**Logotipo

Descripción generada automáticamente**

**Modelo prescriptivo para la mejora del score crediticio personalizado**

Alan Stiven Camacho Restrepo e Isabel Cristina Suárez Roldán

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

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2025

|  |  |
| --- | --- |
| **Cita** | (Camacho Restrepo & Suárez Roldán, 2025) |
| **Referencia**  **Estilo APA 7 (2020)** | Camacho Restrepo, A. S., & Suárez Roldán, I. C. (2025). *Modelo prescriptivo para la mejora del score crediticio personalizado*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteIX.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Texto de agradecimientos centrado.

**Tabla de contenido**

[Resumen 9](#_heading=h.3rdcrjn)

[Abstract 10](#_heading=h.26in1rg)

[1.](#_heading=h.35nkun2) Descripción del problema 11

[1.1.](#_heading=h.1ksv4uv) Problema de negocio 11

[1.2.](#_heading=h.44sinio) Aproximación desde la analítica de datos 12

[1.3.](#_heading=h.z337ya) Origen de los datos 12

[1.4.](#_heading=h.3j2qqm3) Métricas de desempeño [1](#_heading=h.z337ya)2

[2.](#_heading=h.4i7ojhp) Objetivos 14

[2.1.](#_heading=h.2xcytpi) Objetivo general [1](#_heading=h.4i7ojhp)4

[2.2.](#_heading=h.3whwml4) Objetivos específicos [1](#_heading=h.4i7ojhp)4

[3.](#_heading=h.2bn6wsx) Datos [1](#_heading=h.4i7ojhp)5

[3.1.](#_heading=h.qsh70q) Datos originales 15

[3.2.](#_heading=h.3as4poj) Datasets 20

[3.3.](#_heading=h.1pxezwc) Analítica descriptiva 21

[4.](#_heading=h.49x2ik5) Proceso de analítica 24

[4.1.](#_heading=h.2p2csry) Pipeline principal 24

[4.2.](#_heading=h.147n2zr) Preprocesamiento 24

[4.3.](#_heading=h.3o7alnk) Modelos 24

[4.4.](#_heading=h.23ckvvd) Métricas 25

[5.](#_heading=h.ihv636) Metodología 26

[5.1.](#_heading=h.32hioqz) Baseline 26

[5.2.](#_heading=h.1hmsyys) Validación 26

[5.3.](#_heading=h.41mghml) Iteraciones y evolución 26

[5.4 Herramientas](#_heading=h.2grqrue) 26

[6.](#_heading=h.vx1227) Resultados y discusión 27

[6.1.](#_heading=h.3fwokq0) Métricas 27

[6.2.](#_heading=h.1v1yuxt) Evaluación cualitativa 27

[6.3.](#_heading=h.4f1mdlm) Consideraciones de producción 27

[7.](#_heading=h.kgcv8k) Conclusiones 28

[8.](#_heading=h.34g0dwd) Recomendaciones 29

[Referencias](#_heading=h.1jlao46) 30

[Anexos](#_heading=h.43ky6rz) 32

**Lista de tablas**

[**Tabla 1**](#_heading=h.2u6wntf) Descripción de tipos de columnas [21](#_heading=h.3tbugp1)

**Lista de figuras**

[**Figura 3.1**](#_heading=h.2r0uhxc) Distribuciones de variables numéricas21

[**Figura 3.2**](#_heading=h.2r0uhxc) Distribuciones de variables numéricas luego del procesamiento22

[**Figura 3.3**](#_heading=h.2r0uhxc) Distribuciones de variable dicotómicas22

[**Figura 3.4**](#_heading=h.2r0uhxc) Matriz de correlación de las variables numéricas y variable respuesta23

**Siglas, acrónimos y abreviaturas**

**Cross-validation** Validación cruzada

**Grid Search** Búsqueda en rejilla

**Accuracy** Exactitud / Precisión

**ROC** Receiver Operating Characteristic  
**AUC** Area Under the Curve (Área bajo la curva)

**MSE**  Mean Squared Error (Error cuadrático medio)

**RMSE** Root Mean Squared Error (Raíz del error cuadrático medio)

**MAE** Mean Absolute Error (Error absoluto medio)

**SVM** Support Vector Machine (Máquinas de soporte vectorial)

**SQL** Structured Query Language (Lenguaje de consulta estructurado)

# Resumen

Este proyecto busca diseñar y construir un modelo prescriptivo que genere recomendaciones personalizadas para mejorar el score crediticio de los clientes, utilizando variables financieras relevantes. Para ello, se implementaron y evaluaron múltiples enfoques de machine learning, como máquinas de soporte vectorial (SVM), métodos de *boosting*, árboles de decisión y técnicas de aprendizaje por refuerzo, con el propósito de identificar patrones complejos en el comportamiento financiero individual. El modelo se entrenó utilizando una muestra representativa de 1.300.000 registros anonimizados, extraída de una base de datos proporcionada por Bancolombia que contiene más de 700 millones de observaciones sobre ingresos, egresos, endeudamiento, ahorro, inversiones y transacciones. Este volumen de datos exigió una rigurosa etapa de limpieza, transformación y agregación para resolver problemas como registros duplicados, valores nulos y presencia de *outliers*. Como componente clave de la solución, se desarrolló un *dashboard* interactivo que permite a los usuarios visualizar las recomendaciones generadas por el modelo, así como los principales indicadores relacionados con su perfil financiero y la evolución de su *score*. Este panel presenta de manera clara y dinámica el estado del cliente en las variables utilizadas, facilitando la comprensión de los factores que influyen en su puntaje y mostrando acciones concretas sugeridas con base en su historial. Los resultados obtenidos demuestran que el modelo no solo permite segmentar eficazmente a los clientes, sino que también entrega recomendaciones precisas y accionables, fortaleciendo la gestión del riesgo bancario, promoviendo la educación financiera y fomentando una relación más proactiva entre clientes y entidades financieras.

