Segmentación de clientes y modelado del riesgo de crédito mediante K-Means y XGBoost

Valentina Zapata Alzate
Ivan Mauricio Cuberos Arango

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Gabriel Darío Uribe Guerra, Magister en Matemáticas

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2026

Cita	(Zapata Alzate & Cuberos Arango, 2026)
Referencia **Estilo APA 7 (2020)**	Zapata Alzate, V., & Cuberos Arango, I. M. (2026). Segmentación de clientes y modelado del riesgo de crédito mediante K-Means y XGBoost [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte X.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Tabla de contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

EDA	Análisis exploratorio de datos (Exploratory Data Analysis)
ROC	Curva característica de operación del receptor (Receiver Operating Characteristic)
AUC	Área bajo la curva (Area Under the Curve)
F1-score	Media armónica entre precisión y recall
ML	Aprendizaje automático (Machine Learning)
K-Means	Algoritmo de agrupamiento basado en centroides
XGBoost	Extreme Gradient Boosting
SHAP	SHapley Additive exPlanations
Precision	Precisión del modelo
Recall	Sensibilidad del modelo

Resumen

Este trabajo trata el problema del riesgo de crédito, el cual es fundamental en la toma de decisiones dentro del sector financiero. El objetivo principal es evaluar si la segmentación de clientes mediante el algoritmo K-Means aporta valor descriptivo y predictivo al modelado del incumplimiento utilizando XGBoost. Para ello, se utilizó un dataset público con más de 32.000 registros que incluye variables sociodemográficas y financieras de personas solicitantes de crédito. La metodología se basó en un proceso de limpieza y transformación de datos, seguido de un análisis exploratorio, la aplicación de K-Means para identificar perfiles de clientes y la construcción de dos modelos: uno base y otro que incluye la variable de cluster. Durante el desarrollo se presentaron dificultades relacionadas con la calidad de los datos, como valores nulos, errores en escalas y presencia de valores atípicos, los cuales fueron tratados mediante transformación y ajustes en las variables. Los resultados muestran que el modelo base alcanzó un ROC-AUC cercano a 0.95, mientras que el modelo segmentado obtuvo un desempeño muy similar, sin mejoras relevantes. Esto sugiere que la segmentación no aporta valor predictivo adicional, aunque sí contribuye al entendimiento del comportamiento de los clientes.

Palabras clave: Riesgo de crédito, Segmentación de clientes, K-Means, XGBoost, Aprendizaje automático

Abstract

This study addresses the problem of credit risk, which is a key factor in decision-making within the financial sector. The main objective is to evaluate whether customer segmentation using the K-Means algorithm provides descriptive and predictive value in modeling credit default using XGBoost. A public dataset with more than 32,000 records was used, including sociodemographic and financial variables of loan applicants. The methodology involved data cleaning and transformation, followed by exploratory data analysis, the application of K-Means to identify customer profiles, and the development of two predictive models: a baseline model and a model incorporating cluster information. During the process, several data quality issues were addressed, including missing values, inconsistent scales, and outliers. The results show that the baseline model achieved a ROC-AUC close to 0.95, while the segmented model obtained a very similar performance, with no significant improvement. These findings suggest that segmentation does not provide additional predictive value, although it contributes to a better understanding of customer behavior.

Keywords: Credit Risk, Customer Segmentation, K-Means, XGBoos, Machine Learning.

1. Introducción

El crédito es uno de los principales mecanismos de colocación de recursos en el sistema financiero y, al mismo tiempo, una de las fuentes más importantes de riesgo para las entidades. Una mala evaluación del riesgo puede generar pérdidas significativas y afectar la estabilidad financiera, por lo que estimar correctamente la probabilidad de incumplimiento es clave en la toma de decisiones. En este contexto, la ciencia de datos ha tomado un papel importante, ya que permite analizar grandes volúmenes de información e identificar patrones en el comportamiento de los clientes (Lessmann et al., 2015).

Tradicionalmente, el riesgo de crédito se ha evaluado mediante modelos estadísticos como la regresión logística, los cuales son fáciles de interpretar y aplicar. Sin embargo, estos modelos tienen limitaciones al momento de capturar relaciones más complejas entre variables (Hand & Henley, 1997). Con el avance de la tecnología y la disponibilidad de datos, los modelos de machine learning han empezado a utilizarse con mayor frecuencia, mostrando mejores resultados en la predicción del incumplimiento. De hecho, diferentes estudios han evidenciado que algoritmos como Random Forest, Support Vector Machines y modelos de boosting pueden superar a los enfoques tradicionales en este tipo de problemas (Lessmann et al., 2015).

Dentro de estos modelos, XGBoost se ha convertido en una de las herramientas más utilizadas debido a su buen desempeño y eficiencia. Este modelo permite trabajar con diferentes tipos de variables y capturar relaciones más complejas, lo que lo hace adecuado para problemas como el riesgo de crédito (Chen & Guestrin, 2016). Sin embargo, una de sus principales limitaciones es que no siempre es fácil de interpretar, ya que en muchos casos funciona como una “caja negra”. Esto representa un reto en el sector financiero, donde es importante entender cómo se toman las decisiones. Para abordar este problema, se han desarrollado herramientas como SHAP, que permiten analizar la importancia de cada variable en las predicciones del modelo (Lundberg & Lee, 2017).

Por otro lado, el análisis del riesgo no solo depende de la predicción, sino también de entender mejor a los clientes. En este sentido, las técnicas de segmentación, como el algoritmo K-Means, permiten agrupar personas con características similares, facilitando la identificación de perfiles de riesgo (Jain, 2010; MacQueen, 1967). Este tipo de análisis puede ayudar a entender mejor el comportamiento de los clientes y a identificar grupos con mayor o menor riesgo.

A pesar de que tanto los modelos predictivos como las técnicas de segmentación han sido ampliamente estudiados, no siempre se analiza su uso conjunto. En muchos casos, la segmentación se utiliza solo para describir los datos, pero no se evalúa si realmente mejora el desempeño de los modelos. Esto genera una duda importante: ¿la

segmentación aporta valor real en la predicción o su principal aporte es ayudar a entender mejor los datos?

A partir de lo anterior, surge la siguiente pregunta de investigación: ¿la segmentación de clientes mediante K-Means aporta valor descriptivo y/o predictivo al modelado del riesgo de crédito utilizando XGBoost? En este sentido, el objetivo de este trabajo es evaluar si incluir esta segmentación mejora el desempeño del modelo y permite entender mejor los perfiles de riesgo.
Para ello, se utiliza un dataset público de riesgo crediticio que contiene información sociodemográfica y financiera de los solicitantes. La metodología incluye un proceso de limpieza y transformación de los datos, seguido de un análisis exploratorio, la aplicación de K-Means para segmentar a los clientes y la construcción de modelos predictivos mediante XGBoost (Chen & Guestrin, 2016). Se comparan dos enfoques: un modelo base y un modelo que incluye la variable de cluster, utilizando métricas como ROC-AUC, precisión, recall y F1-score.

Finalmente, el documento se organiza de la siguiente manera: en la sección 2 se presentan los datos y la metodología; en la sección 3 se muestran los resultados y hallazgos; y en la sección 4 se presentan las conclusiones del estudio.

2. Materiales y Métodos

2.1 Descripción de los datos

Para el desarrollo de este estudio se utilizó un dataset abierto obtenido de la plataforma Kaggle, denominado “Credit Risk Dataset” (Lao Tse, 2020). Este conjunto de datos contiene información sociodemográfica y financiera de solicitantes de crédito, lo que permite aproximarse al análisis del riesgo crediticio en un contexto aplicado (ver Tabla 1).

El dataset cuenta con 32.581 registros y 12 variables, que incluyen características como edad del solicitante, ingreso anual, antigüedad laboral, monto del crédito, tasa de interés y longitud del historial crediticio. Estas variables permiten capturar tanto la capacidad de pago como el comportamiento financiero de los individuos, elementos fundamentales en la evaluación del riesgo de crédito. La variable objetivo del estudio es “loan_status”, donde el valor 0 representa el cumplimiento del crédito y el valor 1 el incumplimiento. En el conjunto de datos, aproximadamente el 78% de los casos corresponden a clientes sin incumplimiento, mientras que el 22% presentan default, lo que evidencia un desbalance en la variable objetivo.

