Segmentación de cartera y modelo predictivo de mora en propiedad horizontal mediante técnicas de machine learning

Augusto Lis Moncaleano

Asesor

Felipe Alexander Pipicano Guzmán

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Nota de Aceptación

_____

Felipe Alexander Pipicano Guzmán

_____ _____

Jurado Jurado

2026

Resumen

Este proyecto aplicó técnicas de ciencia de datos y aprendizaje automático para analizar la cartera en mora de un conjunto residencial de propiedad horizontal conformado por 576 apartamentos en Bogotá, Colombia, donde la gestión de recaudo se basaba exclusivamente en procedimientos operativos y reportes descriptivos, limitando el aprovechamiento analítico de la información. Bajo la metodología CRISP-DM, y empleando datos anonimizados sobre detalle de deuda, antigüedad de mora y comportamiento de pago, se realizó análisis exploratorio, ingeniería de variables, segmentación de cartera mediante K-Means —identificando unidades de bajo riesgo y riesgo crítico— y modelado predictivo supervisado con Regresión Logística, Random Forest y XGBoost, siendo este último superado por Random Forest, que alcanzó un F1-score de 0.880 y un ROC-AUC de 0.901. Los resultados permitieron identificar las variables asociadas al deterioro de la cartera y priorizar acciones de recaudo basadas en evidencia, concluyendo que la ciencia de datos constituye una herramienta efectiva para fortalecer la gestión administrativa y financiera en propiedad horizontal, contribuyendo a una toma de decisiones más objetiva y eficiente.

Palabras clave: aprendizaje automático, predicción de mora, segmentación de cartera, propiedad horizontal, ciencia de datos, Random Forest, K-Means, evaluación del riesgo crediticio.

Abstract

This project applied data science and machine learning techniques to analyze the delinquent portfolio of a horizontal property residential complex comprising 576 apartments in Bogotá, Colombia, where collection management relied exclusively on operational procedures and descriptive reports, limiting the analytical use of available information. Following the CRISP-DM methodology, and using anonymized data on debt detail, delinquency aging, and payment behavior, exploratory data analysis, feature engineering, portfolio segmentation using K-Means —identifying low-risk and critical-risk units— and supervised predictive modeling with Logistic Regression, Random Forest, and XGBoost were conducted, with Random Forest achieving the best performance, reaching an F1-score of 0.880 and a ROC-AUC of 0.901. The results enabled the identification of variables associated with portfolio deterioration and the prioritization of evidence-based collection actions, concluding that data science constitutes an effective tool for strengthening administrative and financial management in horizontal property, contributing to more objective and efficient decision-making.

Keywords: machine learning, delinquency prediction, portfolio segmentation, horizontal property, data science, Random Forest, K-Means, credit scoring.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Planteamiento del Problema

En los conjuntos residenciales de propiedad horizontal, la gestión de la facturación, el recaudo y la mora constituye un componente fundamental para garantizar la sostenibilidad financiera, el cumplimiento presupuestal y la transparencia administrativa. En el caso del conjunto residencial objeto de estudio, conformado por 576 apartamentos en la ciudad de Bogotá, la administración cuenta con registros históricos de facturación y pagos que permiten dar cumplimiento a las obligaciones operativas mensuales; sin embargo, esta información no se aprovecha mediante herramientas analíticas que permitan anticipar comportamientos financieros y respaldar la toma de decisiones.

Actualmente, se dispone de datos relacionados con pagos, mora, cuotas extraordinarias y ejecución presupuestal; no obstante, estos registros son utilizados principalmente para reportes descriptivos y control operativo. En consecuencia, la administración no cuenta con un enfoque analítico que facilite la identificación de patrones históricos de pago, el análisis del comportamiento de la mora y la generación de información útil para la gestión financiera del conjunto residencial.

La ausencia de este enfoque limita la capacidad de anticipar escenarios de incumplimiento en los pagos, reconocer variables asociadas al riesgo de mora y utilizar el comportamiento histórico del recaudo como insumo para la toma de decisiones administrativas. Esta situación evidencia una brecha entre la disponibilidad de datos financieros y su aprovechamiento estratégico mediante técnicas de ciencia de datos y analítica.

En un contexto donde la analítica de datos se ha consolidado como soporte para la gestión organizacional, resulta pertinente formular un enfoque que permita transformar la información histórica de facturación y pagos en conocimiento útil para la toma de decisiones

administrativas. Por ello, surge la necesidad de aplicar técnicas de análisis exploratorio de datos, segmentación y modelado predictivo que permitan comprender el comportamiento financiero del conjunto, identificar perfiles de riesgo y estimar la probabilidad de mora crítica.

En consecuencia, el problema de investigación se centra en determinar cómo aplicar técnicas de ciencia de datos para identificar patrones de comportamiento de cartera, estimar el riesgo de mora y generar insumos analíticos que apoyen la toma de decisiones financieras y administrativas en un conjunto residencial de propiedad horizontal en la ciudad de Bogotá, a partir del análisis de datos históricos de facturación y pagos.

A partir de lo anterior, se formula la siguiente pregunta de investigación: *¿De qué manera la aplicación de técnicas de ciencia de datos y aprendizaje automático —incluyendo análisis exploratorio, segmentación de cartera y modelos supervisados de clasificación— permite identificar patrones de comportamiento financiero, estimar el riesgo de mora crítica y generar insumos analíticos que apoyen la toma de decisiones de recaudo en un conjunto residencial de propiedad horizontal en la ciudad de Bogotá, a partir del análisis de datos históricos de cartera?*

Justificación

En los conjuntos residenciales de propiedad horizontal, la gestión de la cartera y el recaudo constituye un componente esencial para garantizar la sostenibilidad financiera, la continuidad operativa y la adecuada administración de los recursos comunes. En el caso del conjunto residencial objeto de estudio, conformado por 576 apartamentos en la ciudad de Bogotá, la administración dispone de información asociada al comportamiento de la cartera, incluyendo saldos por unidad, composición de la deuda, antigüedad de mora y estado de cobro. No obstante, estos datos se utilizan principalmente con fines operativos y descriptivos, sin un aprovechamiento analítico que permita comprender de manera más profunda los patrones de mora y apoyar la toma de decisiones frente a la gestión del recaudo.

Actualmente, la información disponible permite identificar valores adeudados, fechas del último pago, componentes de deuda e incluso niveles de antigüedad de cartera; sin embargo, su uso se limita al seguimiento administrativo de las obligaciones pendientes. En consecuencia, no se cuenta con herramientas analíticas que faciliten la caracterización de perfiles de deudores, la identificación de variables asociadas a mayores niveles de criticidad y la priorización de unidades con mayor riesgo de encontrarse en estados de cobro prejurídico o jurídico.

La ausencia de este enfoque limita la capacidad de la administración para diferenciar tipos de deudores según su comportamiento financiero, reconocer patrones relevantes en la evolución de la mora y orientar de manera más estratégica las acciones de recaudo. Esta situación evidencia una brecha entre la disponibilidad de datos de cartera y su aprovechamiento como insumo para la generación de conocimiento útil mediante técnicas de ciencia de datos y analítica.

En este sentido, el desarrollo del proyecto se justifica por la necesidad de aplicar técnicas de análisis exploratorio de datos, ingeniería de variables, segmentación y modelos supervisados de clasificación que permitan transformar la información histórica disponible en herramientas de apoyo para la gestión administrativa y financiera del conjunto residencial. Este enfoque permite pasar de una revisión principalmente descriptiva de la cartera a una interpretación analítica orientada a la identificación de perfiles de riesgo, priorización de cobro y generación de alertas tempranas.

Desde el punto de vista práctico, el proyecto aporta una metodología que puede apoyar a la administración en la toma de decisiones basadas en evidencia, permitiendo focalizar esfuerzos sobre las unidades con mayor criticidad y diseñar estrategias diferenciadas de recaudo. Desde el punto de vista académico, el trabajo demuestra la pertinencia de la ciencia de datos y el aprendizaje automático en un contexto poco explorado como la propiedad horizontal, aportando una aplicación concreta de modelos analíticos para la gestión de cartera y riesgo de mora.

Objetivos

Objetivo General

Desarrollar un modelo de clasificación predictiva que permita estimar el riesgo de mora en propiedad horizontal, a partir de variables históricas de deuda y comportamiento de pago, con el fin de apoyar la toma de decisiones y la priorización de estrategias de recaudo en un conjunto residencial de la ciudad de Bogotá.

Objetivos Específicos

Caracterizar el comportamiento de la cartera mediante técnicas de análisis exploratorio de datos, identificando patrones y variables relevantes asociadas al riesgo de mora.

Preparar los datos históricos de cartera mediante procesos de limpieza, integración, transformación e ingeniería de variables analíticas que representen el comportamiento financiero de las unidades residenciales.

Aplicar técnicas de segmentación mediante K-Means para identificar perfiles de deudores según características de riesgo y comportamiento de mora.

Construir modelos supervisados de clasificación, incluyendo Regresión Logística, Random Forest y XGBoost, para estimar el riesgo de mora en propiedad horizontal y comparar su desempeño predictivo.

Analizar la importancia de las variables predictoras y el desempeño de los modelos desarrollados, con el propósito de generar recomendaciones orientadas a la gestión y priorización del recaudo.

Marcos de Referencia

Marco Conceptual

Para el desarrollo del presente proyecto, es necesario precisar los conceptos fundamentales que orientan su enfoque analítico y metodológico, en relación con la segmentación de cartera y el modelo predictivo de mora en propiedad horizontal mediante técnicas de machine learning.

Ciencia de Datos

Es una disciplina interdisciplinaria que combina estadística, programación, análisis computacional y conocimiento del contexto de aplicación, con el propósito de extraer patrones, generar conocimiento útil y apoyar la toma de decisiones a partir de los datos (Géron, 2022). En este proyecto, la ciencia de datos constituye la base para el análisis de la cartera en mora y la construcción de modelos orientados a la identificación de perfiles de deudores y al riesgo de mora.

Aprendizaje Automático

Corresponde a un conjunto de técnicas y algoritmos que permiten a los sistemas aprender patrones a partir de los datos, sin necesidad de ser programados explícitamente para cada caso particular. Su propósito es identificar relaciones, clasificar observaciones o generar predicciones a partir del comportamiento histórico de las variables (James et al., 2023). En el presente estudio, se utilizó para segmentar la cartera y construir un modelo predictivo de mora en el contexto de la propiedad horizontal.

Análisis Exploratorio de Datos

Es una fase del análisis de datos orientada a comprender la estructura, calidad, distribución y comportamiento de la información disponible. Incluye técnicas estadísticas y

visuales que permiten identificar tendencias, relaciones entre variables, valores atípicos, inconsistencias y patrones preliminares (Han et al., 2022). En este proyecto, el análisis exploratorio fue fundamental para caracterizar la cartera, reconocer comportamientos de pago y establecer insumos para la etapa de modelado.

Propiedad Horizontal

Es una forma de organización jurídica y administrativa en la cual coexisten bienes privados y bienes comunes, regulados por normas que establecen derechos y obligaciones para los copropietarios (Ministerio de Justicia y del Derecho, 2012). En este contexto, la administración del conjunto residencial depende, entre otros aspectos, del pago oportuno de cuotas de administración y demás obligaciones económicas por parte de los residentes, lo cual da lugar a procesos de recaudo y gestión de cartera.

Cartera

Se entiende como el conjunto de obligaciones económicas pendientes de pago a favor de una organización. En el ámbito de la propiedad horizontal, la cartera está conformada por las deudas que registran los propietarios o residentes por conceptos como cuotas de administración, intereses de mora, parqueaderos, cuotas extraordinarias u otros cobros asociados. Su análisis resulta fundamental para evaluar la salud financiera de la administración.

Cartera en Mora

Hace referencia al conjunto de obligaciones vencidas cuyo pago no ha sido realizado dentro del plazo establecido. La mora representa una condición de incumplimiento que puede variar en severidad según el tiempo transcurrido, el monto adeudado y la reincidencia del deudor. En este proyecto, la cartera en mora constituye el objeto principal de análisis.