*https://github.com/Alan-Stiven-CR/TrabajoGrado.*

***Palabras clave*:** *Prescripción, crédito, aprendizaje, riesgo.*

# Abstract

This project aims to design and build a prescriptive model that generates personalized recommendations to improve clients’ credit scores using relevant financial variables. To achieve this, various machine learning approaches were implemented and evaluated, including Support Vector Machines (SVM), boosting methods, decision trees, and reinforcement learning techniques, with the goal of identifying complex patterns in individual financial behavior. The model was trained using a representative sample of 1,500,000 anonymized records, extracted from a dataset provided by Bancolombia containing over 700 million observations on income, expenses, debt, savings, investments, and transactions. The volume and complexity of the data required a thorough preprocessing phase, including cleaning, transformation, and aggregation to resolve issues such as duplicate entries, missing values, and outliers. A key component of the solution is an interactive dashboard that allows users to visualize the recommendations generated by the model, along with key indicators related to their financial profile and credit score evolution. This dashboard clearly and dynamically presents the client’s status across the variables used in the model, enabling users to understand the factors influencing their credit score and view specific actions suggested based on their financial history. The results show that the model not only enables effective customer segmentation but also delivers precise and actionable recommendations, strengthening credit risk management, promoting financial education, and fostering a more proactive relationship between clients and financial institutions.

*Keywords***:** *Prescription, credit, learning, risk.*

# Descripción del problema

Comprender los factores que inciden en la evolución del *score* crediticio y su impacto en la relación entre los clientes y las entidades bancarias representa el desafío central que da origen a este proyecto. A partir de esta problemática, se plantean diversos interrogantes relacionados con los comportamientos financieros que podrían estar asociados a una disminución en el *score*, las posibles señales tempranas de riesgo y las variables que explican dichos cambios. En este contexto, se considera que los modelos prescriptivos pueden ofrecer un enfoque adecuado para responder de manera concreta y fundamentada a estas inquietudes.

## Problema de negocio

El sector bancario colombiano enfrenta actualmente el desafío de mejorar la salud financiera de sus clientes, en un entorno marcado por altos niveles de morosidad, baja fidelización y limitada inclusión financiera. En particular, Bancolombia, como una de las principales entidades financieras del país, dispone de grandes volúmenes de información sobre sus usuarios, pero enfrenta dificultades para traducir esos datos en conocimiento útil que permita entender con precisión las necesidades individuales de cada cliente. Esta falta de comprensión personalizada dificulta la identificación temprana de comportamientos financieros que puedan representar un riesgo, y limita la capacidad del banco para anticiparse a situaciones críticas. Además, los modelos de evaluación crediticia utilizados actualmente en el sector bancario tienden a ser generales y estáticos, lo que restringe la posibilidad de actuar de manera proactiva ante señales de deterioro en el perfil crediticio del cliente. (World Bank, 2023). En este contexto, se vuelve crucial que el sector bancario no solo detecte riesgos, sino también oriente a sus clientes en el mejoramiento de su *score* crediticio, promoviendo así la inclusión financiera y fortaleciendo la gestión de riesgos.

## Aproximación desde la analítica de datos

Se desarrollarán modelos prescriptivos que permitirán detectar patrones en los datos financieros y transaccionales de los clientes, con el objetivo de anticipar comportamientos asociados al riesgo crediticio. Estos modelos servirán para identificar variables críticas que afectan la salud financiera de cada usuario, generar alertas tempranas y, de forma central, recomendar acciones personalizadas que orienten a los clientes en la mejora de su *score* crediticio, facilitando a las entidades bancarias priorizar la gestión de sus productos. Además, permitirán segmentar a los clientes con base en su perfil de riesgo, dando lugar a una comprensión más granular del ecosistema financiero y proporcionando insumos clave para diseñar estrategias diferenciales por grupo de comportamiento.

## Origen de los datos

Los datos utilizados en este proyecto provienen directamente de Bancolombia y fueron obtenidos a través de consultas SQL sobre múltiples tablas internas del sistema de información de la entidad. Estas tablas representan diferentes dimensiones del comportamiento financiero de los clientes, como ingresos, morosidad, transacciones, productos financieros y clasificaciones de riesgo. La información fue suministrada por el equipo de analítica del banco en un entorno seguro y bajo condiciones de confidencialidad, ya que se trata de datos sensibles y privados. Para garantizar la protección de la identidad de los clientes, los registros fueron anonimizados y las variables clave fueron adaptadas. Los datos abarcan un periodo comprendido entre agosto de 2021 y marzo de 2025, y su integración implicó procesos de limpieza, transformación y unión de tablas mediante llaves comunes, lo que permitió construir una base consolidada y lista para el análisis posterior.