Este tipo de estructura es consistente con los problemas de riesgo de crédito, donde la ocurrencia de incumplimiento suele ser menor en comparación con los casos de pago normal. En este sentido, como señalan (Bedoya Leiva, 2020), el análisis del riesgo

crediticio se basa en la identificación de características que permiten diferenciar entre clientes con alta y baja probabilidad de incumplimiento, considerando variables financieras, demográficas y de comportamiento.

***Tabla 1**. Detalles de las variables del proyecto*

Variable	Descripción	Tipo	Rol en análisis
person_age	Edad del solicitante	numérica	predictora
person_income	Ingreso anual	numérica	predictora
person_home_ownership	Tipo de tenencia de vivienda	categórica	predictora
person_emp_length	Años de empleo	numérica	predictora
loan_intent	Propósito del crédito	categórica	predictora
loan_grade	Calificación del préstamo	categórica	predictora
loan_amnt	Monto del préstamo	numérica	predictora
loan_int_rate	Tasa de interés	numérica	predictora
loan_status	Incumplimiento del préstamo (0 = no default, 1 = default)	binaria	objetivo
loan_percent_income	Proporción del ingreso comprometida con el préstamo	numérica	predictora
cb_person_default_on_file	Historial de default	categórica	predictora
cb_person_cred_hist_length	Longitud del historial crediticio	numérica	predictora

2.2 Análisis exploratorio (EDA)

La exploración inicial del dataset constituye la primera fase del trabajo y una de las más importantes, ya que permite evaluar la calidad de los datos y detectar posibles inconsistencias que pueden afectar el desempeño de los modelos. En este sentido, el análisis exploratorio de datos (EDA) no solo facilita la limpieza de la información, sino que también permite identificar patrones, formular hipótesis y sustentar decisiones analíticas con mayor rigor (Dunkeld, 2024).

En el presente proyecto se realizaron gráficos de distribución para variables numéricas como ingreso, monto del préstamo y proporción del ingreso comprometido, así como gráficos de frecuencia para variables categóricas como el propósito del crédito y la variable objetivo (ver Figuras 2–7). Este análisis permitió comprender la variabilidad y comportamiento de los datos, así como detectar posibles anomalías y sesgos en su distribución (Ulloa, 2025). Adicionalmente, se construyó una matriz de correlación entre las variables numéricas mediante un mapa de calor que utiliza el coeficiente de correlación de Pearson para medir la intensidad y dirección de la relación entre variables. Los valores cercanos a 1 (rojo) indican una fuerte correlación positiva, mientras que los valores cercanos a -1 (azul) señalan relaciones inversas, permitiendo identificar los factores con mayor influencia en el comportamiento del préstamo. (ver

Figura 1), con el objetivo de analizar las relaciones lineales existentes entre ellas y su posible impacto en el modelado. Este análisis permite identificar dependencias entre variables, así como posibles problemas de multicolinealidad. A partir de la matriz, se observa una alta correlación entre la edad del solicitante y la longitud del historial crediticio (0.88), lo cual es consistente desde el punto de vista financiero, ya que a mayor edad es esperable un mayor historial de crédito. Asimismo, se evidencia una relación moderada entre el monto del préstamo y la proporción del ingreso comprometido (0.50), indicando que créditos más altos tienden a representar una mayor carga financiera para el solicitante.

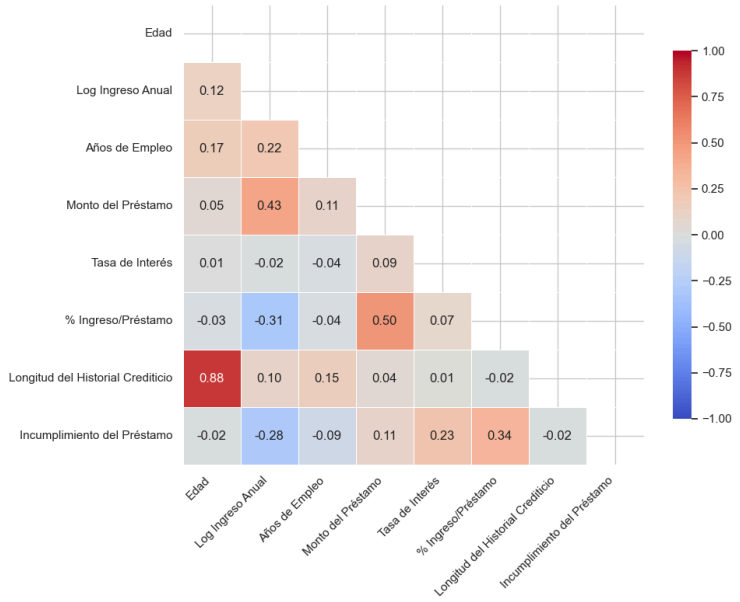

Figura 1*.Gráfica de matriz de correlación de variables numéricas.*

En cuanto a la variable objetivo (loan_status), se identifican correlaciones positivas con la tasa de interés (0.23) y con la proporción del ingreso comprometido (0.34), lo que sugiere que mayores niveles de costo financiero y carga de deuda están asociados con un mayor riesgo de incumplimiento. Por otro lado, se observa una correlación negativa con el ingreso (-0.28), indicando que mayores ingresos podrían estar relacionados con una menor probabilidad de default. En general, las correlaciones entre variables explicativas son moderadas o bajas, lo cual es favorable para el modelado, ya que reduce el riesgo de multicolinealidad y permite que los modelos capturen relaciones más estables entre las variables.

En esta exploración se identificaron varios problemas de calidad de datos que debían ser tratados antes del modelado. En primer lugar, se encontraron 165 registros duplicados, los cuales fueron eliminados para garantizar la consistencia del dataset. Se identificaron valores nulos en las variables “loan_int_rate” y “person_emp_length”, con 3116 y 895 registros respectivamente, los cuales fueron imputados con su mediana correspondiente. Durante la auditoría de calidad de datos también se identificaron inconsistencias en la

escala de algunas variables, las cuales fueron corregidas para garantizar su correcta interpretación y uso en los modelos.

La variable “loan_percent_income”, que representa la proporción del ingreso comprometida con el préstamo, se encontraba expresada en términos porcentuales enteros (por ejemplo, 13 representaba 13%). Para garantizar coherencia con otras variables de tipo razón y facilitar su interpretación en los modelos, se transformó a una escala proporcional dividiendo sus valores entre 100, llevándola a un rango entre 0 y 1. De manera consistente, la variable “loan_int_rate”, correspondiente a la tasa de interés, también presentaba una escala inflada respecto a su valor real, por lo que se aplicó la misma transformación dividiendo sus valores entre 100.

La variable “person_emp_length”, que representa la experiencia laboral en años, presentaba valores inconsistentes en su escala, con observaciones que no guardaban coherencia con la edad de los solicitantes. Tras un análisis de distribución y validación contextual, se identificó que la variable se encontraba sobredimensionada, por lo que se ajustó dividiendo sus valores entre 10. Esta transformación permitió obtener una distribución más coherente, con una mediana de 4 años, un percentil 75 de 7 años y un valor máximo de 41 años.

El análisis de la variable “person_age” evidenció la presencia de valores extremos superiores a 100 años, los cuales no resultan plausibles en el contexto del análisis crediticio. Por esta razón, se decidió eliminar dichas observaciones para evitar distorsiones en el modelado, conservando al mismo tiempo valores altos pero realistas dentro de la población.

En cuanto a la variable “person_income”, se observó una distribución fuertemente sesgada hacia la derecha, con una alta concentración de valores bajos y una cola extensa de ingresos elevados. Dado que estos valores no representan errores sino variabilidad real, se optó por aplicar una transformación logarítmica mediante la función ‘log1p’, con el fin de reducir la asimetría y mejorar el comportamiento de la variable en los modelos.

Finalmente, el análisis descriptivo de las variables según el estado del préstamo (loan_status) permitió identificar diferencias claras entre los perfiles de riesgo. Los individuos en estado de incumplimiento (loan_status = 1) tienden a presentar mayores montos de crédito, con un promedio de 10,854 dólares, a pesar de contar con menores niveles de ingreso y menor estabilidad laboral en comparación con aquellos que cumplen sus obligaciones (loan_status = 0). Esta diferencia se refuerza con una tasa de interés promedio más alta (12.18%) y una mayor proporción del ingreso comprometido (22.37%), lo que sugiere una mayor presión financiera. Por el contrario, variables como la edad y el historial crediticio no muestran diferencias significativas entre ambos grupos (Tabla 2).