Mora

Es el retraso o incumplimiento en el pago oportuno de una obligación financiera. En términos administrativos y financieros, la mora refleja un deterioro en el comportamiento de pago y puede generar efectos como intereses moratorios, acumulación de deuda y escalamiento a procesos de cobro más estrictos. Su estudio permite identificar patrones de incumplimiento y niveles de criticidad dentro de la cartera.

Antigüedad de Mora

Corresponde al tiempo durante el cual una obligación ha permanecido vencida sin ser cancelada. Suele expresarse en tramos o rangos, como 1 a 30 días, 31 a 60 días, 61 a 90 días y más de 90 días. Esta variable es especialmente relevante en el análisis de cartera, ya que permite medir la severidad temporal del incumplimiento y establecer prioridades de gestión de recaudo.

Comportamiento de Pago

Se refiere a la forma en que una unidad residencial atiende sus obligaciones financieras a lo largo del tiempo, considerando variables como frecuencia de pago, oportunidad, valor cancelado, retrasos y recurrencia de mora. El análisis del comportamiento de pago permite identificar patrones diferenciales entre unidades y constituye una base relevante para la segmentación y predicción.

Segmentación de Cartera

Es el proceso analítico mediante el cual se agrupan unidades o deudores con características similares, utilizando variables relacionadas con deuda, antigüedad de mora, frecuencia de pago o nivel de criticidad. La segmentación permite clasificar la cartera en perfiles homogéneos, facilitando la interpretación de los datos y la definición de estrategias diferenciadas de recaudo (Han et al., 2022).

Modelo Predictivo

Es una representación matemática, estadística o algorítmica construida a partir de datos históricos, cuyo propósito es estimar el comportamiento de una variable de interés. En este proyecto, el modelo predictivo se orienta a estimar el riesgo de mora a partir de características asociadas a la deuda, su antigüedad y el comportamiento de pago registrado en la información analizada.

Riesgo de Mora

Se entiende como la probabilidad de que una unidad residencial presente un comportamiento de incumplimiento o un nivel de mora asociado a mayor criticidad. Este riesgo puede estimarse a partir de variables históricas que reflejen el estado de la deuda, la antigüedad de la mora, los intereses acumulados y la dinámica de pago. Su análisis busca apoyar la priorización de acciones de recaudo.

Variable Objetivo o Dependiente

Es la variable que se busca explicar, clasificar o predecir dentro de un modelo analítico. En el marco del presente proyecto, la variable objetivo estará asociada al riesgo de mora, entendido como el resultado que se pretende estimar mediante técnicas de machine learning.

Variables Explicativas o Independientes

Son aquellas características que aportan información para explicar el comportamiento de la variable objetivo. En este estudio pueden incluirse variables como deuda total, intereses de mora, antigüedad de la deuda, días sin pago, valor del último pago y composición de la cartera por concepto. Estas variables permiten construir perfiles analíticos y alimentar el modelo predictivo.

Ingeniería de Variables

Es el proceso mediante el cual se transforman o construyen nuevas variables a partir de los datos originales, con el fin de mejorar la capacidad explicativa de los modelos. En este proyecto, la ingeniería de variables puede incluir indicadores como días sin pago, proporción de deuda en mora superior a 90 días, concentración de intereses o número de conceptos con saldo pendiente.

Clasificación

Es una técnica de aprendizaje supervisado cuyo objetivo es asignar cada observación a una categoría previamente definida, con base en patrones aprendidos a partir de los datos (James et al., 2023). En el presente estudio, la clasificación fue utilizada para estimar el nivel de riesgo de mora de las unidades residenciales, de acuerdo con su comportamiento histórico.

Recaudo

Corresponde al valor efectivamente recibido por concepto de pagos realizados por los residentes en un periodo determinado. En términos financieros, el recaudo constituye un indicador clave para la sostenibilidad de la propiedad horizontal, ya que influye directamente en la capacidad operativa y presupuestal de la administración. Aunque el proyecto se centra en la cartera en mora, el recaudo representa un referente importante para interpretar el comportamiento de pago.

En síntesis, el marco conceptual del proyecto se sustenta en la integración de conceptos propios de la ciencia de datos, la analítica y la gestión de cartera en propiedad horizontal, con el fin de establecer una base teórica clara para el análisis, la segmentación y la predicción del comportamiento de la mora.

Marco Teórico

El presente proyecto se fundamenta en los principios de la ciencia de datos aplicada al análisis de cartera, entendida como un campo interdisciplinario que integra métodos estadísticos, técnicas computacionales y conocimiento del dominio para extraer información útil a partir de los datos (Géron, 2022). Desde esta perspectiva, la ciencia de datos permite transformar registros operativos en conocimiento analítico orientado a comprender fenómenos, reconocer patrones y apoyar la toma de decisiones en contextos donde el comportamiento financiero constituye un factor determinante para la gestión administrativa.

En el ámbito de la propiedad horizontal, la administración financiera depende en gran medida del cumplimiento oportuno de las obligaciones económicas por parte de los copropietarios o residentes. La presencia de cartera en mora afecta la liquidez, limita la capacidad operativa de la administración y genera la necesidad de fortalecer los procesos de recaudo mediante herramientas que permitan no solo describir la deuda existente, sino también comprender su comportamiento y establecer criterios de priorización. En este contexto, la aplicación de la ciencia de datos resulta pertinente para abordar la cartera desde una perspectiva analítica, permitiendo examinar su composición, antigüedad, comportamiento de pago y nivel de criticidad.

Uno de los componentes fundamentales en proyectos de ciencia de datos es el análisis exploratorio de datos (EDA), el cual permite examinar la estructura, calidad, distribución y relaciones presentes en la información antes de la construcción de modelos analíticos. A través de herramientas estadísticas y visuales, el EDA facilita la identificación de patrones, valores atípicos, inconsistencias y tendencias relevantes en variables asociadas a la deuda (Han et al., 2022), la mora y el comportamiento de pago. En el presente proyecto, esta fase constituye un

paso esencial para caracterizar la cartera y establecer una base sólida para las etapas posteriores de segmentación y modelado predictivo.

De manera complementaria, el proceso de preparación de los datos y la construcción de variables analíticas adquieren especial relevancia dentro del marco teórico del proyecto. La calidad de los resultados en ciencia de datos depende en gran medida de la capacidad para integrar, depurar y transformar la información disponible en variables que representen adecuadamente el fenómeno de estudio. En este sentido, la ingeniería de variables permite construir indicadores derivados, tales como días sin pago, concentración de deuda en tramos de mora avanzada, proporción de intereses sobre el total adeudado o número de conceptos con saldo pendiente, los cuales aportan una representación más significativa del comportamiento de la cartera.

Otro referente central es el aprendizaje automático o machine learning, entendido como el conjunto de técnicas algorítmicas que permiten identificar patrones en los datos y generar modelos capaces de clasificar, agrupar o estimar comportamientos a partir de información histórica (James et al., 2023). Dentro de este campo, el proyecto articula dos enfoques analíticos complementarios: la segmentación y el modelado predictivo. La segmentación permite agrupar unidades con características similares en función de variables relacionadas con la deuda, la antigüedad de mora y la dinámica de pago, favoreciendo la identificación de perfiles homogéneos de cartera. Esta aproximación resulta útil para comprender la heterogeneidad de los deudores y establecer estrategias diferenciadas de gestión.

Por su parte, el modelado predictivo constituye una herramienta orientada a estimar el comportamiento de una variable de interés a partir de patrones aprendidos en los datos históricos. En el contexto del presente estudio, el modelo predictivo busca estimar el riesgo de mora a partir

de variables relacionadas con el monto adeudado, la severidad temporal de la deuda, la composición de la cartera y el comportamiento de pago registrado. Este enfoque se sustenta en técnicas de aprendizaje supervisado, las cuales utilizan una variable objetivo definida para entrenar algoritmos capaces de clasificar nuevas observaciones según su nivel de riesgo.

Dentro de los modelos supervisados, técnicas como la regresión logística, los bosques aleatorios (Random Forest) y algoritmos de gradiente reforzado como XGBoost han sido ampliamente utilizados en problemas de clasificación, debido a su capacidad para modelar relaciones entre variables y generar estimaciones útiles para procesos de decisión (Chen & Guestrin, 2016; Breiman, 2001). Su pertinencia en estudios de cartera radica en que permiten evaluar la contribución de distintos factores al comportamiento de la mora y construir herramientas orientadas a la priorización de casos con mayor nivel de criticidad.

Asimismo, el proyecto reconoce la importancia metodológica de la segmentación mediante algoritmos no supervisados, como K-Means, los cuales permiten identificar grupos naturales dentro de los datos sin necesidad de una variable objetivo previa (Han et al., 2022). Esta técnica resulta especialmente útil cuando se busca clasificar unidades según perfiles de deuda y comportamiento, aportando una visión complementaria al modelo predictivo. Mientras la segmentación favorece la comprensión estructural de la cartera, la clasificación supervisada aporta una estimación del riesgo, de modo que ambos enfoques fortalecen de manera conjunta el análisis propuesto.

En línea con lo anterior, la predicción de mora se relaciona con los modelos de riesgo financiero y los sistemas de clasificación de deudores, cuyo propósito es estimar la probabilidad de incumplimiento de una obligación económica a partir del comportamiento histórico de pago y variables asociadas a la deuda. En este contexto, el credit scoring constituye una metodología

orientada a clasificar unidades según su nivel de riesgo mediante el uso de información cuantitativa y modelos estadísticos o de aprendizaje automático (Thomas et al., 2002).

Aunque el presente proyecto no corresponde a un entorno bancario, el principio analítico es similar, ya que busca identificar patrones de incumplimiento y clasificar unidades residenciales según su riesgo de mora en propiedad horizontal. Variables como antigüedad de la deuda, mora superior a 90 días, intereses acumulados y número de tramos vencidos permiten representar el deterioro financiero de la cartera y constituyen insumos relevantes para el modelado predictivo.

Los modelos de clasificación de cartera permiten diferenciar unidades con comportamiento financiero regular de aquellas con mayor probabilidad de incumplimiento, facilitando la priorización de acciones de recaudo y la generación de alertas tempranas. En este tipo de problemas, donde suele existir desbalance entre clases, métricas como precisión, recall, F1-score y ROC-AUC resultan más apropiadas que la exactitud global para evaluar el desempeño de los modelos.

En consecuencia, la incorporación de técnicas de segmentación y modelos predictivos fortalece el análisis de cartera en propiedad horizontal, permitiendo transformar información histórica en herramientas de apoyo para la toma de decisiones administrativas y financieras basadas en evidencia.

Desde una perspectiva metodológica, el desarrollo del proyecto se apoyó en el modelo CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente reconocido en proyectos de minería de datos y ciencia de datos por su carácter estructurado y flexible (Chapman et al., 2000). Esta metodología organiza el proceso analítico en fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y

despliegue o comunicación de resultados. Su adopción permite asegurar coherencia entre el problema planteado, la naturaleza de los datos disponibles y las técnicas analíticas seleccionadas, contribuyendo a la rigurosidad del estudio.

En el análisis de cartera, también resulta relevante considerar que los problemas de clasificación suelen presentar desbalance entre categorías, especialmente cuando los casos críticos o de mayor severidad son menos frecuentes que los casos normales (Géron, 2022). Desde el punto de vista teórico, esta condición implica la necesidad de utilizar métricas de evaluación que no se limiten a la exactitud global, sino que incorporen medidas como precisión, recall, F1-score y curvas ROC-AUC, con el fin de valorar adecuadamente la capacidad del modelo para identificar observaciones de interés (Géron, 2022; Velásquez & Ramírez, 2021). Este aspecto es particularmente importante en contextos de cartera, donde la correcta identificación de unidades de mayor riesgo tiene un valor operativo superior a una clasificación general aparentemente alta.