## Métricas de desempeño

Para evaluar el desempeño del modelo prescriptivo propuesto, se considerarán dos tipos de métricas: técnicas relacionadas con la precisión del modelo de machine learning, y de negocio asociadas al impacto práctico de su implementación.  
Desde el enfoque técnico, se emplearán métricas de evaluación tanto para problemas de clasificación como de regresión, según la naturaleza del modelo implementado. Para los modelos clasificatorios, se utilizarán métricas como AUC, precisión y exactitud. Dado que se trabaja con datos potencialmente desbalanceados (por ejemplo, clientes con y sin sobreendeudamiento), la AUC se considera la métrica principal, ya que mide la capacidad del modelo para distinguir entre clases, independientemente de su proporción. Adicionalmente, en los escenarios que involucren predicciones numéricas, como la estimación de ingresos, *score* crediticio proyectado o capacidad de pago, se utilizarán métricas de regresión como el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE), que permiten cuantificar la precisión de las estimaciones continuas. Se espera alcanzar un valor mínimo de 0.75 en AUC, como umbral para considerar que el modelo tiene un desempeño aceptable en términos predictivos. Este estándar se fundamenta en referencias del sector y en la revisión bibliográfica (Noriega et al., 2023; Shi et al., 2022).

En cuanto a las métricas de negocio, se evaluará el impacto potencial de las recomendaciones generadas por el modelo prescriptivo, asumiendo escenarios simulados en los que los clientes siguen dichas sugerencias. Se medirán indicadores como la mejora proyectada en el *score* crediticio promedio, la reducción estimada en los niveles de mora y el porcentaje de clientes con recomendaciones de alto valor (es decir, aquellas que, si se implementaran, permitirían pasar de un grupo de riesgo alto a uno inferior, como de G7 a G6). Bajo este enfoque, se considera que si al menos un 15% de las recomendaciones emitidas por el modelo logran proyectar una mejora significativa en el *score* crediticio, se podría alcanzar un retorno sobre la inversión (ROI) favorable, al anticipar una disminución de provisiones por riesgo y una mayor eficiencia en la asignación de crédito. Asimismo, se estimará un potencial incremento del 10% en la retención de clientes, basado en el valor informativo del *dashboard* y la utilidad percibida de las recomendaciones, aunque este impacto también será evaluado mediante simulaciones y análisis contrafactuales.

En conclusión, el modelo será considerado viable para su despliegue si alcanza un AUC ≥ 0.75 en su desempeño predictivo y si al menos el 15% de las recomendaciones generadas proyectan una mejora significativa en el *score* crediticio de los clientes, como el cambio de un grupo de riesgo alto a uno inferior. Además, se espera que dichas recomendaciones, analizadas en escenarios simulados, evidencien un impacto financiero potencial positivo, ya sea por la reducción estimada del riesgo de mora o por el aumento proyectado en la fidelización, lo cual justificaría los costos asociados a la implementación, mantenimiento y operación del sistema prescriptivo.

# Objetivos

## Objetivo general

Analizar el comportamiento financiero de los clientes de Bancolombia para desarrollar un modelo que permita generar recomendaciones personalizadas orientadas a la mejora del *score* crediticio, contribuyendo así a la optimización de la gestión del riesgo por parte de la entidad financiera.

## Objetivos específicos

* Recolectar y preparar un conjunto de datos basado en la actividad transaccional de clientes proporcionado por Bancolombia, correspondiente al periodo comprendido entre los años 2021 y 2025, mediante la integración de múltiples tablas relacionadas con ahorro, inversiones y productos financieros, realizando procesos de limpieza, transformación y estandarización para garantizar la calidad y consistencia del base de datos final.
* Diseñar y construir un modelo prescriptivo capaz de generar recomendaciones personalizadas orientadas a la mejora del score crediticio, utilizando variables financieras significativas y explorando enfoques de *machine learning* como máquinas de soporte vectorial (SVM), métodos de *boosting*, árboles de decisión y aprendizaje por refuerzo para identificar patrones relevantes en los datos.
* Validar el modelo prescriptivo empleando datos históricos o simulados, aplicando técnicas como la validación cruzada (*cross-validation*) y la búsqueda de hiper parámetros mediante *grid search*, y evaluando su desempeño a través de curvas ROC, así como métricas cuantitativas como *accuracy*, AUC, MSE, RMSE y MAE, según la naturaleza de las predicciones.
* Desarrollar un *dashboard* interactivo que visualice las recomendaciones generadas por el modelo, junto con los principales indicadores del perfil financiero del cliente y la evolución de su *score* crediticio, permitiendo una interpretación clara de los factores que influyen en el puntaje y facilitando la comprensión de las acciones sugeridas.

# Datos

## Datos originales

**3.1.1** **Origen y naturaleza de los datos**

El proyecto utiliza una base de datos proporcionada por Bancolombia, construida a partir de múltiples fuentes internas que consolidan información financiera y transaccional de los clientes. Esta base fue diseñada específicamente para reflejar el comportamiento financiero individual de cada usuario en una frecuencia mensual, lo cual permite realizar un análisis longitudinal de su situación crediticia. La información fue recolectada y tratada mediante consultas SQL, que permitieron tanto la ingesta como la depuración y transformación inicial de los datos.