Se analizaron las distribuciones de las variables numéricas como el ingreso anual, el monto del préstamo y la proporción del ingreso comprometida (ver Figuras 2, 3 y 4), las cuales son representadas mediante histograma y diagrama de caja (boxplot). Asimismo, se analizaron variables categóricas relevantes como el propósito del crédito y el tipo de tenencia de vivienda (ver Figuras 5 y 6). La combinación de estas gráficas permite identificar la concentración de observaciones en montos bajos y medianos (asimetría positiva), mientras que los puntos aislados en el boxplot señalan valores atípicos.

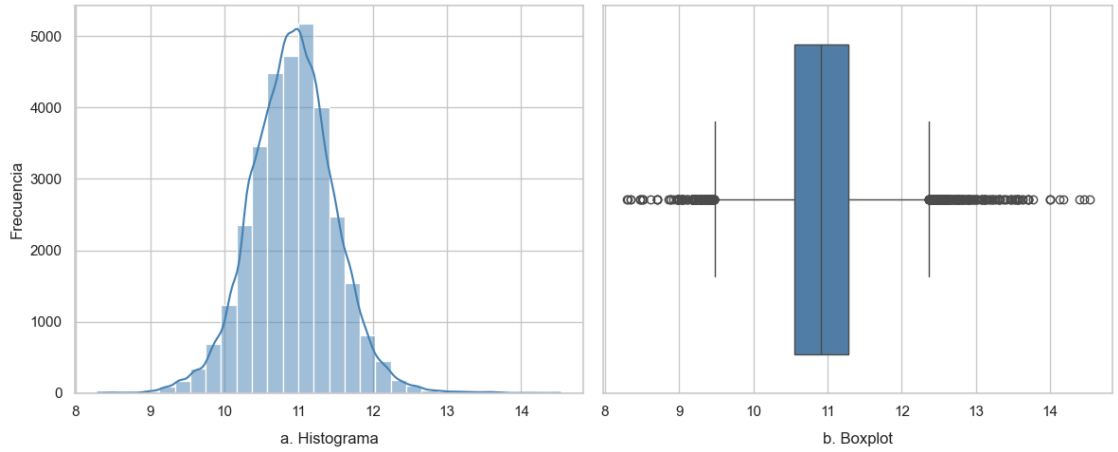

Figura 2. *Gráfica de distribución de la variable 'person_income_log' en escala logarítmica (Ingreso anual)*

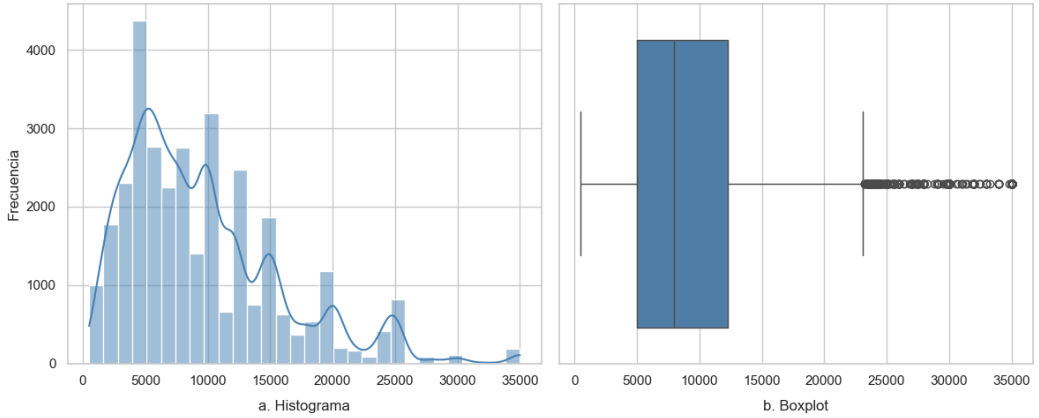

Figura 3. *Gráfica de distribución de variable 'loan_amnt' (Monto del préstamo)*

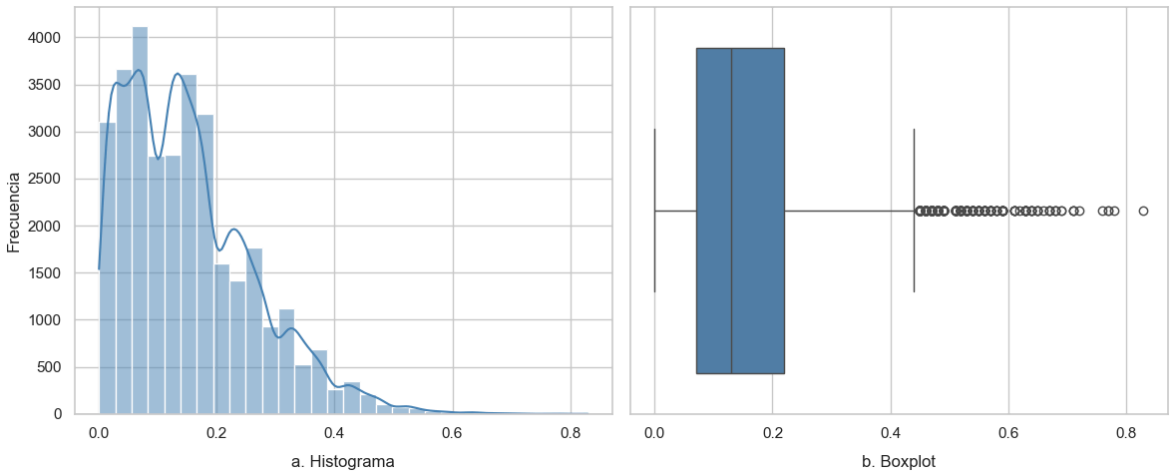

Figura 4. *Gráfica de distribución de variable 'loan_percent_income' (Proporción del ingreso comprometida con el préstamo)*

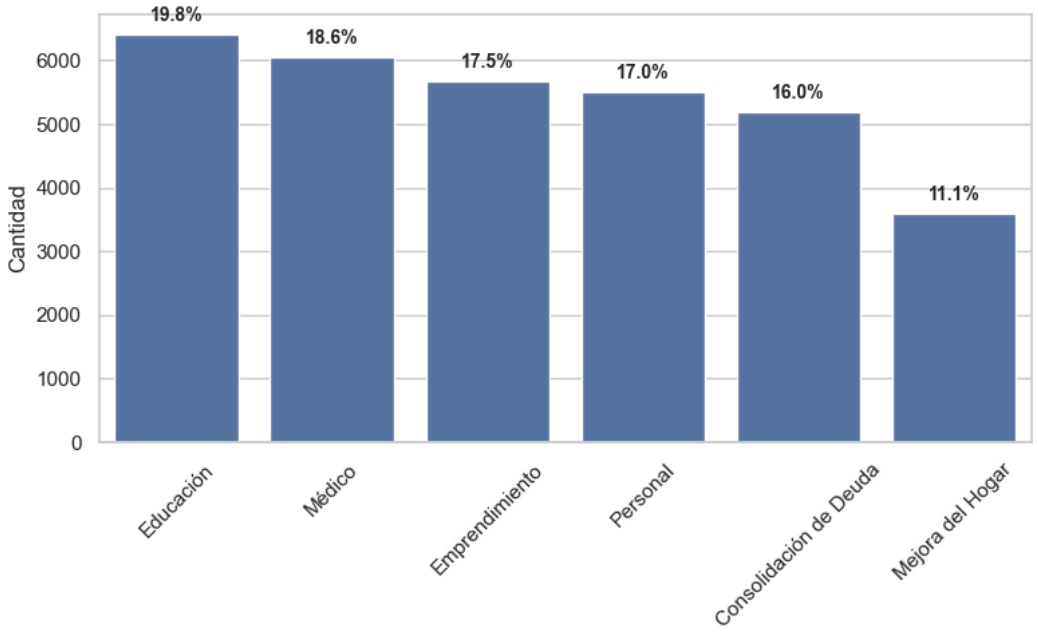

Figura 5. *Gráfica de distribución de variable 'loan_intent' (Propósito del crédito)*