En el contexto de la propiedad horizontal, la incorporación de técnicas de ciencia de datos, segmentación y aprendizaje automático representa una alternativa pertinente para fortalecer la gestión del recaudo. Más allá del uso tradicional de reportes descriptivos, estas herramientas permiten avanzar hacia una comprensión más profunda del comportamiento de la cartera, facilitando la identificación de perfiles de deudores, la estimación del riesgo de mora y la definición de estrategias de recaudo basadas en evidencia. De esta manera, la analítica se convierte en un soporte para una gestión más objetiva, eficiente y orientada a la toma de decisiones.

En consecuencia, el marco teórico del proyecto sustenta la integración del análisis exploratorio de datos, la ingeniería de variables, la segmentación y el modelado predictivo como un enfoque analítico aplicable al estudio de la cartera en mora en propiedad horizontal. Así, la

ciencia de datos y el machine learning no se conciben únicamente como herramientas tecnológicas, sino como fundamentos metodológicos que permiten transformar datos históricos en insumos estratégicos para la administración y la gestión del recaudo.

Metodología

El presente proyecto se desarrolló bajo un enfoque cuantitativo y aplicado, orientado al análisis de datos históricos relacionados con la cartera en mora en un conjunto residencial de propiedad horizontal. La investigación tuvo como propósito caracterizar el comportamiento de la cartera, identificar perfiles de deudores y desarrollar un modelo predictivo de mora mediante técnicas de machine learning, a partir de información asociada a la deuda, la antigüedad de mora y el comportamiento de pago de las unidades residenciales.

Para el desarrollo metodológico se adoptó el modelo CRISP-DM (Cross Industry Standard Process for Data Mining), ampliamente utilizado en proyectos de ciencia de datos por su capacidad para estructurar de manera sistemática las etapas de comprensión del problema, tratamiento de los datos, construcción de modelos y evaluación de resultados. En este proyecto, la metodología se organizó en seis fases interrelacionadas:

Fase 1 Comprensión del Negocio

En esta etapa se realizó la caracterización del contexto administrativo y financiero de la propiedad horizontal, con énfasis en la gestión de cartera y recaudo. Se identificó la problemática asociada al manejo de la mora, la necesidad de contar con herramientas analíticas para segmentar la cartera y la importancia de estimar el riesgo de mora como apoyo a la toma de decisiones. Asimismo, se definieron el objetivo general, los objetivos específicos y las preguntas que orientaron el desarrollo del proyecto.

Fase 2 Comprensión de los Datos

En esta fase se efectuó la revisión y exploración inicial de las fuentes de información disponibles, correspondientes al detalle de cartera por unidad y a la antigüedad de mora por tramos. Se verificó la estructura, calidad, consistencia y completitud de los datos, así como la

correspondencia entre las variables presentes en ambas fuentes. De manera complementaria, se aplicaron técnicas de análisis exploratorio de datos (EDA), incluyendo estadísticas descriptivas y visualizaciones, con el fin de identificar patrones de deuda, distribución de la mora, relaciones entre variables, valores atípicos y comportamientos relevantes para el análisis.

Fase 3 Preparación de los Datos

Esta fase comprendió la limpieza, integración, transformación y estructuración de la información para su posterior uso en los procesos de segmentación y modelado predictivo. Se realizó el tratamiento de valores faltantes, la revisión de inconsistencias, la depuración de registros y la normalización de variables cuando sea requerido por los algoritmos seleccionados. Adicionalmente, se llevó a cabo la construcción de variables derivadas o analíticas, tales como días sin pago, concentración de deuda en mora avanzada, proporción de intereses sobre la deuda total y número de conceptos con saldo pendiente, con el fin de enriquecer la representación del fenómeno estudiado.

Fase 4 Modelado

En esta etapa se desarrollaron dos componentes analíticos complementarios. En primer lugar, se aplicaron técnicas de segmentación para identificar grupos de cartera con características similares en términos de deuda, antigüedad de mora y comportamiento de pago. Para ello, se emplearon algoritmos no supervisados, como K-Means, con el propósito de establecer perfiles homogéneos de deudores.

En segundo lugar, se construyó un modelo predictivo de mora mediante técnicas de aprendizaje supervisado, orientado a estimar el riesgo de mora a partir de las variables históricas disponibles. Para esta finalidad, se evaluaron algoritmos como regresión logística, Random Forest y XGBoost, seleccionando el más adecuado según su desempeño, capacidad de

generalización e interpretabilidad. Dado que la variable objetivo puede presentar desbalance entre categorías, se contempló la aplicación de técnicas de ajuste como ponderación de clases o métodos de sobremuestreo, cuando sea metodológicamente pertinente.

Fase 5 Evaluación

Los modelos desarrollados fueron evaluados mediante métricas acordes con la naturaleza de cada técnica. En el caso de la segmentación, se emplearon indicadores como el índice de silhouette y criterios de interpretabilidad de los grupos obtenidos. Para el modelo predictivo, se utilizaron métricas de clasificación como precisión, recall, F1-score, matriz de confusión y curva ROC-AUC, prestando especial atención a la capacidad del modelo para identificar correctamente las unidades con mayor nivel de riesgo. Esta fase permitió valorar la utilidad analítica y práctica de los modelos en el contexto de la gestión de cartera.

Fase 6 Interpretación de Resultados y Recomendaciones

Finalmente, los resultados obtenidos fueron interpretados a la luz del contexto administrativo y financiero del conjunto residencial. A partir de los hallazgos derivados del análisis exploratorio, la segmentación y el modelo predictivo, se formularon recomendaciones orientadas a fortalecer la gestión del recaudo, priorizar estrategias de cobro y mejorar el aprovechamiento de la información histórica para la toma de decisiones. De esta manera, el proyecto buscó aportar una base analítica que respalde una gestión de cartera más objetiva, eficiente y sustentada en evidencia.

Caracterización y Descripción de la Base de Datos

Presentación General de la Base de Datos

La base de datos utilizada en el presente proyecto representa el estado financiero de las unidades residenciales frente a sus obligaciones económicas con la administración del conjunto. Su contenido permite analizar el comportamiento de la cartera, la antigüedad de la mora, la acumulación de intereses y los estados de cobro asociados a cada apartamento.

La unidad de análisis corresponde a cada apartamento o unidad residencial, por lo que cada registro del dataset representa una unidad financiera independiente. Esto permitió estudiar el comportamiento individual de pago, identificar patrones de incumplimiento y construir perfiles de riesgo asociados a la mora.

El propósito analítico de esta base de datos fue servir como insumo para el análisis exploratorio de datos, la segmentación de cartera y el desarrollo de modelos supervisados de clasificación orientados a estimar el riesgo de mora crítica.

Fuente de Datos

La información fue suministrada por la administración de un conjunto residencial de propiedad horizontal ubicado en la ciudad de Bogotá, conformado por 576 apartamentos. Los datos fueron entregados en formato Excel y anonimizados para proteger la identificación directa de las unidades residenciales.

Las fuentes principales utilizadas fueron:

Tabla 1

Fuentes de Datos Utilizadas en el Proyecto

Fuente	Descripción
CART_MORA_apto_apto.xlsx	Contiene información general de cartera, saldos pendientes, intereses, estado jurídico y comportamiento financiero de las unidades.
Cart_Mora_Tiempo.xlsx	Contiene la distribución de la deuda por tramos de antigüedad de mora.

Nota. Los archivos fueron suministrados por la administración del conjunto residencial. Ambas fuentes fueron integradas mediante el identificador interno de cada unidad residencial.

Ambas fuentes fueron integradas mediante el identificador interno de cada unidad residencial, consolidando un dataset analítico único para las etapas de limpieza, transformación, análisis exploratorio, segmentación y modelado predictivo.

Periodo Temporal de los Datos

La información analizada corresponde a un corte histórico de cartera suministrado por la administración del conjunto residencial durante el año 2025. Los registros reflejan el estado financiero acumulado de las unidades residenciales al momento de la extracción de la información.

Cantidad de Registros y Variables

Tabla 2

Cantidad de Registros y Variables

Fuente	Registros	Variables
CART_MORA_apto_apto.xlsx	578	21
Cart_Mora_Tiempo.xlsx	578	14

Nota. Los registros corresponden al estado financiero por unidad al momento del corte de información.

Después del proceso de integración y depuración, se consolidó un dataset final compuesto por 578 registros y 35 variables analíticas.

Tipo de Variables

Las 35 variables del dataset consolidado fueron clasificadas de acuerdo con su naturaleza estadística y su función dentro del análisis. Se identificaron variables numéricas continuas, relacionadas principalmente con valores monetarios de deuda, intereses y saldos por tramos de mora; variables numéricas discretas, asociadas a conteos o medidas de tiempo como meses de mora, días sin pago y número de tramos vencidos; variables categóricas, correspondientes a estados administrativos o identificadores de las unidades residenciales; y variables binarias, utilizadas para representar la presencia o ausencia de condiciones específicas de mora.

Esta clasificación permitió comprender mejor la estructura de los datos y definir el tratamiento adecuado para cada variable durante las etapas de limpieza, transformación, segmentación y modelado predictivo. Las variables monetarias permitieron medir la magnitud de la deuda, las variables temporales facilitaron el análisis de la antigüedad de la mora, las variables

categóricas aportaron contexto administrativo y las variables binarias permitieron construir indicadores analíticos y la variable objetivo del modelo.

Para el modelado supervisado no se utilizaron las 35 variables originales, sino una selección de 11 variables predictoras con mayor capacidad discriminante, evitando redundancias y posibles problemas de colinealidad. Entre ellas se incluyeron variables como cuotas_adm, interes_mora, promedio, dias_sin_pago, mas_90_dias, ratio_mora_90, num_tramos_mora y prop_interes, las cuales fueron relevantes para diferenciar unidades de bajo y alto riesgo. La variable tiene_mora_90 fue excluida por ser componente binario directo de la variable objetivo, mientras que tiene_mora fue excluida por ser proxy redundante del mismo fenómeno.

Descripción Operacional de las Variables

Con el propósito de comprender adecuadamente el comportamiento financiero de las unidades residenciales y fortalecer la interpretación analítica del modelo predictivo, se realizó una descripción operacional de las variables utilizadas en el estudio. Esta etapa permitió definir el significado funcional de cada variable, su forma de interpretación dentro del contexto de cartera y, en los casos correspondientes, el procedimiento utilizado para su construcción mediante ingeniería de variables.

La descripción operacional resulta relevante dentro del proceso metodológico, ya que facilita la trazabilidad de los datos y permite comprender cómo cada variable representa aspectos específicos relacionados con la deuda, la antigüedad de mora, los intereses acumulados y el comportamiento histórico de pago. Asimismo, esta caracterización aporta claridad sobre la función analítica de las variables dentro de las etapas de segmentación y modelado predictivo.

Las variables utilizadas en el proyecto pueden clasificarse en dos grupos principales: variables originales provenientes de las fuentes de información suministradas por la

administración del conjunto residencial y variables derivadas construidas durante la etapa de preparación de datos e ingeniería de características.

Variables Originales

Las variables originales corresponden a los campos financieros y administrativos presentes en las fuentes de cartera y antigüedad de mora. Estas variables describen directamente el estado de deuda de cada unidad residencial.

Tabla 3

Descripción de Variables Originales

Variable	Definición operacional	Interpretación analítica
cuotas_adm	Valor pendiente por concepto de cuotas de administración	Representa el principal componente de deuda de la unidad
interes_mora	Valor acumulado por intereses generados debido al incumplimiento en pagos	Refleja deterioro financiero y persistencia de mora
_1_30_dias	Saldo correspondiente a obligaciones vencidas entre 1 y 30 días	Identifica mora temprana
_31_60_dias	Saldo correspondiente a obligaciones vencidas entre 31 y 60 días	Representa mora intermedia
_61_90_dias	Saldo correspondiente a obligaciones vencidas entre 61 y 90 días	Indica deterioro progresivo de cartera
mas_90_dias	Saldo correspondiente a obligaciones vencidas por más de 90 días	Representa mora crítica y mayor riesgo financiero

Variable	Definición operacional	Interpretación analítica
promedio	Promedio de antigüedad de mora expresado en meses	Permite medir persistencia temporal de la deuda
juridico	Estado administrativo de la unidad respecto al proceso de cobro	Permite identificar casos jurídicos o prejurídicos

Nota. Las variables corresponden a los campos financieros y administrativos de las fuentes originales de cartera.