Es importante destacar que, por tratarse de datos sensibles provenientes de una entidad financiera, todo el conjunto fue anonimizado, eliminando cualquier posibilidad de identificar a los clientes. Además, los nombres originales de las tablas fueron modificados por razones de confidencialidad.

**3.1.2 Formato y estructura de los datos**

Los datos se encuentran almacenados en formato .parquet, lo cual facilita una gestión eficiente de grandes volúmenes de información. Cada registro representa la situación financiera de un cliente en un mes específico. Para la integración de las distintas tablas, se utilizan como claves la fecha (año y mes), el tipo de documento y el número de documento. El periodo analizado abarca desde agosto de 2021 hasta marzo de 2025, permitiendo un análisis temporal detallado sobre la evolución financiera de los clientes.

La base de datos original contiene aproximadamente 740 millones de registros distribuidos en 92 columnas, lo que representa un volumen considerable de información. Debido a esta magnitud, se optó por extraer una muestra aleatoria simple que mantiene la representatividad del conjunto original. Esta muestra está compuesta por 1.598.523 registros y tiene un tamaño total de 213,7 MB. Para facilitar su manejo en el repositorio del proyecto, fue dividida en 11 archivos .parquet, cada uno con un tamaño aproximado de 21,5 MB. Esta segmentación permite un procesamiento más ágil durante el desarrollo, análisis y entrenamiento de modelos, sin comprometer la validez estadística del estudio.

**3.1.3 Estructura de tablas y variables**

La base de datos está compuesta por 12 tablas principales, cada una con un enfoque temático:

* ***clientes***: Contiene información general del cliente, incluyendo tipo de cliente, estado y segmento económico.
* ***riesgo\_credito***: Almacena información sobre el comportamiento crediticio, incluyendo fechas de desembolso y días de mora.
* ***seguimiento\_desempleo***: Indicador binario que refleja si el cliente estaba desempleado (1) o empleado (0) en cada periodo.
* ***ingresos\_personales***: Detalla los ingresos mensuales por aportes a seguridad social, pagos variables y transacciones fijas.
* ***cuenta\_ahorro\_clientes***: Registra los saldos mensuales mínimos, máximos y promedio en cuentas de ahorro, expresados en salarios mínimos.
* ***modulos\_subcategorías***: Refleja el gasto mensual del cliente según categorías de consumo (compras básicas y no básicas).
* ***inversiones***: Registra si el cliente tuvo inversiones activas y cuántas realizó en el periodo correspondiente.
* ***cuotas\_financieras\_externas*** y ***cuotas\_financieras\_internas***: Reportan las cuotas mensuales de créditos contratados fuera del banco y dentro de la entidad, respectivamente.
* ***modulos\_categorias***: Contiene los montos transaccionados por canal (app, cajeros, tarjetas) y permite caracterizar el comportamiento digital del cliente.
* ***estimadores\_ingreso*s**: Incluye estimaciones mensuales de ingreso realizadas por diferentes metodologías internas.
* ***cupos***: Presenta información sobre el cupo total disponible en tarjetas de crédito, así como compras y avances realizados.
* ***grupos\_riesgo***: Clasifica a los clientes en niveles de riesgo del G1 (muy bajo) al G8 (muy alto), asignado por el área de riesgo del banco.

Estas tablas se cruzan utilizando las variables ***cod\_tipo\_doc, num\_doc, year*** *y* ***month***, lo que permite una integración mensual precisa para cada cliente en el universo de análisis.

**3.1.4** **Transformaciones y variables derivadas**

A partir de estas tablas se generaron más de 30 variables derivadas, diseñadas para enriquecer el análisis y alimentar los modelos prescriptivos. Estas variables incluyen indicadores de:

* **Razones de sobreendeudamiento** (por tipo de crédito y total).
* **Capacidad de pago** (real y estimada).
* **Estabilidad de ingresos y gastos**, calculada con ventanas móviles de seis meses.
* **Uso del crédito**, incluyendo razones de compras y avances frente a ingresos o cupo disponible.
* **Ahorro**, tanto en volumen como en proporción del ingreso.

Asimismo, se crearon *flags* binarios para representar situaciones críticas o deseables, como sobreendeudamiento alto, estabilidad financiera o comportamiento saludable de ahorro.