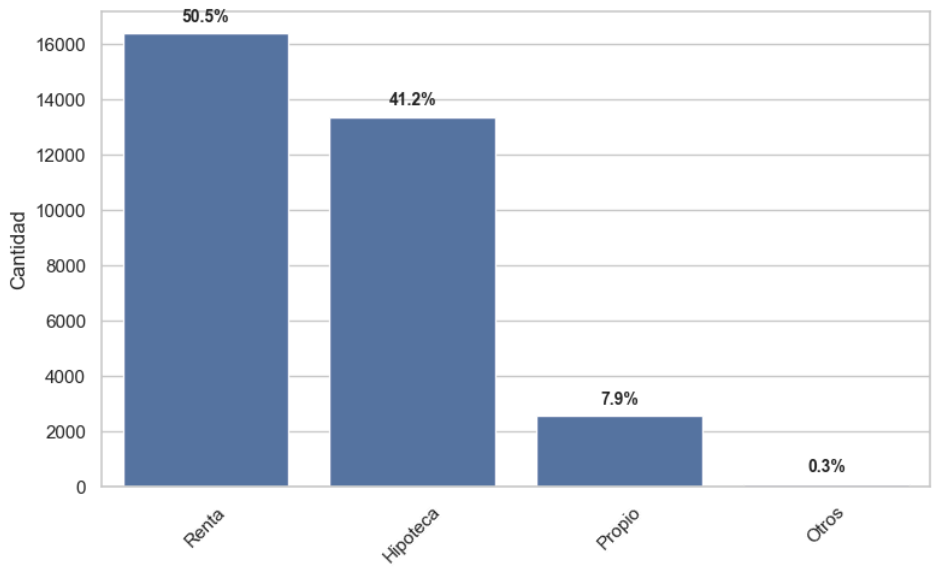

Figura 6. *Distribución de la variable 'person_home_ownership' (tipo de tenencia de vivienda)*

En la figura 5 se observa que ciertos propósitos de crédito presentan mayor frecuencia, lo que puede influir en el perfil de riesgo de los solicitantes. Por su parte, la figura 6 muestra que la mayoría de los clientes se encuentran en condición de renta o hipoteca, mientras que una proporción menor posee vivienda propia. Esta distribución sugiere diferencias en la estabilidad financiera de los individuos, lo cual podría estar relacionado con el riesgo de incumplimiento.

Por último, al analizar los promedios de las variables numéricas según el estado del préstamo, se evidencia un perfil financiero diferenciado entre clientes que cumplen y aquellos que incumplen (ver Tabla 2).

Tabla 2. *Promedios de las variables numéricas según el estado del préstamo*

Loan_Status	Edad del solicitante	Ingreso anual	Años de empleo	Monto del préstamo	Tasa de interés	Proporción del ingreso comprometida con el préstamo	Longitud del historial crediticio
1	27.80	11.01	4.94	9,239	10.03	0.13	5.85
0	27.48	10.62	4.11	10,854	12.18	0.22	5.69

Uno de los hallazgos más significativos se observa en la tasa de interés asignada al crédito. Los clientes con incumplimiento en el crédito (1) registraron una tasa de interés promedio notablemente más alta (12.18%) en comparación con el grupo que no presenta

fallos de pago (10.03%). Este comportamiento es consistente con las políticas financieras tradicionales, donde las entidades bancarias imponen tasas más elevadas a los perfiles percibidos con mayor riesgo para compensar la probabilidad de impago.

Por otra parte, al observar las variables de ingresos y montos otorgados, se identifica que el grupo que cumple con el pago del crédito (0) accedió a montos de préstamo promedio superiores ($10,854) frente a los $9,239 del grupo en incumplimiento. De igual manera, los clientes cumplidores muestran una proporción del ingreso comprometida con el crédito del 22%, superando el 13% que registra el grupo en default (1). Respecto a la variable de ingresos anuales, el promedio del grupo en default se sitúa en 11.01, cifra que debe ser interpretada bajo la escala logarítmica aplicada previamente durante la auditoría de calidad de datos.

Finalmente, variables como la edad del solicitante, los años de empleo y la longitud del historial crediticio muestran un comportamiento homogéneo entre ambos segmentos. Los promedios de edad se mantienen en torno a los 27 años, mientras que la antigüedad del historial crediticio se ubica cerca de los 5.7 años para ambos casos. Esta similitud sugiere que, de manera aislada, estas características demográficas y de antigüedad no representan factores determinantes para discriminar de forma contundente el riesgo de impago en este conjunto de datos.

2.3 Segmentación K-Means.

El agrupamiento K-Means es uno de los algoritmos de aprendizaje no supervisado más utilizados en la ciencia de datos, debido a su simplicidad, eficiencia y facilidad de implementación. Su objetivo principal es segmentar un conjunto de datos en K grupos (clusters) mutuamente excluyentes, agrupando observaciones con características similares a partir de la distancia entre ellas, generalmente medida mediante distancia euclidiana. En este sentido, (Chong, 2021, pág. 37) señala que el algoritmo K-Means “tiene como ventajas una idea simple, alta eficiencia y facilidad de implementación”. En el contexto del dataset utilizado “Credit Risk”, esta técnica permite identificar perfiles de clientes con comportamientos financieros similares sin la necesidad de contar con una variable objetivo, lo cual resulta útil para explorar estructuras subyacentes en los datos.

El funcionamiento del algoritmo es iterativo y se basa en dos etapas principales. En la primera, cada observación es asignada al centroide más cercano, es decir, al punto medio del cluster. En la segunda, se recalculan los centroides como la media de las observaciones asignadas a cada grupo. Este proceso se repite hasta alcanzar la

convergencia, es decir, cuando las asignaciones de los clusters dejan de cambiar. A pesar de su simplicidad, la literatura indica que el desempeño del algoritmo depende de factores clave. Uno de los más importantes es la selección del número óptimo de clusters (K), dado que diferentes valores generan distintas segmentaciones. En este sentido, se reconoce que "la selección del valor de K es desafiante" (Chong, 2021, pág. 37), por lo que en este estudio se utilizaron métricas como el método del codo, el coeficiente de silueta, el índice de Davies-Bouldin y el índice de Calinski-Harabasz para determinar el valor más adecuado.

Otro aspecto relevante es la sensibilidad del algoritmo a valores atípicos y ruido en los datos, lo cual puede afectar la calidad de los clusters obtenidos. Como se menciona en la literatura, los resultados del agrupamiento "pueden verse afectados por ruido y valores atípicos" (Chong, 2021, pág. 37). En este sentido, se realizó un proceso previo de estandarización de las variables numéricas mediante StandardScaler, con el fin de garantizar que todas las variables contribuyeron de manera equilibrada al cálculo de distancias. Para la construcción del modelo, se seleccionaron variables numéricas relevantes relacionadas con características sociodemográficas y financieras del solicitante. Posteriormente, estas variables fueron escaladas y utilizadas como entrada del algoritmo K-Means.

Adicionalmente, se aplicó una reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), con el objetivo de facilitar la visualización de los clusters en un espacio bidimensional y analizar su separación. A partir de la evaluación de las métricas mencionadas, se definió un número final de tres clusters (K=3), priorizando interpretabilidad y estabilidad en los resultados. Finalmente, los clusters obtenidos fueron incorporados como una nueva variable categórica en el conjunto de datos, permitiendo su inclusión en el modelo supervisado XGBoost. De esta manera, se buscó evaluar si la segmentación aporta información adicional en la predicción del incumplimiento crediticio, considerando que cada cluster representa perfiles de riesgo diferenciados dentro de la población analizada.

2.4 Modelo predictivo XGBoost

En estudios de otorgamiento de crédito los algoritmos de aprendizaje supervisado son muy usados debido a que entregan resultados confiables y predictivos de

incumplimientos de pagos. En el presente estudio el algoritmo XGBoost (Extreme Gradient Boosting) se utilizó con la finalidad de predecir la probabilidad de incumplimiento del crédito contando con las variables sociodemográficas, financieras y del préstamo incluidas en el conjunto de datos. Este modelo de machine learning ha sido ampliamente utilizado en problemas de clasificación, en nuestro caso otorgamiento de crédito, gracias a su alto desempeño predictivo y su capacidad para recopilar relaciones no lineales entre las variables.

De acuerdo con (Zúñiga, 2020, pág. 3), XGBoost "es considerada el estado del arte en la evolución de estos algoritmos", dado que se basa en un bosquejo de ensamblado secuencial de árboles de decisión que corrigen repetidamente los errores de los modelos anteriores. En el presente estudio se construyeron dos modelos: un modelo base y un modelo extendido que integra la variable "cluster" obtenida mediante el modelo no supervisado K-Means, con la finalidad de evaluar si la segmentación de los datos aporta valor predictivo adicional para el cumplimiento o no del crédito que se va a adquirir.