Variables Derivadas o Construidas Mediante Ingeniería de Variables

Con el fin de enriquecer la capacidad explicativa del modelo y representar de manera más precisa el comportamiento financiero de las unidades residenciales, se construyeron variables derivadas a partir de la información original. Estas variables permitieron capturar relaciones, proporciones y patrones de mora relevantes para el análisis predictivo.

Tabla 4

Descripción de Variables Derivadas

Variable	Definición operacional	Forma de cálculo / interpretación
dias_sin_pago	Número de días transcurridos desde el último pago registrado	Calculada a partir de la diferencia entre la fecha de corte y la fecha del último pago
ratio_mora_90	Proporción de la deuda total ubicada en mora superior a 90 días	Se calcula dividiendo mas_90_dias sobre el total de cartera
num_tramos_mora	Cantidad de tramos de mora con saldo activo	Conteo de tramos con valores mayores a cero

Variable	Definición operacional	Forma de cálculo / interpretación
prop_interes	Proporción de intereses sobre el total adeudado	Relación entre interes_mora y deuda total
tiene_mora	Indicador binario de existencia de mora activa	Valor 1 si la unidad presenta saldo vencido; 0 en caso contrario
target_riesgo_mora	Variable objetivo-utilizada en el modelo predictivo	Toma valor 1 cuando la unidad presenta mora superior a 90 días o estado jurídico/prejurídico

Nota. mediante ingeniería de variables aplicada durante la etapa de preparación de datos.

La construcción de estas variables permitió representar de manera más adecuada el fenómeno de estudio, especialmente en relación con la severidad temporal de la mora, la concentración de deuda crítica y el deterioro progresivo del comportamiento de pago. Variables como ratio_mora_90, num_tramos_mora y prop_interes aportaron capacidad discriminante adicional al modelo predictivo, facilitando la diferenciación entre unidades con bajo riesgo y unidades con alta probabilidad de incumplimiento.

Finalmente, la definición operacional de las variables permitió establecer coherencia entre las etapas de análisis exploratorio, segmentación y modelado supervisado, garantizando que cada variable utilizada tuviera una interpretación clara, una función analítica definida y una relación consistente con el problema de investigación planteado.

Variables Predictoras y Variable Objetivo

Para el desarrollo del modelo predictivo de clasificación se definió una variable objetivo y un conjunto de variables predictoras seleccionadas a partir del análisis exploratorio de datos, la

revisión de correlaciones y la capacidad discriminante observada en las etapas preliminares del análisis. Esta selección permitió estructurar el problema como una tarea de aprendizaje supervisado orientada a estimar el riesgo de mora crítica en las unidades residenciales.

Variable Objetivo

La variable objetivo utilizada en el proyecto fue target_riesgo_mora, construida como un indicador binario orientado a identificar unidades residenciales con riesgo financiero elevado.

La variable fue definida operacionalmente de la siguiente manera:

Tabla 5

Descripción de la Variable Objetivo

Valor	Interpretación
0	Unidad sin riesgo crítico de mora
1	Unidad con riesgo alto de mora

Nota. La variable objetivo fue construida integrando los estados jurídicos, pre jurídico y saldo activo en mora superior a 90 días.

La clasificación de riesgo alto se asignó a las unidades que cumplían al menos una de las siguientes condiciones:

Presentar saldo en mora superior a 90 días.

Encontrarse en estado jurídico.

Encontrarse en estado prejurídico.

Esta definición permitió representar de manera práctica y operativa los casos con mayor criticidad financiera dentro de la cartera del conjunto residencial. Asimismo, facilitó el

entrenamiento de modelos supervisados capaces de identificar patrones asociados al deterioro del comportamiento de pago.

Durante el análisis se identificó un desbalance entre las categorías de la variable objetivo, debido a que las unidades clasificadas como riesgo alto representaban una proporción considerablemente menor frente a las unidades sin riesgo crítico. Esta condición motivó el uso de métricas como Recall, F1-score y ROC-AUC, además de estrategias de ajuste de pesos de clase para fortalecer la capacidad predictiva de los modelos frente a la clase minoritaria.

Variables Predictoras

Las variables predictoras corresponden al conjunto de características utilizadas como entrada para el entrenamiento de los modelos de clasificación. Estas variables fueron seleccionadas considerando su relevancia analítica, su relación con el fenómeno de mora y su capacidad para diferenciar unidades con distintos niveles de riesgo financiero.

Inicialmente, el dataset consolidado contaba con 35 variables disponibles; sin embargo, para la etapa de modelado supervisado se seleccionaron 11 variables predictoras con mayor capacidad discriminante y menor redundancia estadística. Durante este proceso se evaluaron relaciones de colinealidad entre variables, evitando incorporar variables altamente correlacionadas que pudieran afectar la estabilidad e interpretabilidad del modelo.

La Tabla 6 presenta las variables predictoras utilizadas en el modelo final:

Tabla 6

Descripción de las Variables Predictoras

Variable	Descripción operacional	Tipo
cuotas_adm	Saldo pendiente por cuotas de administración	Continua
interes_mora	Intereses acumulados por mora	Continua

Variable	Descripción operacional	Tipo
promedio	Antigüedad promedio de mora en meses	Discreta
dias_sin_pago	Días transcurridos desde el último pago	Continua
_1_30_dias	Saldo en mora temprana	Continua
_31_60_dias	Saldo en mora intermedia	Continua
_61_90_dias	Saldo en mora moderada	Continua
mas_90_dias	Saldo en mora crítica superior a 90 días	Continua
ratio_mora_90	Proporción de deuda en mora crítica	Continua
num_tramos_mora	Número de tramos de mora activos	Discreta
prop_interes	Proporción de intereses sobre deuda total	Continua

Nota. Se excluyeron tiene_mora_90 (componente binario directo de la variable objetivo) y tiene_mora (proxy redundante) para prevenir fuga de información.

La variable binaria tiene_mora_90 fue excluida del conjunto final de predictores por ser componente directo de la definición de target_riesgo_mora, lo que generaría fuga de información (data leakage). De igual manera, tiene_mora fue excluida por constituir un proxy redundante del mismo fenómeno. Las variables continuas mas_90_dias y ratio_mora_90 se conservaron por aportar información cuantitativa sobre la magnitud y proporción de la deuda crítica, aunque su proximidad conceptual con el target fue documentada y considerada en la interpretación de los resultados.

Finalmente, la definición de una variable objetivo clara y la selección controlada de variables predictoras permitieron garantizar coherencia metodológica entre el análisis

exploratorio, la ingeniería de variables y el modelado supervisado, fortaleciendo la capacidad interpretativa y predictiva de los resultados obtenidos.

Herramientas Tecnológicas Utilizadas

Para el desarrollo del presente proyecto se utilizaron diferentes herramientas tecnológicas orientadas al procesamiento, análisis y modelado de datos, permitiendo implementar de manera integral las etapas del enfoque CRISP-DM.

Python

Python fue utilizado como lenguaje principal de programación para el análisis exploratorio de datos, ingeniería de características, segmentación de cartera y construcción de modelos predictivos. Su flexibilidad y amplio ecosistema de librerías permitieron desarrollar procesos de limpieza, transformación y análisis estadístico de manera eficiente.

Google Colab

Google Colab se empleó como entorno de ejecución en la nube para el desarrollo del proyecto. Esta plataforma facilitó la ejecución de notebooks interactivos, el procesamiento de datos y la visualización de resultados sin necesidad de infraestructura local especializada.

Jupyter Notebook

El proyecto fue estructurado mediante notebooks de Jupyter, permitiendo organizar de forma secuencial las etapas de carga de datos, análisis exploratorio, modelado y evaluación de resultados. Esta herramienta favorece la trazabilidad y reproducibilidad del proceso analítico.

Librerías Utilizadas

Para el desarrollo de los modelos y visualizaciones se emplearon diferentes librerías especializadas del ecosistema Python:

Pandas: manipulación y análisis de datos tabulares.

NumPy: operaciones matemáticas y manejo de arreglos numéricos.

Matplotlib: generación de gráficos estadísticos y visualizaciones analíticas.

Seaborn: visualización avanzada de datos y análisis exploratorio.

Scikit-learn: implementación de algoritmos de Machine Learning, validación cruzada y métricas de evaluación.

XGBoost: construcción de modelos de clasificación basados en gradient boosting.

SciPy: apoyo en operaciones estadísticas y análisis numérico.

Aprendizaje Automático y Modelado Predictivo

Los modelos predictivos y de segmentación fueron desarrollados mediante técnicas de aprendizaje supervisado y no supervisado, utilizando algoritmos como Regresión Logística, Random Forest, XGBoost y K-Means, orientados a la identificación de patrones de riesgo de mora en las unidades residenciales.

Visualización de Datos

Las visualizaciones analíticas fueron desarrolladas mediante gráficos estadísticos y comparativos que permitieron interpretar el comportamiento de la cartera, validar los modelos y comunicar los resultados obtenidos durante el proceso de investigación.

Resultados

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos corresponde a la Fase 2 del modelo CRISP-DM y tuvo como propósito examinar la estructura, calidad, distribución y comportamiento de las fuentes de información disponibles antes de avanzar hacia la etapa de segmentación y modelado predictivo. Este análisis permitió identificar patrones relevantes de mora, inconsistencias, valores atípicos y variables con potencial predictivo para la clasificación del riesgo de cartera.

Integración y Calidad de los Datos

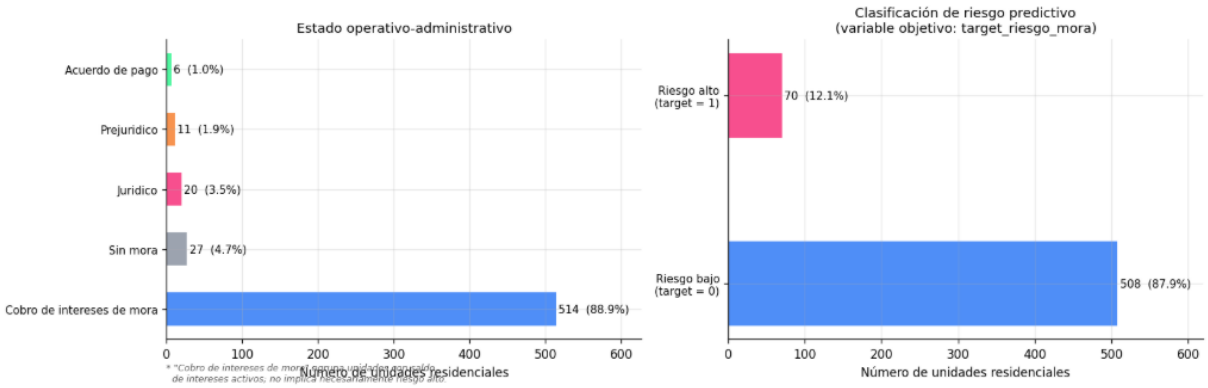
Las fuentes CART_MORA_apto_apto.xlsx y Cart_Mora_Tiempo.xlsx fueron integradas mediante el código interno de cada unidad residencial. La primera fuente contiene 578 registros y 21 variables relacionadas con la composición de la deuda, estado de cobro y comportamiento de pago; la segunda aporta información por tramos de antigüedad de mora. Como resultado, se consolidó un dataset de 578 unidades con 35 variables disponibles para el análisis.

Durante la revisión de calidad se identificaron 27 valores nulos en la variable juridico, equivalentes al 4.7%, los cuales fueron imputados como “Sin mora” al corresponder a unidades con saldo igual o inferior a cero. También se encontraron 102 registros con valores negativos en la cartera total, asociados a anticipos que superan el saldo adeudado, por lo que fueron tratados como unidades sin mora activa. Adicionalmente, 321 unidades sin registros en la fuente de antigüedad recibieron valor cero en todos los tramos, dado que no presentaban deuda en mora.