|  |  |  |
| --- | --- | --- |
| **Tipo de columna** | **Descripción** | **Tipo de dato** |
| **cliente\_id** | Identificador único de cada cliente, compuesto por tipo y número de documento. | Categórico |
| **fecha** | Fecha de referencia del registro, expresada en año y mes. | Categórico |
| **segmento\_cliente** | Clasificación del cliente según nivel de ingresos o tipo (Personal, Preferencial, Plus, etc.). | Categórico |
| **estado\_laboral** | Indicador binario que señala si el cliente estuvo desempleado en el periodo (1: desempleado, 0: empleado). | Numérico |
| **ingresos\_mensuales** | Ingresos del cliente durante el mes, incluyendo ingresos por seguridad social, transacciones fijas y variables, y estimaciones por modelos internos. | Numérico |
| **saldos\_ahorro** | Información de saldos mensuales en cuentas de ahorro (mínimo, máximo y promedio), expresados en salarios mínimos. | Numérico |
| **inversiones\_cliente** | Registro de inversiones activas y cantidad total realizadas por el cliente en el periodo. | Numérico |
| **cuotas\_credito** | Cuotas mensuales pagadas por el cliente, separadas por productos internos y externos (hipoteca, vehículo, tarjeta). | Numérico |
| **gastos\_transaccionales** | Gastos mensuales del cliente en compras básicas y no básicas, expresados en salarios mínimos. | Numérico |
| **uso\_credito\_tarjeta** | Información del cupo disponible, compras realizadas y avances en tarjetas de crédito. Incluye razones de uso frente a cupo o ingreso. | Numérico |
| **grupo\_riesgo** | Nivel de riesgo asignado por el área de riesgo del banco, clasificado en grupos G1 (riesgo bajo) a G8 (riesgo alto). | Categórico |
| **capacidad\_pago** | Razones que relacionan ingresos con consumo básico y obligaciones crediticias, tanto internas como externas. | Numérico |
| **sobreendeudamiento** | Razones que expresan el nivel de endeudamiento del cliente respecto a sus ingresos estimados o consolidados. | Numérico |
| **estabilidad\_financiera** | Indicadores calculados con ventanas móviles de seis meses que evalúan la consistencia de ingresos y gastos (transacciones, ahorro, etc.). | Numérico |
| **indicadores\_flag\_binarios** | Variables binarias derivadas que indican condiciones específicas como ahorro regular, uso excesivo del crédito, sobreendeudamiento o estabilidad mensual. | Numérico |

**Tabla 1.** Descripción de tipos de columnas.

**3.1.5 Etiquetado para aprendizaje supervisado**

En el contexto de aprendizaje automático supervisado, la variable objetivo corresponde a la clasificación del cliente en grupos de riesgo crediticio, representada por la variable *valor\_g* en la base de datos final, codificada de G1 (riesgo bajo) a G8 (riesgo alto).

**3.1.6** **Modo de acceso y restricciones**

El acceso a los datos se realizó mediante consultas SQL en entornos controlados por la entidad bancaria. El código fuente de cada etapa de procesamiento se encuentra documentado y almacenado en un repositorio Git, con referencias cruzadas a cada tabla (Ver Anexo 1).

Por tratarse de información confidencial, el conjunto de datos está restringido exclusivamente al equipo del proyecto, sin posibilidad de reutilización externa. Además, cualquier divulgación debe cumplir con los lineamientos legales y contractuales del banco, asegurando el respeto por la privacidad de los datos de los clientes.

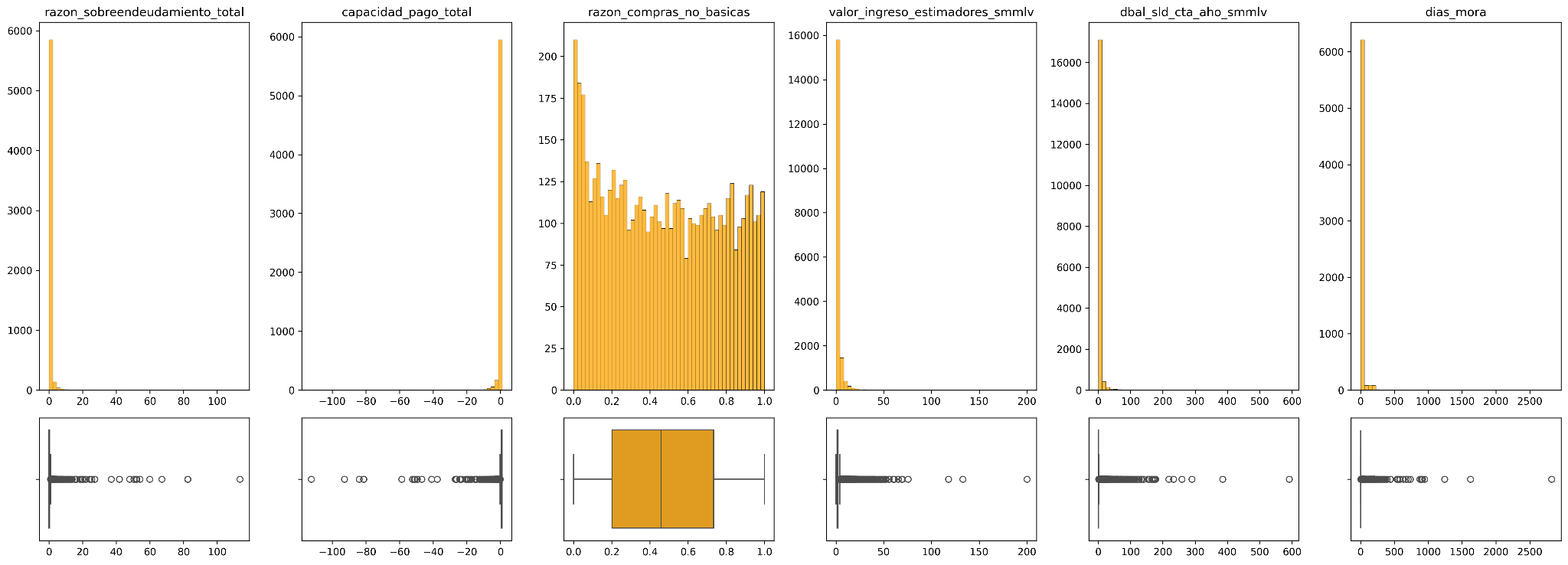
## Datasets

Antes de construir los datasets de entrenamiento y validación, se eliminaron los registros con valores nulos en la variable objetivo *valor\_g*, ya que imposibilitan el entrenamiento supervisado del modelo al no contar con una clase de referencia.