Para el entrenamiento de los modelos, la base de datos fue dividida en muestras de entrenamiento y prueba por medio de la partición estratificada, asegurando la proporción de la variable objetivo en ambos subconjuntos de datos. Luego, se realizó la transformación de las variables categóricas y numéricas mediante un "ColumnTransformer", ambos subconjuntos se trataron de forma diferenciada, incluyendo la codificación "one-hot" para las variables categóricas. Dado que se presenta un desbalance de la variable objetivo, se añadió el parámetro "scale_pos_weight dentro del modelo XGBoost, el cual fue calculado entre las observaciones negativas y positivas del conjunto de entrenamiento. Por consiguiente se realizó de igual manera la optimización de los hiperparámetros mediante RandomizedSearchCV, utilizando la validación cruzada en cinco particiones y el área bajo la curva ROC (ROC-AUC) como métrica principal, se acepta que para el modelo XGBoost "se deben ajustar correctamente los parámetros del algoritmo a fin de minimizar el error y evitar sobreajuste del modelo" (Zúñiga, 2020, pág. 4).

Finalmente, el desempeño de los modelos se evaluó sobre el conjunto de prueba mediante métricas de clasificación, incluyendo ROC-AUC, precisión, recall y F1-score,

lo que permitió analizar el comportamiento del modelo tanto en términos globales como en la identificación de la clase minoritaria

3. Resultados y Discusión

3.1 Resultados del EDA.

La etapa de análisis exploratorio de datos (EDA) permitió identificar patrones relevantes en el conjunto de datos antes de la implementación de los modelos de segmentación y clasificación. En primer lugar, se observó que la variable objetivo "loan_status" presenta un desbalance de clases, con mayor medida hacia los clientes con incumplimiento del pago de crédito (78,1%) frente a aquellos que cayeron en default del crédito (21,9%) (ver Figura 7). Esta distribución es importante para este trabajo, ya que un conjunto de datos desbalanceado puede provocar sesgos en el entrenamiento del modelo, favoreciendo la clase mayoritaria y afectando la capacidad predictiva sobre los casos de menor riesgo de incumplir el crédito.

Adicionalmente, en la exploración de los datos mostraron diferencias importantes entre los clientes con y sin default crediticio. En primera instancia, las personas sin default en el crédito presentan mayores niveles de ingresos anuales, con un promedio del logaritmo natural de 11.01 ($60,400) frente a 10.62 ($41,000) del grupo de incumplimiento de crédito, indicando una mayor capacidad de pago para asumir obligaciones crediticias. En cambio, las personas con default señalan montos promedios de préstamos superiores ($10,854 frente a $9,239), además, de una mayor variación en la distribución de los diagramas de caja, indicando una mayor dispersión en los niveles de endeudamientos adquiridos por estas personas.

Así mismo, la tasa de interés muestra diferencias evidentes entre ambos grupos, con un promedio de 12.18% para las personas con default del crédito frente a 10.03% para aquellos que cumplen con la obligación financiera que adquirieron. Este comportamiento es lógico con el riesgo crediticio, donde los perfiles que se comprenden como más riesgosos enfrentan mayores costos de financiamiento (la tasa de interés a la que se les presta). Sin embargo, se debe tener en cuenta que esta variable podría capturar parcialmente decisiones previas de evaluación del riesgo por parte de las entidades financieras, y no únicamente características inherentes del cliente.

Por último, la proporción del ingreso comprometida con el crédito adquirido señala una de las diferencias más evidentes entre estos dos grupos. Las personas con default presentan un promedio de 22% frente a 13% del grupo de personas sin incumplimiento de pago del crédito, además de mostrar una mediana y dispersión más elevada. Esta tendencia señala una mayor presión financiera relativa sobre los ingresos disponibles, dejando claro una señal descriptiva muy importante entre estos dos grupos. En conjunto, estos hallazgos indican que el default tiende a asociarse con perfiles caracterizados por menor capacidad de ingreso, mayor carga financiera relativa y condiciones de financiamiento menos favorables. (Ver Figura 8).

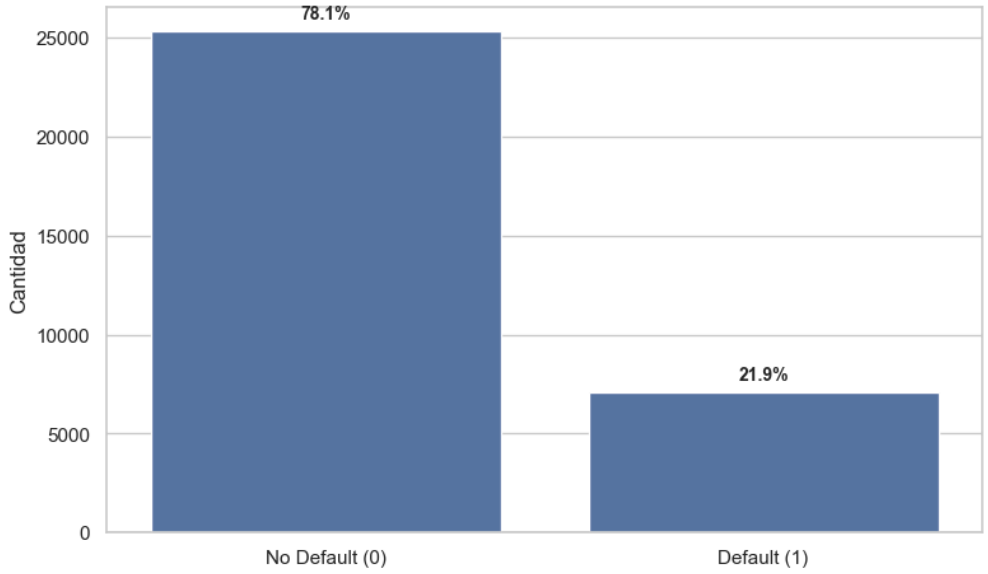

Figura 7. *Gráfica de distribución de variable objetivo 'loan_status' (Incumplimiento del préstamo)*