Patrones de Mora y Distribución de la Deuda

Figura 1

Distribución de las Unidades según Estado de Cartera



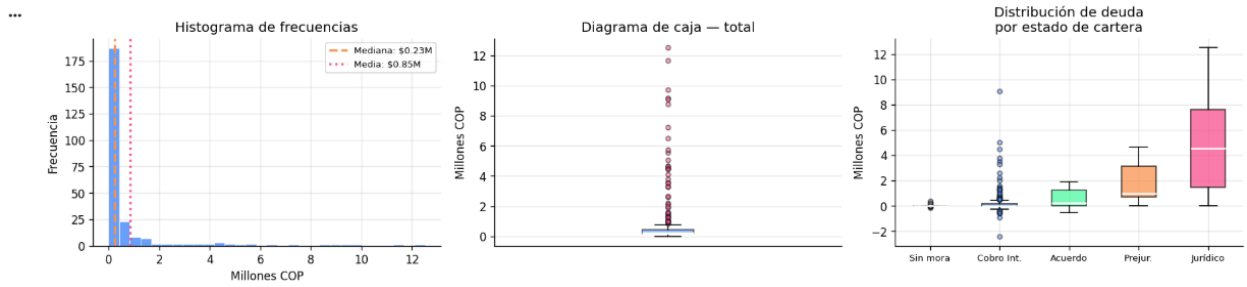
Nota. El panel izquierdo presenta la clasificación operativo-administrativa de las unidades residenciales y el panel derecho muestra la distribución de la variable objetivo target_riesgo_mora utilizada en el modelado predictivo.

Se evidencia una alta concentración de unidades en estado de cobro de intereses de mora, correspondiente a unidades con saldo de intereses activo, aunque no necesariamente asociadas a situaciones de riesgo crítico. Por su parte, los estados Jurídico y Prejurídico representan una proporción reducida del total de unidades residenciales (5,4 %), pero concentran los casos con mayor nivel de criticidad financiera y antigüedad de deuda. El panel derecho muestra la variable objetivo construida para el modelado predictivo (target_riesgo_mora), la cual integra las condiciones asociadas al riesgo alto y evidencia un desbalance de clases, con una proporción de 11,2 % de casos positivos, aspecto que fue considerado durante el entrenamiento y evaluación de los modelos supervisados.

Distribución estadística de la deuda en unidades con mora activa

Figura 2

Distribución de la Deuda Total en Unidades con Mora Activa



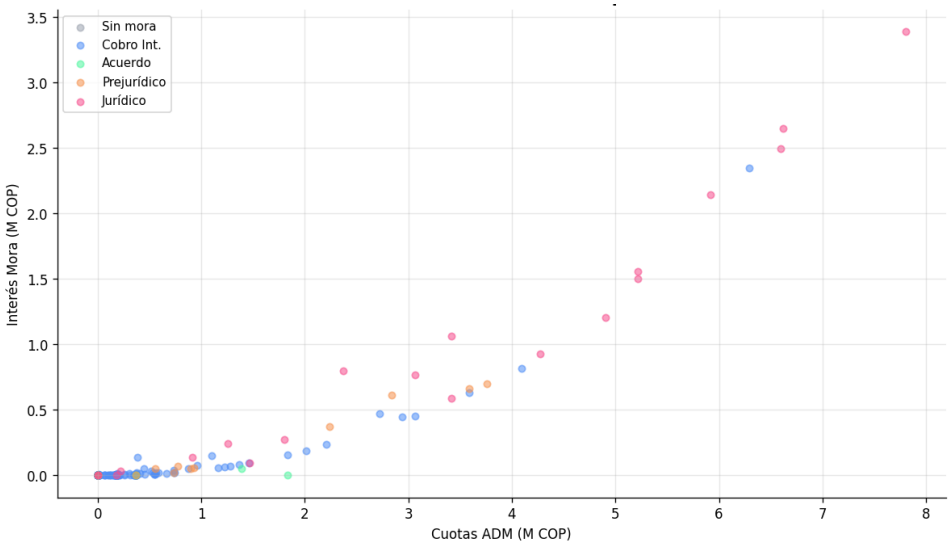
Nota. El panel izquierdo presenta el histograma de frecuencias con indicadores de media y mediana; el panel central muestra el diagrama de caja general; y el panel derecho compara la distribución de la deuda según el estado de cartera.

La Figura 2 muestra la distribución de la deuda total de las unidades con mora activa. Los resultados evidencian una marcada asimetría positiva, caracterizada por una alta concentración de unidades con valores de deuda relativamente bajos y un número reducido de casos con montos significativamente superiores. El diagrama de caja confirma la presencia de valores atípicos y una elevada dispersión en la parte superior de la distribución. Adicionalmente, la comparación por estado de cartera evidencia que las categorías Jurídico y Prejurídico concentran los mayores niveles de deuda y la mayor variabilidad, mientras que los estados Sin mora y Cobro de intereses de mora presentan montos considerablemente menores. Estos hallazgos sugieren que la severidad de la deuda se incrementa progresivamente conforme avanza el proceso de gestión de cartera.

Relación entre cuotas de administración e intereses de mora

Figura 3

Relación entre Cuotas de Administración e Intereses de Mora por Estado de Cartera



Nota. Se observa una tendencia creciente entre ambas variables; las unidades en estado Jurídico y Prejurídico concentran los niveles más altos de deuda e intereses, confirmando el mayor deterioro financiero de estos casos.

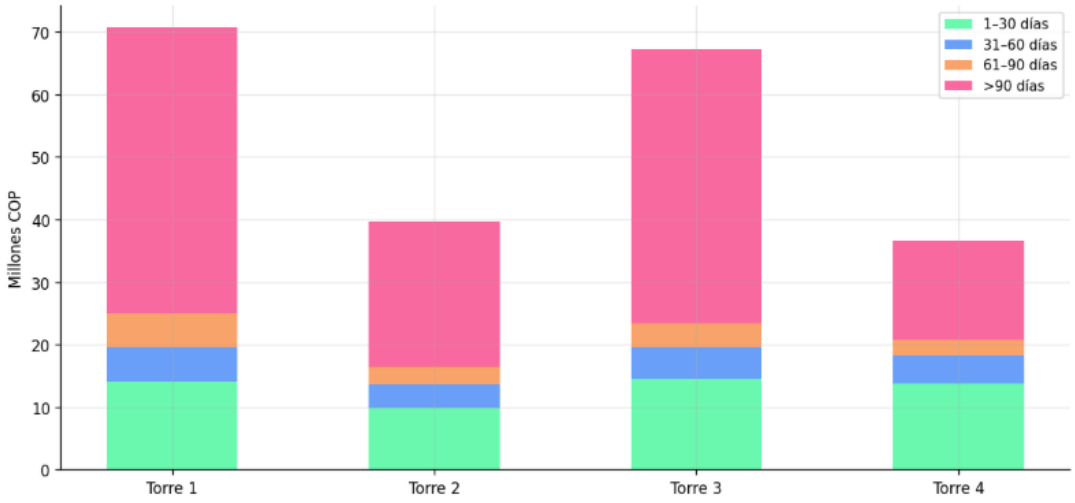
La Figura 3 presenta la relación entre el valor adeudado por cuotas de administración y los intereses de mora acumulados, segmentados según el estado de cartera de cada unidad residencial. Se observa una tendencia creciente entre ambas variables, evidenciando que el incremento de las cuotas pendientes se asocia directamente con mayores valores de interés de mora.

Asimismo, las unidades clasificadas en estado Jurídico y Prejurídico concentran los niveles más altos de deuda e intereses, reflejando un mayor deterioro financiero y antigüedad de cartera. En contraste, las unidades en estados operativos presentan valores significativamente menores y una menor dispersión. Este comportamiento respalda la relevancia de las variables financieras utilizadas durante el proceso de modelado predictivo.

El estado Cobro de intereses de mora agrupa las unidades que registran saldo de intereses activo según el sistema administrativo, categoría que incluye tanto unidades con mora reciente como condiciones de facturación ordinaria; por ello, la proporción de unidades con mora activa confirmada (43.8%) difiere del peso relativo de dicho estado en la distribución administrativa presentada en la Figura 1.

Figura 4

Distribución de la Cartera Acumulada por Torre y Tramo de Antigüedad de Mora



Nota. Las torres 1 y 3 concentran el mayor volumen de deuda; el tramo superior a 90 días representa la mayor proporción de cartera en todas las torres.

La Figura 4 presenta la distribución de la cartera acumulada por torre y tramo de antigüedad de mora. Se observa que las torres 1 y 3 concentran el mayor volumen de deuda dentro del conjunto residencial. Asimismo, el tramo superior a 90 días representa la mayor proporción de la cartera en todas las torres, lo que evidencia una persistencia significativa de obligaciones vencidas de difícil recuperación. Este comportamiento confirma que la antigüedad de la deuda constituye uno de los principales factores asociados al riesgo financiero y respalda la

importancia predictiva de las variables relacionadas con mora prolongada dentro del modelo de clasificación desarrollado.

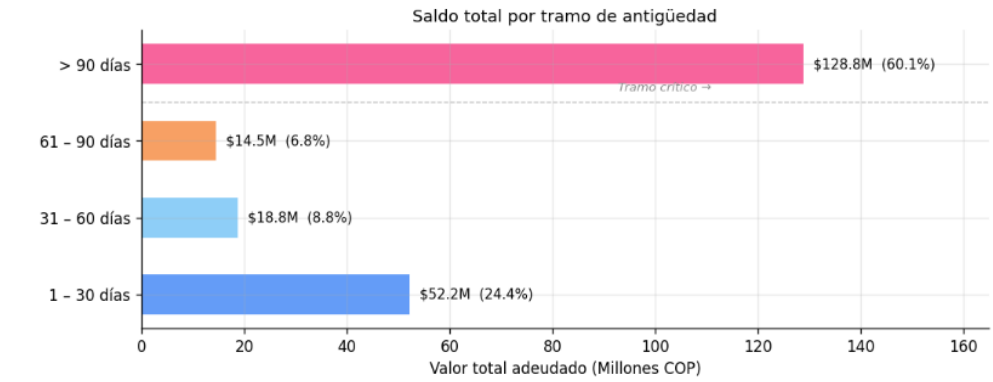
Los estados Jurídico y Prejurídico representan el 5.4% del total de unidades, con 20 y 11 casos, respectivamente. Aunque corresponden a una minoría, concentran una proporción significativa del valor adeudado debido a la acumulación de intereses de mora, honorarios de abogados y costas judiciales. Este comportamiento evidencia que la gestión de cartera no debe basarse únicamente en el número de unidades morosas, sino también en la criticidad y antigüedad de la deuda.

Para efectos del modelado predictivo, se construyó la variable objetivo target_riesgo_mora, que clasifica como riesgo alto (target = 1) las unidades en estado Jurídico, Prejurídico o con saldo activo en el tramo de antigüedad superior a 90 días. Bajo esta definición, el 11.2% de las unidades presenta riesgo alto, configurando un marcado desbalance de clases que fue considerado en el diseño y evaluación de los modelos predictivos.