Para construir los *datasets* de entrenamiento y validación, se partió de la base de datos final ya procesada, la cual contiene la variable objetivo *valor\_g* con ocho clases posibles (de G1 a G8). Con el fin de garantizar que la distribución de las clases fuera representativa y balanceada en ambos subconjuntos, se aplicó una técnica de muestreo estratificado utilizando la función *train\_test\_split* de *scikit-learn.* (Pedregosa et al., 2011)

Este proceso permitió dividir los datos en un 70 % para entrenamiento y 30 % para validación, preservando la proporción original de cada grupo de riesgo crediticio. De esta manera, se asegura que el modelo pueda aprender adecuadamente de todas las clases y que su desempeño sea evaluado de forma justa sobre un conjunto con la misma estructura de clases que el original.

## Analítica descriptiva

A continuación, se presentan algunas visualizaciones exploratorias del conjunto de datos utilizados para analizar el comportamiento de las variables numéricas y dicotómicas. Se incluyen histogramas que permiten observar la distribución general de estas variables, así como gráficos de cajas (*boxplots*) para detectar la presencia de valores atípicos (ver Figura 3.1). Adicionalmente, se muestran gráficos actualizados tras aplicar técnicas de eliminación de *outliers*, lo que permite una representación más precisa y equilibrada de la información (ver Figura 3.2). Estas visualizaciones son fundamentales para comprender la dispersión, la asimetría y la concentración de los datos, aspectos clave en la preparación del modelo prescriptivo.

***Figura 3.1*** *Histogramas de variables numéricas.*

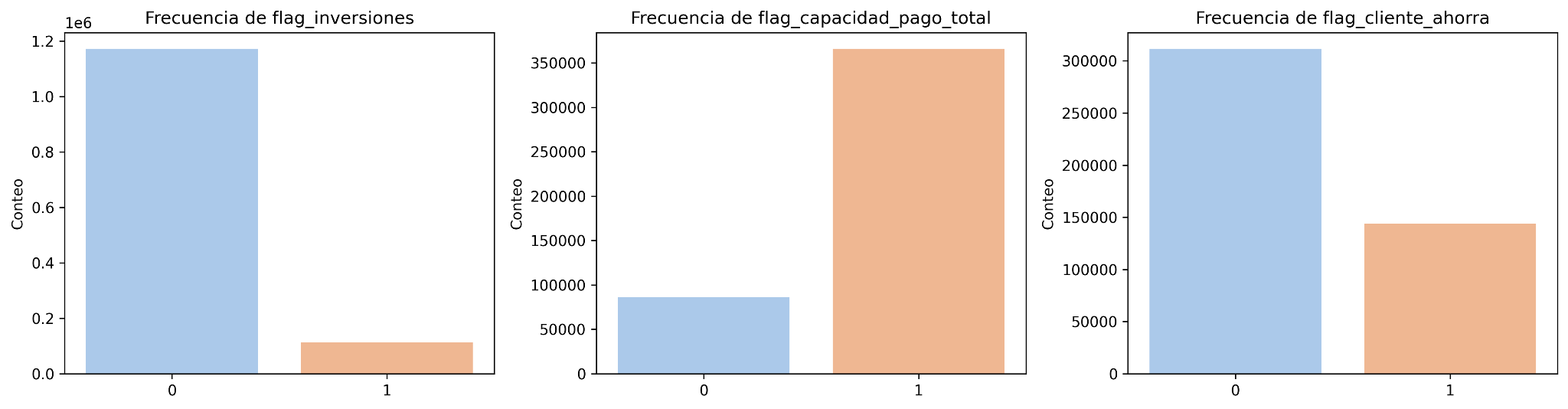
Las variables numéricas presentan distribuciones altamente sesgadas hacia la derecha, con la mayoría de los datos concentrados cerca de cero y presencia significativa de valores atípicos, especialmente en variables como ingresos, saldo, días de mora y sobreendeudamiento (ver Figura 3.1). Esto sugiere que, aunque la mayoría de los individuos tienen niveles bajos en estas métricas, existen casos extremos que podrían influir en los análisis. Solo una variable muestra una distribución más uniforme (*razon\_compras\_no\_basicas*), mientras que otras evidencian asimetría y dispersión elevadas, lo cual indica la necesidad de tratar *outliers* y considerar transformaciones para análisis posteriores.