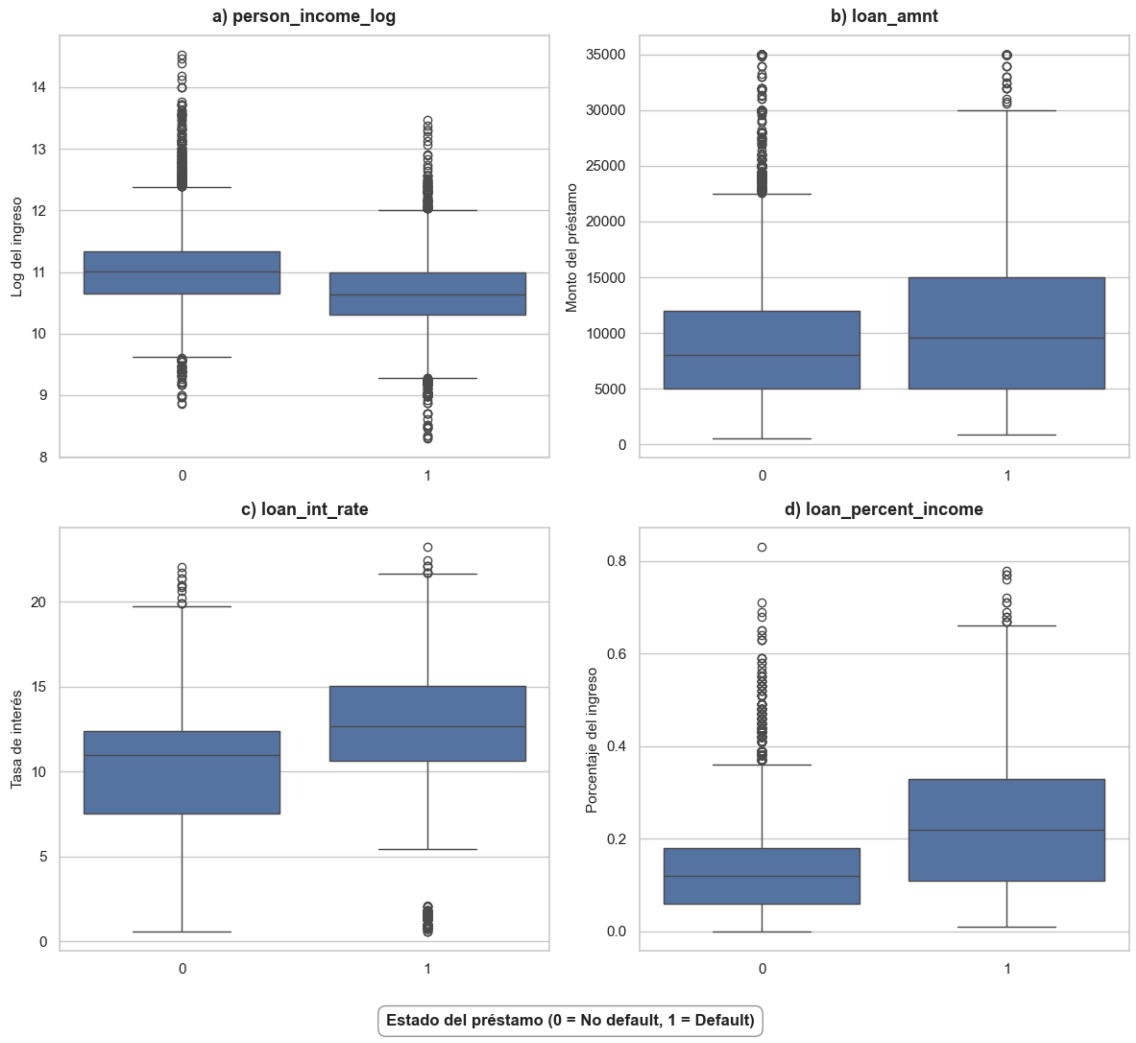

***Figura 8**. Distribución de variables ingreso, monto, tasa y capacidad de pago segmentadas por la variable objetivo 'loan_status' (Incumplimiento del préstamo)*

3.2 Resultados de la segmentación K-Means

Con el propósito de identificar patrones estructurales dentro de la población analizada en el dataset, se implementó el algoritmo de aprendizaje automático no supervisado K-Means (clustering) sobre las variables numéricas seleccionadas de la ya mencionada base de datos. La selección del número de clusters se apoyó en métricas de evaluación interna como el coeficiente de silhouette, el índice de Davies-Bouldin y el índice de Calinski-Harabasz, las cuales permitieron seleccionar una solución de tres grupos (K=3), al ofrecer un balance adecuado entre calidad del agrupamiento e interpretabilidad analítica. La distribución final de los clusters correspondió a 5.756 clientes en el perfil de menor riesgo, 17.976 en un perfil intermedio y 8.677 en el perfil de mayor riesgo.

El análisis descriptivo de los clusters evidenció diferencias relevantes en términos de comportamiento financiero y riesgo crediticio de cada uno de los grupos. El cluster identificado como perfil de mayor riesgo concentra clientes con condiciones menos favorables en la adquisición del crédito, caracterizados por una mayor proporción del préstamo respecto al ingreso (26.80%), tasas de interés promedio más elevadas (11.47%), una mayor tasa de incumplimiento (35.90%) y montos promedios más altos; estos resultados ya mencionados son significativamente superior a los observados en los dos grupos restantes.

En contraste, el perfil de menor riesgo agrupa clientes con mejores condiciones financieras generales, incluyendo mayor antigüedad crediticia promedio (12.18 años), una carga relativa de endeudamiento de 12.60% y la menor tasa de incumplimiento observada (16.20%). Por su parte, el perfil intermedio presenta características parcialmente similares al grupo de menor riesgo en términos de incumplimiento, aunque con diferencias más positivas en variables como nivel del ingreso comprometido con el crédito y tipo de vivienda, lo que permite considerarlo como un segmento con comportamiento diferenciado, aunque menos claramente separado desde una perspectiva de riesgo.

Al complementar el análisis con variables categóricas, se observó que el perfil de mayor riesgo concentra una mayor proporción de clientes con antecedentes de incumplimiento registrados y una distribución menos favorable en la calificación del préstamo, con menor presencia de categorías crediticias altas (B y C). Asimismo, dentro de este grupo de mayor riesgo se identificó una proporción de clientes con vivienda en arriendo de 49.50%, lo cual podría reflejar condiciones económicas menos estables en comparación con los perfiles de menor riesgo y un 19.60% de propósito del crédito para educación.

Desde una perspectiva visual, la representación de los clusters mediante reducción de dimensionalidad con PCA permitió observar una separación general entre los grupos identificados, aunque con cierto grado de superposición entre los perfiles de menor riesgo e intermedio, lo cual es consistente con los resultados descriptivos obtenidos (ver Figura 9). En términos analíticos, esta segmentación aporta una visión complementaria al problema del riesgo crediticio, permitiendo caracterizar perfiles de clientes con comportamientos financieros diferenciados más allá del enfoque estrictamente

predictivo. Aunque el clustering no constituye un mecanismo de clasificación supervisada, sí proporciona información útil para comprender la heterogeneidad de la población analizada y apoyar procesos de segmentación y toma de decisiones.

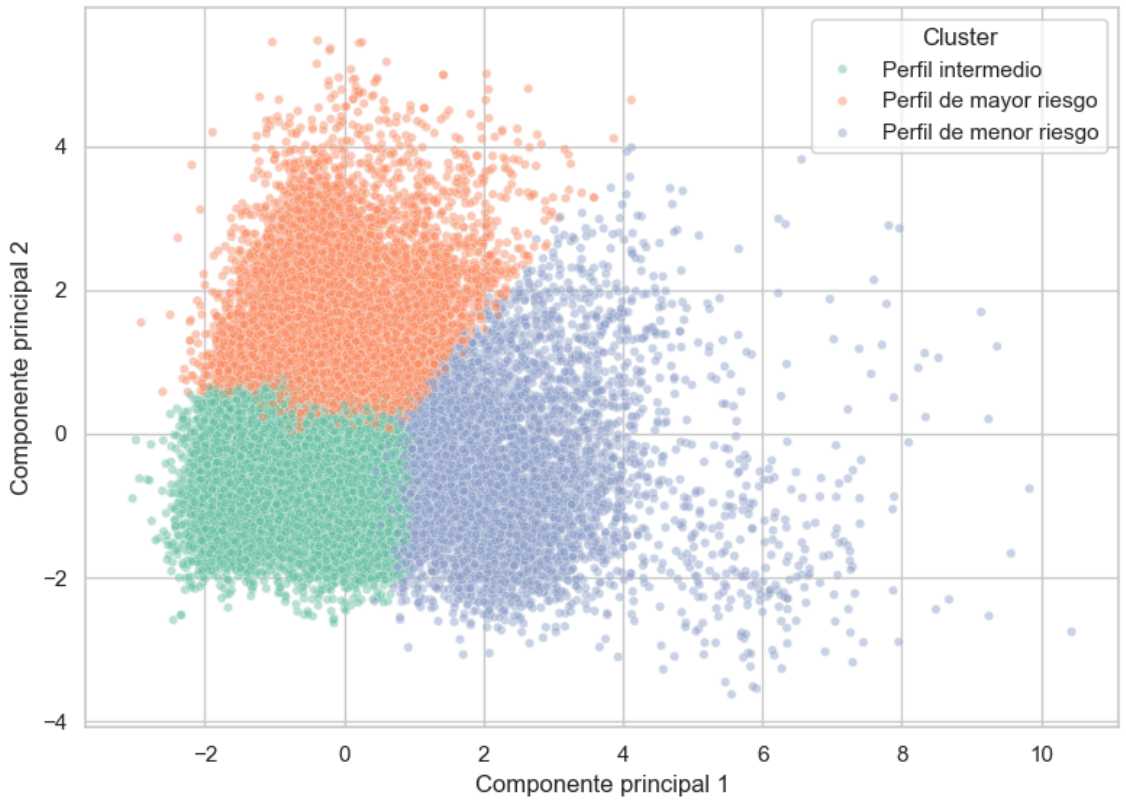

Figura 9. *Visualización 2D de clusters usando PCA*

3.3 Resultados del modelo XGBoost

Con el objetivo de evaluar el desempeño predictivo del riesgo de crédito, se implementaron dos modelos de clasificación basados en XGBoost: un modelo base construido con las variables originales del dataset y un segundo modelo que incorporó como característica adicional la variable de cluster obtenida a partir de la segmentación mediante K-Means. En ambos casos, el entrenamiento consideró el desbalance presente en la variable objetivo mediante el parámetro 'scale_pos_weight', calculado a partir de la proporción entre observaciones negativas y positivas en el conjunto de entrenamiento, con el fin de mejorar la identificación de clientes en cumplimiento del crédito.

El proceso de optimización de hiperparámetros mostró que tanto el modelo base como el modelo con cluster convergieron hacia la misma configuración óptima, incluyendo parámetros relacionados con profundidad máxima del árbol, número de estimadores, regularización, tasa de aprendizaje y criterios mínimos de partición. Este comportamiento sugiere que la incorporación de la variable de segmentación no generó cambios relevantes en la estructura de aprendizaje del algoritmo.