Antigüedad de Mora

Figura 5

Distribución de la Cartera por Tramo de Antigüedad de Mora



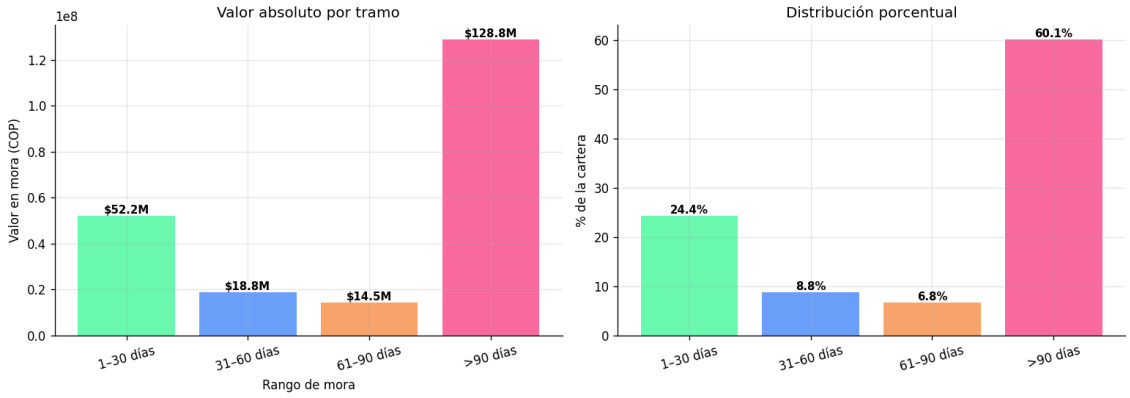
Nota. El panel izquierdo muestra el valor absoluto acumulado por tramo (1–30, 31–60, 61–90 y más de 90 días); el panel derecho presenta las unidades activas por tramo de antigüedad.

El panel izquierdo muestra el valor absoluto acumulado por tramo (1–30, 31–60, 61–90 y más de 90 días); el panel derecho presenta las unidades por tramo de antigüedad. El tramo superior a 90 días concentra más del 60% del valor total adeudado, confirmando el perfil de cartera deteriorada con presencia de deuda crónica.

El análisis por tramos permitió identificar que la mora superior a 90 días concentra más del 60% del valor total adeudado, lo que evidencia un perfil de cartera deteriorada y con presencia de deuda crónica. Además, se identificaron casos con antigüedad de mora de hasta 42 meses, aproximadamente 3.5 años.

Figura 6

Distribución de la Cartera por Tramo de Antigüedad: Valor Acumulado y Unidades



Nota. El panel izquierdo presenta el monto total adeudado en millones de pesos colombianos (M COP) por tramo de antigüedad; el panel derecho muestra la cantidad de unidades con saldo activo en cada tramo.

La concentración del valor adeudado en el tramo superior a 90 días, que supera el 60% del total, confirma que el deterioro financiero de la cartera se manifiesta principalmente a través de obligaciones crónicas de larga data, más que de mora reciente generalizada. Este patrón es

consistente con el comportamiento observado en carteras de propiedad horizontal donde la acumulación de intereses moratorios y la inacción administrativa temprana agravan progresivamente el nivel de endeudamiento (Thomas et al., 2002).

La Figura 6 evidencia que el tramo de mora superior a 90 días concentra el mayor volumen de cartera vencida en términos absolutos, mientras que los tramos de mora temprana (1–30 días) y moderada (31–60 y 61–90 días) representan proporciones significativamente menores del total adeudado. Esta distribución asimétrica respalda la pertinencia de las variables mas_90_dias y ratio_mora_90 como predictores de alta relevancia en el modelo de clasificación, dado que capturan el componente de cartera con mayor impacto financiero y mayor dificultad de recuperación.

Por distribución territorial interna, la Torre 3 concentra el mayor volumen de cartera acumulada, seguida por las Torres 1, 4 y 2. Asimismo, las unidades en estado jurídico presentan una mediana de 17 meses de mora, mientras que las unidades en cobro de intereses registran antigüedades considerablemente menores. Esto confirma que la antigüedad de la deuda es una variable clave para diferenciar niveles de riesgo.

Ingeniería de Variables Analíticas

A partir del dataset integrado se construyeron variables derivadas orientadas a representar mejor el comportamiento financiero de cada unidad residencial. Entre ellas se encuentran dias_sin_pago, ratio_mora_90, num_tramos_mora, prop_interes, tiene_mora, tiene_mora_90 y target_riesgo_mora.

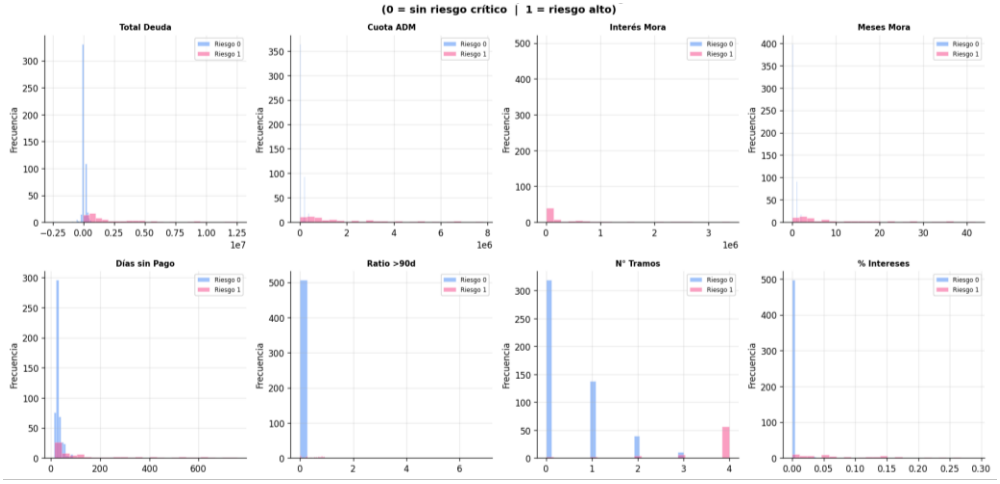
La variable objetivo target_riesgo_mora fue definida con valor 1 cuando la unidad presenta mora superior a 90 días o se encuentra en estado jurídico o prejurídico. Esta definición

permitió estructurar el problema como una tarea de clasificación binaria orientada a identificar unidades con riesgo alto de mora.

Variables Predictoras Identificadas

Figura 7

Distribución de Variables por Clase de Riesgo

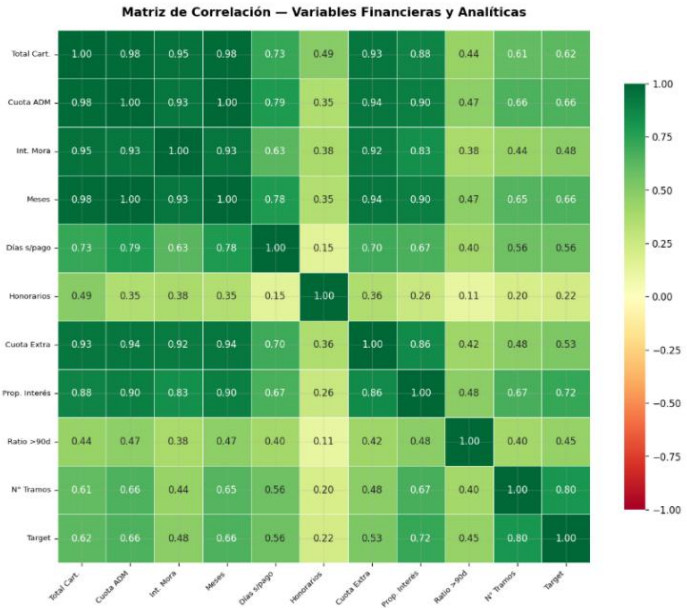


Nota. La figura presenta diagramas de caja comparativos diferenciando unidades sin riesgo crítico (clase 0) y unidades de riesgo alto (clase 1).

La figura presenta diagramas de caja comparativos para las principales variables predictoras, diferenciando entre las unidades sin riesgo crítico (clase 0) y las unidades de riesgo alto (clase 1). Las variables promedio, dias_sin_pago, interes_mora, ratio_mora_90 y num_tramos_mora evidencian mayor diferenciación entre clases, confirmando su pertinencia como predictores del modelo de clasificación.

Figura 8

Matriz de Correlación — Variables Financieras y Analíticas



Nota. La matriz de correlación fue calculada sobre las variables candidatas y la variable objetivo target_riesgo_mora.

La Figura 8 presenta la distribución comparativa de las variables predictoras por clase de riesgo, permitiendo identificar visualmente cuáles variables ofrecen mayor separación entre unidades de bajo y alto riesgo. Esta visualización respalda la selección de las variables con mayor capacidad discriminante para el modelado supervisado.

El análisis de correlación permitió evaluar la relación entre las principales variables financieras y analíticas construidas durante la etapa de preparación de datos. La Figura 8 presenta la matriz de correlación entre las variables candidatas y la variable objetivo target_riesgo_mora.

Como se observa en la Figura 8, se identificó una relación muy alta entre total_cartera, cuotas_adm y promedio, lo que evidencia redundancia parcial entre variables asociadas al valor

acumulado de la deuda. Asimismo, num_tramos_mora presentó la mayor correlación con la variable objetivo, con un valor aproximado de r = 0.80, seguida por prop_interes con r = 0.72, cuotas_adm con r = 0.66 y dias_sin_pago con r = 0.56. Estos resultados permitieron orientar la selección de variables predictoras y confirmar la relevancia de la antigüedad, composición y acumulación de la deuda en la clasificación del riesgo de mora.

Las variables promedio, dias_sin_pago, interes_mora, ratio_mora_90 y num_tramos_mora mostraron una alta capacidad de diferenciación entre unidades de bajo y alto riesgo. Adicionalmente, los valores atípicos identificados en deuda total, intereses de mora y días sin pago no fueron eliminados, ya que corresponden precisamente a los casos de mayor criticidad que el modelo busca detectar.

Finalmente, se identificó un desbalance de clases en la variable objetivo, con una proporción aproximada de 7.3:1 entre unidades sin riesgo crítico y unidades de riesgo alto. Por esta razón, para la etapa de modelado se consideró necesario utilizar métricas como F1-score, Recall y ROC-AUC, además de estrategias de ajuste por pesos de clase.

Modelado Predictivo

El presente capítulo desarrolla los Objetivos Específicos 3, 4 y 5 del proyecto, correspondientes a las Fases 4 (Modelado) y 5 (Evaluación) del modelo CRISP-DM. A partir del conjunto de datos integrado y preparado en el capítulo de Análisis Exploratorio de Datos, se implementan dos componentes analíticos complementarios: la segmentación no supervisada de la cartera mediante el algoritmo K-Means y la construcción y evaluación de modelos supervisados de clasificación orientados a estimar el riesgo de mora por unidad residencial. Los resultados obtenidos constituyen la base analítica para la formulación de recomendaciones dirigidas a fortalecer la gestión de recaudo del conjunto residencial.

Preparación del Dataset Para Modelado

Antes de desarrollar los componentes de segmentación y modelado, se verificó la integridad del dataset consolidado en la etapa anterior. El conjunto de trabajo quedó conformado por 578 registros correspondientes a las unidades residenciales de las cuatro torres del conjunto, con un total de 35 variables disponibles tras el proceso de integración e ingeniería de características.

Para el modelado supervisado se seleccionaron 11 variables predictoras con base en su poder discriminante identificado durante el análisis exploratorio y su pertinencia metodológica para reducir colinealidad y prevenir posibles fugas de información. La variable de mayor correlación con el total de cartera fue la cuota de administración ($r \approx 0.96$), por lo que se conservó esta última como representante del componente principal de deuda y se excluyó el total consolidado para evitar redundancia informativa.

Tabla 7

Variables Predictoras Seleccionadas para Modelado

Variable	Descripción	Tipo
cuotas_adm	Saldo pendiente por cuotas de administración	Continua
interes_mora	Intereses acumulados por mora	Continua
promedio	Antigüedad de mora en meses	Discreta
dias_sin_pago	Días desde el último pago registrado	Continua
_1_30_dias	Saldo en tramo de mora temprana (1–30 días)	Continua
_31_60_dias	Saldo en tramo de mora intermedia (31–60 días)	Continua
_61_90_dias	Saldo en tramo de mora moderada (61–90 días)	Continua
mas_90_dias	Saldo en tramo de mora crítica (>90 días)	Continua
ratio_mora_90	Proporción de deuda en tramo >90 días	Continua
num_tramos_mora	Número de tramos con saldo activo	Discreta
prop_interes	Proporción de intereses sobre deuda total	Continua

Nota. Las variables fueron seleccionadas con base en su poder discriminante identificado en el análisis exploratorio y en criterios de reducción de colinealidad.