# 

***Figura 3.2*** *Histogramas de variables numéricas después de preprocesamiento.*

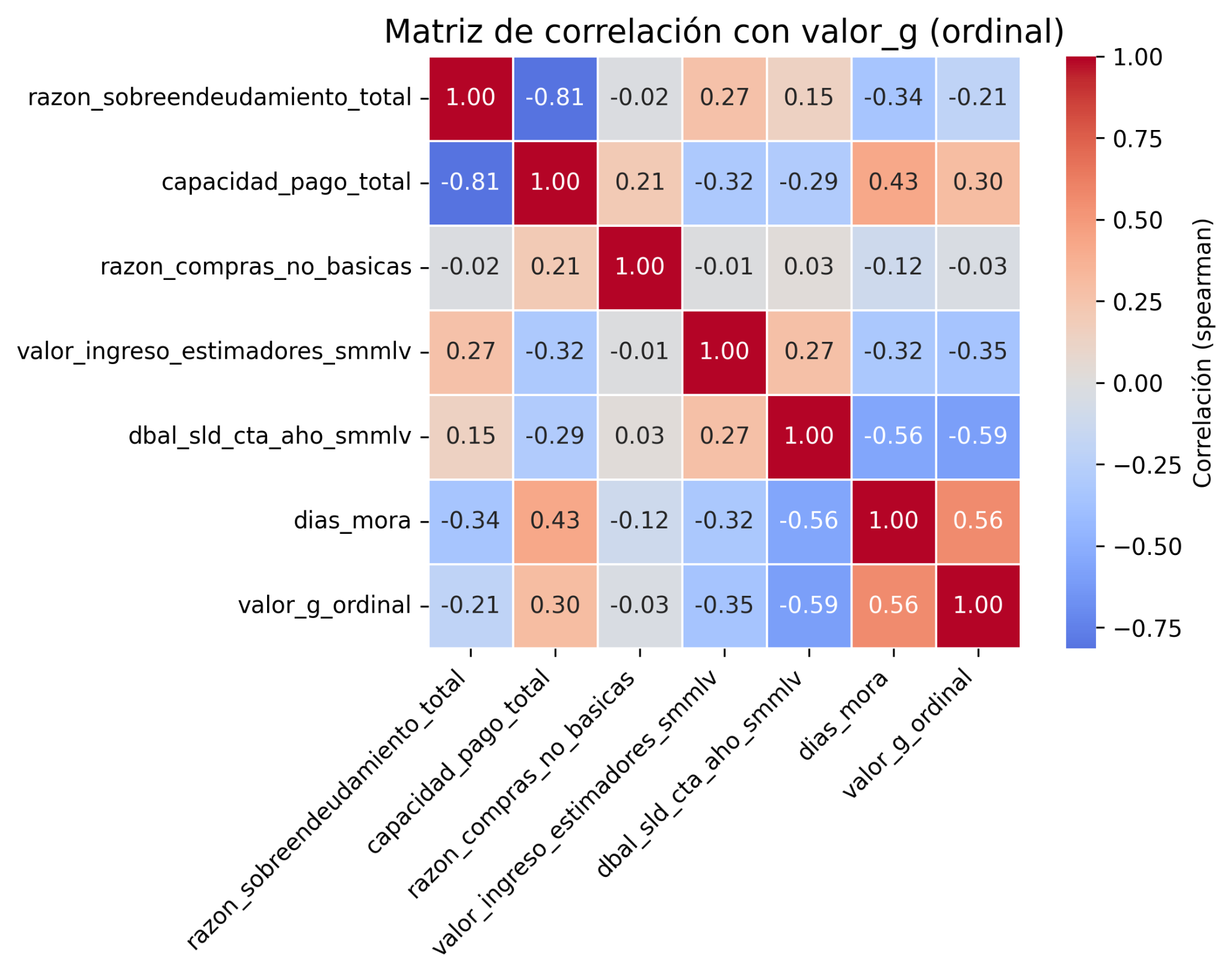
# Tras eliminar los *outliers*, las distribuciones de las variables numéricas se muestran más compactas y simétricas, lo que facilita su análisis e interpretación. Aunque algunas variables mantienen cierta asimetría (especialmente ingresos y saldos), en general se observa una reducción significativa de la dispersión extrema. La mayoría de los datos ahora se agrupan en rangos más coherentes, y los *boxplots* reflejan una menor presencia de valores extremos, lo cual mejora la calidad de los datos para análisis estadísticos o modelado predictivo.

# 



***Figura 3.3*** *Distribuciones de variables dicotómicas.*

A partir de la Figura 3.3, se observa que las variables dicotómicas analizadas muestran distribuciones desbalanceadas. En el caso de *flag\_inversiones*, la mayoría de los individuos no realizan inversiones, mientras que en *flag\_capacidad\_pago\_total*, predomina la población con capacidad de pago adecuada. Por otro lado, en *flag\_cliente\_ahorra*, la proporción de clientes que no ahorran supera ampliamente a quienes sí lo hacen. Estas diferencias reflejan comportamientos financieros contrastantes y podrían tener un impacto importante en la segmentación o el análisis de riesgo financiero.