En términos de desempeño, ambos modelos presentaron resultados altamente competitivos y prácticamente equivalentes. El modelo base alcanzó un AUC-ROC de 0.9493, mientras que el modelo con incorporación del cluster obtuvo un AUC-ROC de 0.9492 (ver Figura 10), evidenciando una capacidad discriminativa muy alta en ambos modelos para diferenciar entre clientes con y sin incumplimiento. De manera similar, las métricas de clasificación muestran un comportamiento estable entre ambos enfoques, con valores de precisión cercanos al 82% y recall del 81% para la clase positiva, lo que indica una capacidad consistente para identificar clientes con riesgo de default.

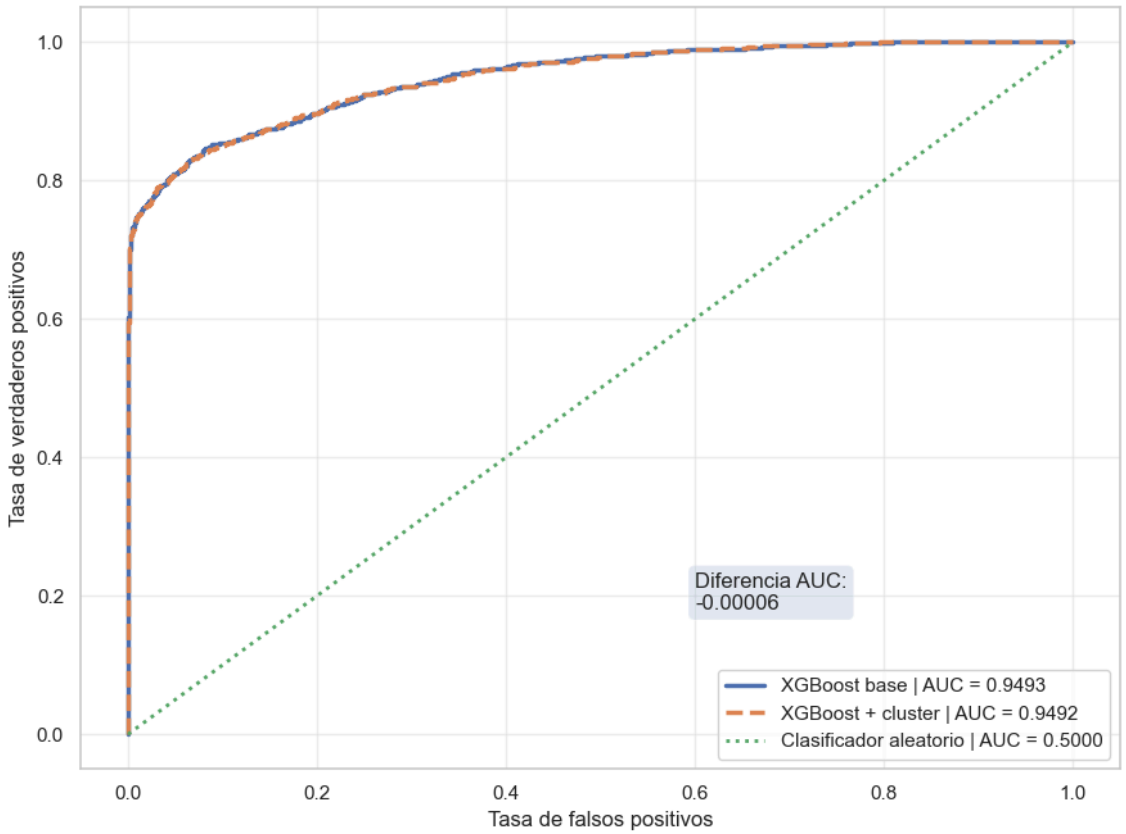

Figura 10. *Curva ROC: comparación de capacidad discriminante entre modelos*

La matriz de confusión refuerza esta estabilidad en el desempeño, mostrando diferencias mínimas entre ambos modelos. En el modelo base se identificaron correctamente 1.148 clientes en incumplimiento, mientras que el modelo con cluster identificó 1.147 casos, manteniendo prácticamente el mismo número de falsos negativos (270 para el modelo base frente a 271 para el modelo con cluster), al igual que los falsos positivos (249 y 250 respectivamente). Estos resultados sugieren que, desde una perspectiva estrictamente predictiva, ambos enfoques ofrecen un comportamiento muy similar.

En conjunto, los resultados evidencian que XGBoost constituye una alternativa robusta para el modelado del riesgo crediticio, logrando capturar patrones complejos dentro de la información financiera y sociodemográfica del dataset. Sin embargo, dado el comportamiento prácticamente equivalente entre ambos modelos, resulta innecesario añadir la incorporación de la segmentación al proceso predictivo ya que no aporta valor incremental significativamente alto para el pronóstico, no obstante, nos ayuda a visualizar más de cerca y tener los perfiles de riesgo crediticio.

Tabla 3. *Comparación desempeño de los modelos XGBoost*

Modelo	ROC AUC	Accuracy	Precision	Recall	F1-Score	Verdaderos Positivos	Verdaderos Negativos
XGBoost base	0.9493	0.9199	0.8218	0.8096	0.8156	1148	4815
XGBoost + cluster	0.9492	0.9196	0.8210	0.8089	0.8149	1147	4814

3.4 Impacto de la incorporación del clustering en el modelo predictivo

Aunque la segmentación mediante K-Means permitió identificar perfiles diferenciados de clientes desde una perspectiva descriptiva, su incorporación como variable adicional dentro del modelo XGBoost no generó mejoras significativas en el desempeño predictivo. Las métricas obtenidas muestran un comportamiento prácticamente idéntico entre el modelo base y el modelo enriquecido con la variable de cluster, evidenciando que la información aportada por la segmentación no incrementó la capacidad discriminativa del algoritmo.

Una posible explicación de este comportamiento es que las variables utilizadas para construir los clusters ya contenían la mayor parte de la información estructural relevante

para la predicción del incumplimiento. Dado que XGBoost tiene la capacidad de capturar relaciones no lineales e interacciones complejas entre variables, es razonable que el modelo ya estuviera incorporando implícitamente los patrones que posteriormente fueron resumidos mediante la clusterización. Adicionalmente, aunque el proceso de segmentación logró identificar con claridad un perfil de mayor riesgo, la separación entre los grupos de riesgo bajo e intermedio presenta cierto grado de superposición, lo que limita el valor informativo adicional del cluster como variable predictiva. Esto sugiere que, en este caso particular, la segmentación aporta mayor valor como herramienta de interpretación y caracterización de clientes que como mecanismo para mejorar directamente el desempeño de clasificación.

Desde una perspectiva aplicada, este hallazgo resulta relevante, ya que demuestra que no toda transformación o enriquecimiento de variables genera necesariamente mejoras predictivas. Sin embargo, el valor del clustering permanece desde el análisis estratégico, al facilitar la identificación de perfiles de comportamiento financiero que pueden ser útiles en procesos de segmentación comercial, gestión del riesgo y diseño de estrategias diferenciadas de intervención.

3.5 SHAP

Dado que los modelos basados en algoritmos de ensamble suelen ofrecer un alto desempeño predictivo, pero menor interpretabilidad frente a modelos más tradicionales se recurrió a SHAP (SHapley Additive exPlanations) como herramienta para explicar la contribución individual de las variables dentro del proceso predictivo (Lundberg & Lee, 2017). Este análisis resulta especialmente relevante en problemas de riesgo crediticio, donde no solo es importante alcanzar una alta capacidad de predicción, sino también comprender los factores que explican las decisiones del modelo.

Los resultados muestran que, a nivel de variables transformadas, los factores con mayor impacto sobre la predicción corresponden al ingreso del solicitante “person_income_log”, la proporción del préstamo respecto al ingreso “loan_percent_income” y la tasa de interés “loan_int_rate”. La figura 11 representa gráfica que explica el impacto individual de cada característica en la salida del modelo predictivo. El eje horizontal mide la magnitud del impacto (SHAP value), mientras que el color indica el valor de la variable: el rojo representa valores altos y el azul valores

bajos, esta gráfica evidencia que menores niveles de ingreso tienden a incrementar la probabilidad estimada de incumplimiento, mientras que una mayor carga financiera relativa y tasas de interés más elevadas generan un efecto positivo sobre el riesgo proyectado por el modelo. Estos hallazgos resultan consistentes con el comportamiento esperado dentro del contexto financiero, donde restricciones en capacidad de pago suelen estar asociadas con una mayor probabilidad de default.

