en variables originales, principalmente *Duration* y *Saving accounts_little*. Las reglas obtenidas presentaron niveles elevados de precisión y *recall*, lo que sugiere que la estructura identificada en el espacio explicativo conserva una correspondencia interpretable con características observables de los solicitan-

tes.

La caracterización final de los grupos permitió distinguir tres perfiles con implicaciones diferenciadas para el análisis del riesgo. El *Cluster 0*, denominado solicitantes de corto plazo con bajo ahorro, agrupa créditos de duración reducida y una elevada presencia de solicitantes con bajo nivel de ahorro. Este perfil refleja una situación financiera ajustada, aunque con una exposición temporal limitada. El *Cluster 1*, denominado solicitantes vulnerables de largo plazo, reúne créditos de mayor duración y concentración total en la categoría de menor ahorro, lo que explica su mayor proporción de observaciones *bad*. El *Cluster 2*, denominado solicitantes con mayor estabilidad de ahorro, no incluye observaciones en la categoría *Saving accounts_little* y presenta una composición más diversificada en términos de ahorro, aun cuando contiene préstamos de mayor duración.

A pesar de estos resultados, el estudio presenta algunas limitaciones que deben considerarse al valorar el alcance de la metodología aplicada. En primer lugar, el conjunto de datos utilizado presenta un desbalance moderado entre clases, dado que la categoría *good* representa el 70 % de las observaciones. Esta distribución condicionó el aprendizaje del modelo, que mostró mayor capacidad para identificar solicitantes de bajo riesgo que para detectar la clase *bad*. Esta misma composición se reflejó posteriormente en la segmentación, donde los tres *clusters* estuvieron dominados por la clase mayoritaria. En consecuencia, los perfiles obtenidos no deben interpretarse como segmentos puros de riesgo observado, sino como perfiles explicativos con distinta proporción interna de solicitantes *good* y *bad*. En segundo lugar, la reducida dimensionalidad del conjunto de datos limitó la diversidad de factores disponibles para explicar el riesgo crediticio. Esto concentró la interpretación del modelo en un número reducido de variables, especialmente la duración del crédito y el nivel de ahorro, lo que permitió obtener reglas simples y comprensibles, pero restringió la posibilidad de capturar perfiles financieros más complejos.

Como líneas futuras de investigación, sería conveniente replicar la metodología en conjuntos de datos de mayor tamaño y mayor riqueza informativa, incorporando variables habituales en contextos crediticios reales, como historial de pagos, ratio de endeudamiento, ingresos recurrentes, estabilidad laboral, comportamiento financiero reciente o información transaccional. La incorporación de estas variables permitiría construir un espacio explicativo más amplio y evaluar si la segmentación supervisada identifica perfiles más granulares y próximos a la heterogeneidad observada en entornos financieros reales. También sería pertinente analizar estrategias alternativas de modelización que mejoren la detección de solicitantes de alto riesgo sin deteriorar la interpretabilidad del procedimiento, por ejemplo mediante funciones de pérdida sensibles al coste del error, calibración de probabilidades, ajuste de umbrales de decisión o validación con métricas orientadas a la clase minoritaria. Finalmente, futuras extensiones podrían comparar distintas configuraciones metodológicas del enfoque propuesto. En particular, sería útil evaluar otros modelos supervisados como base para el cálculo de valores SHAP, contrastar diferentes técnicas de reducción de dimensionalidad y analizar

la estabilidad de los *clusters* ante cambios en los hiperparámetros de UMAP y DBSCAN. Asimismo, la validación de los perfiles obtenidos con datos externos o con indicadores financieros adicionales permitiría valorar con mayor precisión su utilidad operativa.

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

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

Por la presente, yo, Sofia González-Montagut Tanure, estudiante de Programa de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Análisis de perfiles de riesgo crediticio utilizando segmentación supervisada con valores SHAP en solicitantes de crédito", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.
6. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
7. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
8. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.

9. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
10. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
11. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
12. Traductor: Para traducir textos de un lenguaje a otro.

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

Fecha: Junio de 2026

Firma: Sofia González-Montagut Tanure

Referencias

Alonso, A., y Carbó, J. M. (2021). *Understanding the performance of machine learning models to predict credit default: A novel approach for supervisory evaluation* (Inf. Téc. n.º Documentos de Trabajo N.º 2105). Banco de España. Descargado de https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/21/Files/dt2105e.pdf (Madrid, España)

Alonso, A., Durán, D., García-Olmedo, B., y Quesada, M. A. (2024). Basel core principles for effective banking supervision: an update after a decade of experience. *Financial Stability Review*, *46*.