Durante el proceso de selección de variables se realizó una revisión metodológica orientada a prevenir problemas de fuga de información (data leakage). Debido a que la variable objetivo target_riesgo_mora fue construida utilizando criterios relacionados con mora superior a 90 días y estados críticos de cartera, variables binarias derivadas directamente de esta condición, como tiene_mora_90, fueron excluidas del entrenamiento de los modelos supervisados.

Asimismo, se identificó que variables continuas como mas_90_dias y ratio_mora_90 presentan una relación estrecha con el tramo crítico de mora. No obstante, estas variables se conservaron debido a que aportan información cuantitativa sobre la magnitud y proporción de la deuda acumulada, permitiendo representar distintos niveles de severidad financiera sin reproducir explícitamente la etiqueta binaria del target. Sus resultados fueron interpretados bajo criterios de cautela metodológica durante la evaluación de los modelos.

La variable objetivo target_riesgo_mora fue construida como indicador binario que toma el valor de 1 cuando la unidad presenta mora activa en el tramo superior a 90 días o se encuentra en estado jurídico o prejurídico, y 0 en caso contrario. Su distribución refleja un desbalance de clases de aproximadamente 8:1 entre la categoría sin riesgo crítico (clase 0) y la categoría de riesgo alto (clase 1), condición que motivó el uso de ajuste por pesos de clase y métricas orientadas a la detección de la clase minoritaria.

Segmentación de Cartera Mediante K-Means

Fundamento Metodológico

La segmentación de cartera se desarrolló mediante el algoritmo K-Means, técnica de aprendizaje no supervisado orientada a identificar grupos homogéneos de unidades con comportamientos financieros similares. Previamente, las variables fueron estandarizadas mediante StandardScaler, con el fin de evitar sesgos derivados de las diferencias de escala entre los datos. El proceso se realizó sobre 578 unidades residenciales y 11 variables numéricas relacionadas con deuda, antigüedad y comportamiento de mora.

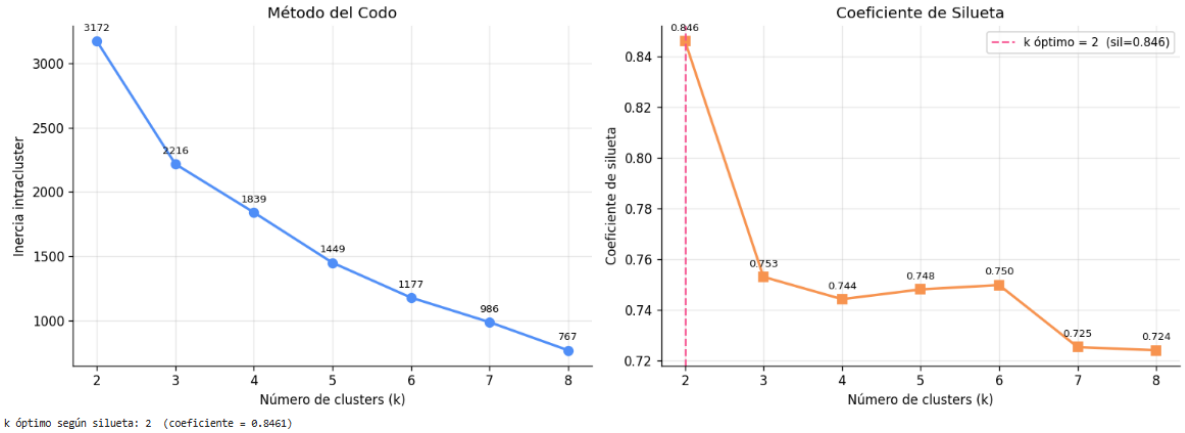
Determinación del Número Óptimo de Clusters

La selección del número adecuado de grupos se realizó mediante dos criterios complementarios. El método del codo (elbow method) evaluó la inercia intracluster en función

del número de grupos, identificando el punto de inflexión a partir del cual la reducción marginal de inercia disminuye notoriamente. El coeficiente de silueta (silhouette score) complementó este análisis midiendo la cohesión interna de cada cluster y su separación respecto a los demás grupos. El número óptimo se determinó como aquel que maximizó el coeficiente de silueta, confirmando la segmentación en k = 2 clusters como la configuración más adecuada para los datos disponibles.

Figura 9

Determinación del Número Óptimo de Clusters mediante K-Means



Nota. El método del codo evidenció disminución progresiva de la inercia intracluster; el coeficiente de silueta alcanzó su valor máximo en k = 2 (0,846), conforme a los estándares de Rousseeuw (1987) y Kaufman y Rousseeuw (2005).

La Figura 9 presenta los resultados obtenidos mediante el método del codo y el coeficiente de silueta para distintos valores de k. El método del codo evidenció una disminución progresiva de la inercia intracluster a medida que aumenta el número de grupos, mientras que el coeficiente de silueta alcanzó su valor máximo en k = 2 (0.846), indicando una alta cohesión interna y adecuada separación entre clusters. Estos resultados confirmaron que la partición en

dos segmentos representa la estructura más consistente para la cartera analizada. Conforme a los estándares de la literatura especializada, valores de silueta superiores a 0.70 se consideran excelentes (Rousseeuw, 1987; Kaufman & Rousseeuw, 2005), lo que otorga sólido respaldo estadístico a la solución obtenida. En carteras de propiedad horizontal con alta concentración de unidades al día y una minoría con mora severa, la estructura bipolar resultante no representa una simplificación sino una descripción fiel de la realidad financiera del conjunto, donde la transición entre cumplimiento y mora crítica tiende a ser abrupta una vez superado el umbral de los 90 días sin pago.

Resultados de la Segmentación

El modelo K-Means ajustado con k = 2 asignó las 578 unidades residenciales a dos segmentos con características marcadamente diferenciadas:

Tabla 8

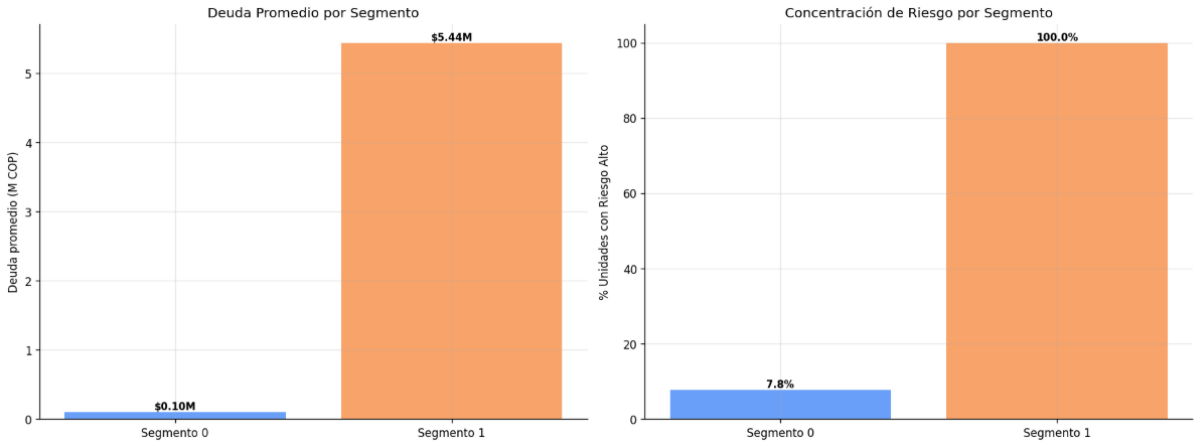
Caracterización de Segmentos Obtenidos mediante K-Means

Segmento	N Unidades	Etiqueta	Deuda Prom. (M COP)	Meses Mora	% Riesgo Alto
Segmento 0	551	Sin Riesgo Relevante	$0.10M	0.5	7.8%
Segmento 1	27	Riesgo Crítico	$5.44M	20.4	100.0%

Nota. Resultados obtenidos mediante el modelo K-Means ajustado con k = 2. M COP = millones de pesos colombianos.

Figura 10

Caracterización de Segmentos de Cartera mediante K-Means



Nota. Resultados obtenidos mediante el algoritmo K-Means con k = 2. El Segmento 0 agrupa la mayoría de las unidades con bajo riesgo, mientras que el Segmento 1 concentra los casos asociados a riesgo crítico de cartera.

La Figura 10 resume las principales diferencias identificadas entre los segmentos obtenidos mediante K-Means. El Segmento 0 concentra la mayoría de las unidades residenciales, con bajos niveles de deuda promedio y una reducida proporción de riesgo alto. En contraste, el Segmento 1 agrupa un número reducido de unidades con elevada deuda acumulada y concentración total de riesgo crítico, confirmando la capacidad del algoritmo para identificar perfiles financieros claramente diferenciados dentro de la cartera analizada.

Segmento 0 Sin Riesgo Relevante (551 unidades, 95.3%): Agrupa la gran mayoría de las unidades del conjunto, caracterizadas por una deuda promedio de $0.10M COP, una antigüedad de mora de apenas 0.5 meses y una concentración de riesgo alto del 7.8%. Este segmento incluye tanto unidades completamente al día como aquellas con saldos menores recientes, para las cuales la gestión preventiva de bajo costo resulta suficiente.

Segmento 1 Riesgo Crítico (27 unidades, 4.7%): Concentra la totalidad de las unidades en estado jurídico o prejurídico, con una deuda promedio de $5.44M COP, una antigüedad de mora de 20.4 meses y un porcentaje de riesgo alto del 100%. Estas 27 unidades representan el foco prioritario de la gestión de recaudo, dado que su nivel de deterioro requiere intervención formal especializada.

La nitidez de esta separación en dos grupos refleja la estructura bimodal de la cartera: una mayoría de unidades con comportamiento de pago regular y una minoría reducida pero de alto impacto financiero con mora crónica consolidada. El coeficiente de silueta obtenido confirmó la validez estadística de esta partición como la más coherente con los datos disponibles.

Modelado Predictivo Supervisado

Diseño Experimental

El modelado predictivo se planteó como una tarea de clasificación binaria supervisada, orientada a estimar el riesgo de mora crítica. Se evaluaron tres algoritmos: Regresión Logística, Random Forest y XGBoost. El dataset fue dividido en conjuntos de entrenamiento (80%, 462 registros) y prueba (20%, 116 registros) mediante partición estratificada, preservando la proporción de clases en ambos subconjuntos. Debido al desbalance identificado en la variable objetivo, se aplicaron mecanismos de ajuste como class_weight='balanced' y scale_pos_weight. Es importante considerar que el conjunto de prueba contiene únicamente 14 unidades de riesgo alto; en este escenario, cada error de clasificación individual altera el Recall en aproximadamente ±7 puntos porcentuales y el F1-score en ±4–5 puntos. Por esta razón, la validación cruzada estratificada de cinco pliegues se adoptó como estimador principal de la capacidad de generalización, siendo más robusta que las métricas del hold-out único.

Resultados por Modelo

Los tres modelos fueron entrenados y evaluados sobre el conjunto de prueba. A continuación, se presentan los resultados individuales para la clase de interés (Riesgo Alto):

Tabla 9

Resultados de Clasificación — Regresión Logística

Clase	Precisión	Recall	F1-score	Soporte
Sin riesgo crítico	0.961	0.971	0.966	102
Riesgo alto	0.769	0.714	0.741	14
Exactitud global	—	—	0.940	116
ROC-AUC	—	—	0.884	—

Nota. Resultados obtenidos sobre el conjunto de prueba (n = 116). F1-score y ROC-AUC calculados con class_weight='balanced'.

Tabla 10

Resultados de Clasificación — Random Forest

Clase	Precisión	Recall	F1-score	Soporte
Sin riesgo crítico	0.971	1.000	0.986	102
Riesgo alto	1.000	0.786	0.880	14
Exactitud global	—	—	0.974	116
ROC-AUC	—	—	0.901	—

Nota. Resultados obtenidos sobre el conjunto de prueba (n = 116). F1-score y ROC-AUC calculados con class_weight='balanced'.