En el caso de las variables categóricas, también se identificó una contribución relevante de factores como el tipo de vivienda, la calificación del préstamo y el propósito del crédito dentro del proceso de clasificación. No obstante, al consolidar la importancia de las variables codificadas hacia sus variables originales, se evidencia que la calificación del préstamo, “loan_grade”, representa el factor con mayor influencia agregada dentro del modelo, seguido por el ingreso del solicitante, el tipo de vivienda y el propósito del crédito. Este comportamiento sugiere que tanto las condiciones financieras objetivas del cliente como atributos asociados al contexto del crédito desempeñan un papel importante en la estimación del riesgo.

En términos generales, los resultados de interpretabilidad confirman que el modelo no basa sus decisiones en patrones arbitrarios, sino en variables con sentido financiero y comportamiento coherente con el problema analizado. Esto fortalece la confianza en el uso del modelo como herramienta de apoyo analítico, al proporcionar no solo una alta capacidad predictiva, sino también explicabilidad sobre los factores que impulsan la clasificación del riesgo crediticio.

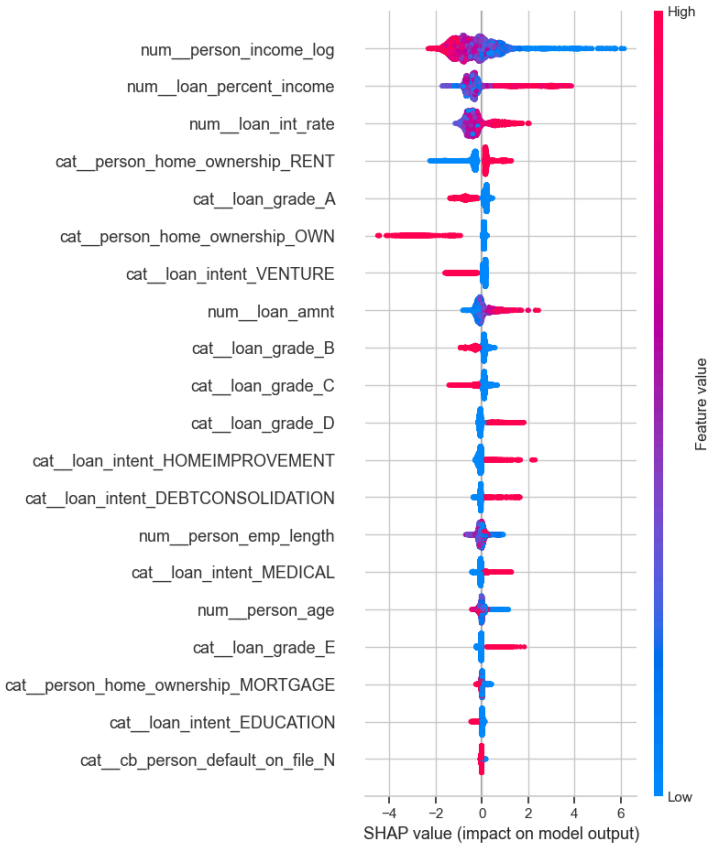

Figura 11. *Importancia e impacto de variables sobre la predicción del modelo XGBoost mediante SHAP*

4. Conclusiones

El presente proyecto tuvo como objetivo segmentar clientes y modelar el riesgo de incumplimiento crediticio mediante el uso de técnicas de análisis exploratorio, clustering y aprendizaje supervisado. Durante la etapa inicial se identificaron inconsistencias relevantes en el dataset, como problemas de escala en variables clave y la presencia de valores extremos no plausibles. La corrección de estos aspectos, junto con transformaciones como el logaritmo del ingreso y la normalización de variables proporcionales, permitió construir una base de datos coherente y adecuada para el modelado. Estas decisiones fueron fundamentales, ya que aseguraron que los algoritmos trabajaran sobre información representativa y evitaron sesgos derivados de errores en los datos.

El análisis exploratorio permitió identificar patrones claros asociados al riesgo de crédito y variables como el ingreso, la carga financiera y la tasa de interés mostraron diferencias significativas entre clientes que incumplen y aquellos que no, evidenciando

su relevancia en la predicción del riesgo. Asimismo, variables categóricas como la calificación del préstamo, el tipo de vivienda y el propósito del crédito mostraron comportamientos diferenciados, aportando información adicional sobre el perfil de los solicitantes.

El uso de K-Means permitió identificar tres grupos de clientes con características diferenciadas: Un grupo de alto riesgo, con menor capacidad de pago y mayor carga financiera; un grupo de bajo riesgo, con mayor estabilidad económica y mejores condiciones crediticias y por último un grupo intermedio, con características mixtas. Aunque esta segmentación no mejoró directamente el desempeño del modelo supervisado, sí aportó valor en la comprensión del comportamiento de los clientes, facilitando la identificación de perfiles de riesgo.

El modelo XGBoost obtuvo un desempeño alto, con un ROC AUC superior a 0.95, lo que indica una excelente capacidad para discriminar entre clientes que incumplen y los que no. Sin embargo, el análisis de métricas evidenció un aspecto importante: el recall de la clase de incumplimiento se ubicó alrededor del 81%, lo que implica que una proporción de clientes riesgosos no es detectada por el modelo. Esto refleja un trade-off entre precisión y sensibilidad que debe ser evaluado según el contexto de negocio. La inclusión de la variable de cluster no generó mejoras significativas en el desempeño, lo que sugiere que el modelo ya captura adecuadamente las relaciones presentes en los datos originales.

El análisis mediante SHAP permitió identificar los principales drivers del riesgo de crédito. Entre las variables más relevantes se destacan:

- La calificación del préstamo (loan_grade), como una variable que resume la evaluación crediticia.
- El ingreso del solicitante (person_income_log), con una relación inversa frente al riesgo.
- El porcentaje del ingreso comprometido (loan_percent_income), como uno de los factores más influyentes.
- La tasa de interés (loan_int_rate) y el monto del préstamo (loan_amnt), asociados positivamente al riesgo.

- Variables de perfil como el tipo de vivienda (person_home_ownership) y el propósito del crédito (loan_intent).

Además, se identificaron relaciones no lineales, particularmente en la carga financiera, donde a partir de ciertos niveles el riesgo aumenta de forma acelerada. También se evidenciaron interacciones entre variables, como el mayor riesgo asociado a clientes en arriendo bajo niveles similares de endeudamiento.

El proyecto demuestra que es posible construir un modelo robusto de riesgo de crédito combinando técnicas de análisis exploratorio, segmentación y machine learning. Si bien la segmentación mediante clustering no mejoró directamente el desempeño predictivo, sí contribuyó al entendimiento del comportamiento de los clientes. Por su parte, el modelo supervisado logró capturar de manera efectiva los factores clave del riesgo, ofreciendo una herramienta útil para la toma de decisiones. Finalmente, el uso de técnicas de interpretabilidad permitió validar el comportamiento del modelo, aumentando la confianza en sus resultados y facilitando su aplicación en contextos reales.

Referencias

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295. https://doi.org/10.3390/electronics9081295

Bao, C. (2021). K-means clustering algorithm: A brief review. Academic Journal of Computing & Information Science, 4(5), 37–40. https://doi.org/10.25236/AJCIS.2021.040506

Borrero, D., & Bedoya, O. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial. Revista UIS Ingenierías, 19(4), 37–52. https://doi.org/10.18273/revuin.v19n4-2020004

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Dunkeld, M. (s. f.). The importance of exploratory data analysis (EDA) in product data science. Medium. https://medium.com/@melissadunkeld/the-importance-of-exploratory-data-analysis-eda-in-product-data-science-fe5308965630

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería Investigación y Tecnología, 21(3), 1–16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer. https://doi.org/10.1007/978-3-319-10247-4

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3.ª ed.). Morgan Kaufmann.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A, 160(3), 523–541. https://doi.org/10.1111/1467-985X.00078

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Lao Tse. (2020). Credit Risk Dataset [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. En Advances in Neural Information Processing Systems (NeurIPS).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

Ulloa, G. C. (s. f.). Distribución normal y su importancia en el análisis de datos y en ciencia de datos. Medium. https://medium.com/@gladyschoqueulloa20/distribuci%C3%B3n-normal-y-su-importancia-en-el-an%C3%A1lisis-de-datos-y-en-ciencia-de-datos-5a7ad784771d