Angelini, E., di Tollo, G., y Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733–755. doi: 10.1016/j.qref.2007.04.001

Ansari, Z., Azeem, M. F., Ahmed, W., y Babu, A. V. (2015). Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. *arXiv preprint arXiv:1507.03340*.

Authority, E. B. (2021, November). *Discussion paper on machine learning for irb models* (Inf. Téc. n.º EBA/DP/2021/04). Autor. Descargado de https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf

Banco de España. (2024). *Informe de estabilidad financiera. abril 2024.* Madrid, España. Descargado de https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesEstabilidadFinanciera/2024/Fich/IEF_0424.pdf (Consultado el 9 de octubre de 2025)

Banco de España. (2025). *Informe de la situación financiera de los hogares y las empresas. 1.º semestre de 2025.* Madrid, España. Descargado de https://www.bde.es/f/webbe/SES/Secciones/Publicaciones/Informesituacionfinancierafamiliasyempresas/2025/S1/Fich/SituacionFinanciera_012025.pdf (Consultado el 9 de octubre de 2025)

Bolton, L. (2009). *Logistic regression and its application in credit scoring* (Tesis de Master, University of Pretoria). Descargado de https://repository.up.ac.za/bitstream/handle/2263/27333/dissertation.pdf?sequence=1 (Master's dissertation in Financial Management)

Borio, C., Furfine, C., Lowe, P., y cols. (2001). Procyclicality of the financial system and financial stability: issues and policy options. *BIS papers*, *1*(3), 1–57.

Cardona Hernández, P. A. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de Estadística*, *27*(2), 139–151. Descargado de https://www.redalyc.org/pdf/899/89927204.pdf

Cooper, A. (2022). *Supervised clustering with shap values.* https://www.aidancooper.co.uk/supervised-clustering-shap-values/. (Accessed: 2025-01-25)

Davies, D. L., y Bouldin, D. W. (1979, April). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. doi: 10.1109/TPAMI.1979.4766909

Dhamangaonkar, S. (2020, December 20). *Ml: Model interpretability methods: Pdp, ice eli5, lime, shap implementation on tabular dataset.* https://dhamangaonkar-s.medium.com/ml-model-interpretability-methods-7c2fc02f51b6. (Accessed [fecha de consulta])

Durand, D. (1941). Summary of findings. En *Risk elements in consumer instalment financing, technical edition* (pp. 1–8). National Bureau of Economic Research. Descargado de https://www.nber.org/chapters/c12948.pdf (In: Risk Elements in Consumer Instalment Financing, NBER, 1941. Available at: https://www.nber.org/books/dura41-1)

EBA. (2021). *Discussion paper on machine learning for irb models.* European Banking Authority.

Echanove Puig, M. (2025). *Segmentación de clientes con clustering supervisado e interpretabilidad shap en fuga de clientes.* Trabajo de Fin de Grado, Facultad de Ciencias Económicas y Empresariales, Universidad Pontificia Comillas. (Dirigido por Jenny Alexandra Cifuentes Quintero. Madrid, Junio 2025)

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., y cols. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. En *kdd* (Vol. 96, pp. 226–231).

Funcas. (2020). Riesgo de crédito y provisiones: prudencia en la banca europea y española. *Funcas Blog*. Descargado de https://www.funcas.es/articulos/riesgo-de-credito-y-provisiones-prudencia-en-la-banca-europea-y-espanola/ (Consultado el 9 de octubre de 2025)

Game of Bits. (2019). *Deriving interpretable rules from classification models.* Descargado de https://medium.com/game-of-bits/deriving-interpretable-rules-from-classification-models-3fda725cc5a3 (Medium article. Accessed:

2026-01-25)

García-Escribano, M., y Han, F. (2015). *Credit expansion in emerging markets: Propeller of growth?* (Working Paper n.º WP/15/212). International Monetary Fund (IMF). Descargado de https://www.imf.org/external/pubs/ft/wp/2015/wp15212.pdf (Consultado el 9 de octubre de 2025)

Gramegna, A., y Giudici, P. (2021). Shap and lime: An evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 1–6. Descargado de https://www.frontiersin.org/articles/10.3389/frai.2021.752558/full doi: 10.3389/frai.2021.752558

Hadji Misheva, B., Hirsa, A., Osterrieder, J., Kulkarni, O., y Lin, S. F. (2021, marzo). Explainable ai in credit risk management. *arXiv preprint*. Descargado de https://www.researchgate.net/publication/349703943_Explainable_AI_in_Credit_Risk_Management (A Preprint)

Hand, D. J., y Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *International Statistical Review*, *64*(3), 243–265. Descargado de https://watermark02.silverchair.com/jrsssa_160_3_523.pdf (PDF available via the International Statistical Review archive) doi: 10.2307/1403773

James, G. (2013). *An introduction to statistical learning with applications in r.* springer.