***Figura 3.4*** *Matriz de correlación.*

La matriz de correlación mostrada en la Figura 3.4, evidencia relaciones importantes entre las variables financieras y el valor objetivo *valor\_g\_ordinal*. Se destacan correlaciones negativas moderadas con *dbal\_sld\_cta\_aho\_smmlv* y *valor\_ingreso\_estimadores\_smmlv*, lo cual sugiere que a mayor saldo en ahorro o ingresos, menor es la categoría ordinal del valor objetivo, asociado a menor riesgo o necesidad financiera. Asimismo, se observa una correlación positiva con *dias\_mora*, lo que indica que a mayor número de días en mora, mayor es el nivel ordinal del valor objetivo, reflejando mayor riesgo. Estas correlaciones ayudan a identificar variables relevantes en la explicación o predicción del comportamiento financiero.

En resumen, las gráficas muestran una alta concentración de valores en los extremos para varias variables numéricas, lo que motivó la eliminación de *outliers* para mejorar la distribución. Las variables dicotómicas presentan desequilibrios importantes en sus frecuencias, y la matriz de correlación revela que factores como la capacidad de pago, los ingresos y los días en mora tienen una relación significativa con la variable objetivo, siendo útiles como primer paso a la etapa de modelamiento.

# 4. Proceso de analítica

## 4.1 Pipeline principal

Describe con un gráfico el flujo de trabajo general de los datos en tu proyecto. Añade o modifica secciones si lo consideras necesario

## 4.2 Preprocesamiento

Describe las alternativas de preprocesamiento de datos que consideraste, aumentación de datos, etc.

## 

## 4.3 Modelos

Describe los distintos modelos que consideraste, sus configuraciones, etc.

## 

## 4.4 Métricas

Describe cómo calculas las métricas de desempeño ML y de negocio (Por ejemplo: con qué funciones de sklearn, tensorflow, etc.)

# 5. Metodología

## 5.1 Baseline

Describe tu primera iteración, sus resultados y los problemas técnicos que tuviste que resolver

## 5.2 Validación

Describe tu proceso de validación, particiones de train/test/validación, etc.

## 5.3 Iteraciones y evolución

Describe en qué te enfocaste en las iteraciones subsiguientes. No es necesario hacer un listado exhaustivo de las iteraciones, pero sí has de describir qué perseguías y qué obtuviste con cada grupo de iteraciones. P.ej. primero hicimos un conjunto de iteraciones para definir las opciones de preprocesamiento, luego hicimos unas iteraciones probando distintos modelos de tal tipo, etc.

## 5.4 Herramientas

Menciona las herramientas que usaste para tu proyecto

# 6. Resultados y discusión

En los resultados se comunican los hallazgos y descubrimientos del estudio. Se incluyen tablas, figuras, diagramas y demás material demostrativo. Al narrar descriptivamente una figura, tabla, etc., en un párrafo, puedes insertar una referencia cruzada, es decir, un hipervínculo al elemento mencionado dentro o fuera de paréntesis, ejemplos: estos resultados se muestran en la **Tabla 1**. Igualmente, los datos son validados con otros instrumentos (**Tabla 2**, **Tabla 3**). Lineamientos que se establecen en la nueva versión de las Normas APA séptima edición (**Figura 1**). La producción intelectual institucional se publica en el Repositorio (**Figura 2**).

La discusión es la interpretación crítica y el análisis de los resultados, que surgen de las preguntas de investigación.

## 6.1 Métricas

Describe los resultados numéricos de las métricas de las iteraciones que consideres más relevantes, junto con las configuraciones. Usa tablas o gráficas siguiendo el formato mostrado a continuación.

## 6.2 Evaluación cualitativa

Realiza una evaluación cualitativa de los resultados, indicando casos de overfitting, underfitting, utilidad de los resultados, relación obtenida entre la métrica de ML y la de negocio, etc.

## 6.3 Consideraciones de producción

Menciona consideraciones técnicas de una posible puesta en producción. Por ejemplo: condiciones de monitoreo del desempeño de los modelos, integración con streams de datos, servicios en la nube, etc.

# 

# 7. Conclusiones

Son las interpretaciones finales que recopilan los datos de la investigación, describe lo que se obtuvo, qué se logró y cuáles son los resultados. Guardan relación directa con lo que se mencionó en el planteamiento del problema y en los objetivos. Haz un breve juicio crítico de tu proyecto.

# 8. Recomendaciones

Las recomendaciones son las futuras y posibles líneas de investigación que llevarán a resolver problemas relacionados con la presente investigación.

# Referencias

World Bank. (2023). *Financial inclusion overview*.   
<https://www.worldbank.org/en/topic/financialinclusion/overview>

Alan-Stiven-CR. (2025). *Monografía entrega 2* [Jupyter notebook]. GitHub.<https://github.com/Alan-Stiven-CR/TrabajoGrado/blob/main/Code/Monografia_entrega_2.ipynb>

Noriega, J. P., Rivera, L. A., & Herrera, J. A. (2023). Machine Learning for Credit Risk Prediction: A Systematic Literature Review. *Data*, 8(11), 169. <https://doi.org/10.3390/data8110169>

Shi, S., Tse, R., Luo, W., D’Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-022-07472-2>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research, 12*, 2825–2830.

<https://jmlr.org/papers/v12/pedregosa11a.html>

# Anexos

1. Anexo 1. Tablas de información utilizadas para el proyecto.

Repositorio: <https://github.com/Alan-Stiven-CR/TrabajoGrado>

Carpeta específica: SQL