Ji, Y. (2021). *Explainable ai methods for credit card fraud detection: Evaluation of lime and shap through a user study.*

Kaggle. (2024). *German credit data with risk.* Descargado de https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk (Accedido el 8 de junio de 2026)

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, *37*(9), 6233–6239. doi: 10.1016/j.eswa.2010.02.101

Kleinbort, M. (2021). *A love letter to data science tooling.* Descargado de https://mkleinbort.medium.com/a-love-letter-to-data-science-tooling-c91ac6a169d5 (Medium article. Accessed: 2026-01-25)

Lundberg, S. M., y Lee, S.-I. (2017). A unified approach to interpreting model predictions. En *Proceedings of the 31st international conference on neural information processing systems (nips 2017)* (pp. 4765–4774).

Menasalvas, E., Ángel Guzmán, M., y Ruíz Bonilla, S. (2022). *Ml aplicado al riesgo de crédito: construcción de modelos explicables* (Inf. Téc. n.os Newsletter Trimestral, 3T22). Cátedra iDanae, Universidad Politécnica de Madrid y Management Solutions. Descargado de https://www.managementsolutions.com/sites/default/

files/publicaciones/ES/idanae-newsletter-ml-riesgo-credito.pdf (Informe técnico sobre explicabilidad e inteligencia artificial aplicada al riesgo de crédito)

Munafo, F. (2019). La importancia de la gestión de datos y su impacto en el riesgo de crédito de instituciones financieras. *Revista de Investigación en Modelos Financieros*, *2*, 25–38.

on Banking Supervision, B. C. (2004). *International convergence of capital measurement and capital standards: A revised framework (basel ii)* (Inf. Téc.). Bank for International Settlements. Descargado de https://www.bis.org/publ/bcbs107.htm

on Banking Supervision, B. C. (2011). *Basel iii: A global regulatory framework for more resilient banks and banking systems* (Inf. Téc.). Bank for International Settlements. Descargado de https://www.bis.org/publ/bcbs189.htm

Parlamento Europeo y Consejo de la Unión Europea. (2016). *Reglamento (ue) 2016/679 del parlamento europeo y del consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (reglamento general de protección de datos).* Descargado de https://eur-lex.europa.eu/eli/reg/2016/679/oj (Consultado el 9 de octubre de 2025)

Pozo Fonseca, J. C. (2024). *Segmentación de clientes utilizando aprendizaje no supervisado para predecir el comportamiento de pago en una cartera de clientes.* (Tesis de Master no publicada). Quito: EPN, 2024.

Rousseeuw, P. J. (1987, November). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi: 10.1016/0377-0427(87)90125-7

Sangacha Ávila, R. A. (2024). *Evaluación de modelos de machine learning aplicados al cálculo de pérdidas esperadas en entidades de microfinanzas.* (B.S. thesis). Quito: EPN, 2024.

Schröder, C. (2024). *Understanding umap.* https://meta.caspershire.net/umap/. (Online resource. Accessed: 21 Jan. 2026)

Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., y Xu, X. (2017). Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems*, *42*(3), 1–21. doi: 10.1145/3068335

Sharma, D. (2011). Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. *SSRN Electronic Journal*. (Available at SSRN: https://ssrn.com/abstract=1861535) doi: 10.2139/ssrn.1861535

Starmer, J. (2018). *Clustering with dbscan, clearly explained!!!* YouTube. (StatQuest with Josh Starmer. Disponible en: https://www.youtube.com/watch?v=RDZUdRSDOok)

Technology, T. (2023, 28 de noviembre). *Light gbm light and powerful gradient boost algorithm.* Medium. Descargado de https://medium.com/@turkishtechnology/

light-gbm-light-and-powerful-gradient-boost-algorithm-eaa1e804eca8 (Accessed: 2025-12-05)

Tiwari, K., Mehta, K., Jain, N., Tiwari, R., y Kanda, G. (2007). Selecting the appropriate outlier treatment for common industry applications. En *Nesug conference proceedings on statistics and data analysis* (pp. 1–5).

Twomey, L. (2025). *Easy umap – explained with an example.* https://biostatsquid.com/umap-simply-explained/. (Accessed: 24 March 2025)

with Josh Starmer, S. (2022, March 7). *Umap dimension reduction, main ideas!!!* YouTube Video. Descargado de https://www.youtube.com/watch?v=eN0wFzBA4Sc ([Video]. Available at: https://www.youtube.com/watch?v=eN0wFzBA4Sc (Accessed: 21 Jan. 2026))