Tabla 11

Resultados de Clasificación — XGBoost

Clase	Precisión	Recall	F1-score	Soporte
Sin riesgo crítico	0.971	1.000	0.986	102
Riesgo alto	1.000	0.786	0.880	14
Exactitud global	—	—	0.974	116

Nota. Resultados obtenidos sobre el conjunto de prueba (n = 116). F1-score y ROC-AUC calculados con class_weight='balanced'.

Comparación y Selección del Modelo Final

Los modelos evaluados mostraron desempeños diferenciados sobre el conjunto de prueba. Random Forest y XGBoost alcanzaron un F1-score de 0.880, un recall de 0.786 y una precisión de 1.000 para la clase de riesgo alto. La Regresión Logística, como modelo de referencia, obtuvo un F1-score de 0.741, un recall de 0.714 y una precisión de 0.769, evidenciando un desempeño inferior frente a los modelos basados en árboles. En cuanto al ROC-AUC, Random Forest obtuvo el valor más alto con 0.901, seguido de XGBoost con 0.890 y Regresión Logística con 0.884.

Tabla 12

Comparación Consolidada de Modelos — Clase: Riesgo Alto

Modelo	Precisión	Recall	F1-score	ROC-AUC
Regresión Logística	0.769	0.714	0.741	0.884
Random Forest	1.000	0.786	0.880	0.901

Modelo	Precisión	Recall	F1-score	ROC-AUC
XGBoost	1.000	0.786	0.880	0.890

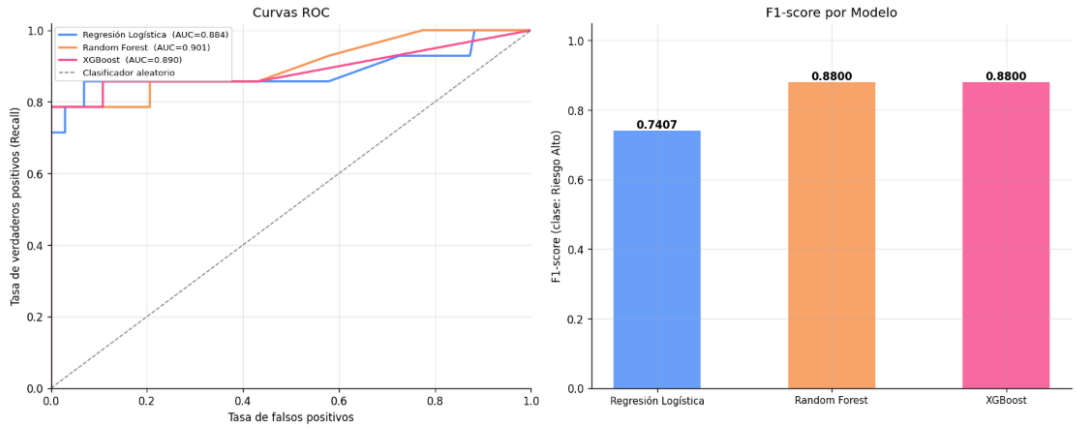
Nota. Resultados sobre el conjunto de prueba (n = 116, 14 positivos). La selección del modelo final se basó en el F1-score para la clase de riesgo alto y el ROC-AUC.

Random Forest fue seleccionado como modelo final por presentar el mayor F1-score (0.880) y el mayor ROC-AUC (0.901) entre los tres modelos evaluados. Su capacidad para capturar relaciones no lineales entre variables financieras, combinada con el análisis de importancia por reducción de impureza (Gini), permite identificar con claridad los factores más determinantes del riesgo de mora, conservando la capacidad interpretativa necesaria para generar recomendaciones operativas orientadas a la administración del conjunto residencial.

El Recall de 0.786 indica que el modelo identifica correctamente el 78.6% de las unidades de riesgo alto en el conjunto de prueba, sin generar falsos positivos (Precisión = 1.000). Dado el tamaño reducido de la clase minoritaria (14 unidades en el conjunto de prueba), este resultado es estadísticamente sólido y operativamente relevante para la gestión de recaudo.

Figura 11

Evaluación Comparativa de Modelos Predictivos

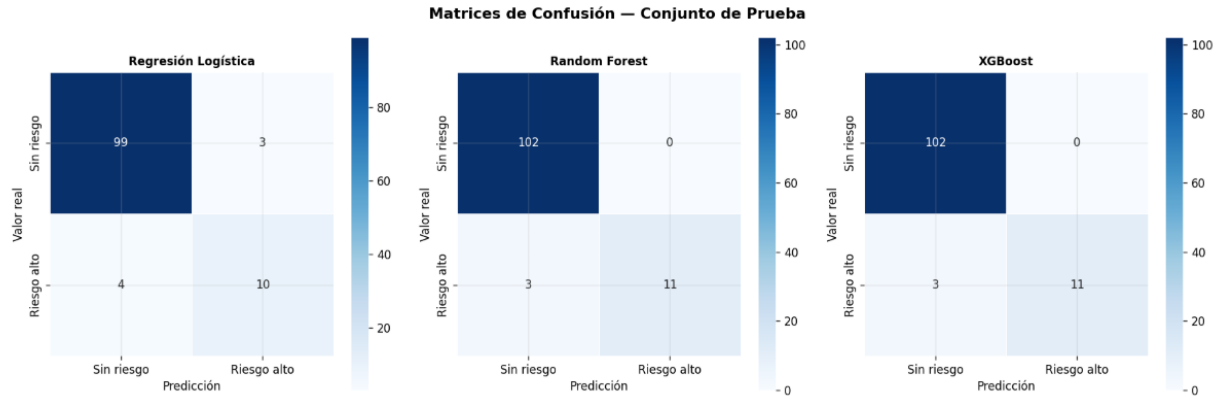


Nota. La figura presenta la comparación del desempeño de los modelos de Regresión Logística, Random Forest y XGBoost mediante las métricas ROC-AUC y F1-score calculadas sobre el conjunto de prueba (n = 116).

La Figura 11 resume visualmente el desempeño de los modelos supervisados evaluados mediante las curvas ROC y el F1-score para la clase de riesgo alto. Random Forest presentó la mayor capacidad discriminativa (ROC-AUC = 0.901), seguido de XGBoost (0.890) y Regresión Logística (0.884). De igual manera, los modelos basados en árboles alcanzaron el mejor equilibrio entre precisión y recall, reflejado en un F1-score de 0.880. Estos resultados respaldan la selección de Random Forest como modelo final para la estimación del riesgo de mora en el conjunto residencial.

Figura 12

Matrices de Confusión de los Modelos Supervisados



Nota. Los modelos Random Forest y XGBoost clasificaron correctamente la totalidad de las unidades sin riesgo crítico y detectaron 11 de las 14 unidades de riesgo alto, generando únicamente tres falsos negativos y ningún falso positivo.

La Figura 12 presenta las matrices de confusión obtenidas para los tres modelos supervisados evaluados. Los resultados muestran que Random Forest y XGBoost clasificaron correctamente la totalidad de las unidades sin riesgo crítico y detectaron 11 de las 14 unidades correspondientes a riesgo alto, generando únicamente tres falsos negativos y ningún falso positivo.

Desde una perspectiva operativa, la ausencia de falsos positivos en los modelos basados en árboles reduce el riesgo de intervenciones innecesarias sobre unidades financieramente estables, mientras que la adecuada detección de casos críticos fortalece la priorización de estrategias de recaudo preventivo y jurídico.

Validación Cruzada Estratificada

Con el propósito de estimar la capacidad de generalización del modelo seleccionado más allá de la partición entrenamiento/prueba, se aplicó validación cruzada estratificada con cinco pliegues (Stratified 5-Fold CV) sobre el dataset completo. Este procedimiento permite evaluar la estabilidad del modelo ante diferentes configuraciones del conjunto de datos.

Tabla 13

Resultados de Validación Cruzada Estratificada (5-Fold) — Random Forest

Pliegue	F1-score	ROC-AUC
Pliegue 1	1.0000	1.000
Pliegue 2	0.9231	0.9160
Pliegue 3	0.9231	0.9086
Pliegue 4	1.0000	1.0000
Pliegue 5	1.0000	1.0000
Media ± Desv. Est.	0.9692 ± 0.0377	0.9649 ± 0.0430

Nota. Validación cruzada estratificada con cinco pliegues aplicada sobre el conjunto de datos completo (n = 578).

Los resultados de la validación cruzada evidencian un desempeño consistente del modelo Random Forest sobre diferentes particiones del dataset. El F1-score promedio de 0.969 ± 0.038 y el ROC-AUC promedio de 0.965 ± 0.043 muestran una adecuada estabilidad en la capacidad predictiva del modelo y una baja variabilidad entre pliegues. No obstante, dado que algunas variables asociadas al tramo crítico de mora presentan una relación estrecha con la construcción

de la variable objetivo, estos resultados deben interpretarse considerando las limitaciones metodológicas inherentes al problema de clasificación planteado.

No obstante, dado que algunas variables asociadas al tramo crítico de mora presentan una relación estrecha con la construcción de la variable objetivo, estos resultados deben interpretarse considerando las limitaciones metodológicas inherentes al problema de clasificación planteado. Aun así, el comportamiento estable entre pliegues respalda la utilidad del modelo como herramienta analítica de apoyo para la gestión de recaudo.

Importancia de Variables

El análisis de importancia Gini realizado sobre el modelo Random Forest —reportado como referencia comparativa— permitió identificar las variables con mayor aporte en la estimación del riesgo de mora. La Tabla 8 presenta el ranking de importancia relativa de las variables predictoras.

Tabla 14

Importancia de Variables Mediante Criterio Gini: Random Forest

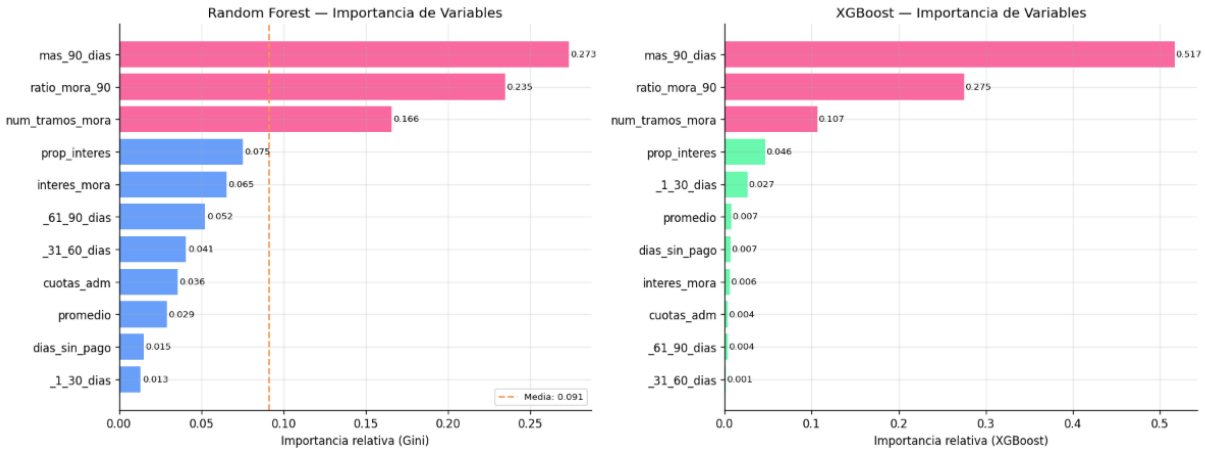
Ranking	Variable	Importancia RF (%)
1	mas_90_dias	27.35
2	ratio_mora_90	23.46
3	num_tramos_mora	16.58
4	prop_interes	7.52
5	interes_mora	6.53
6	_61_90_dias	5.24
7	_31_60_dias	4.07
8	cuotas_adm	3.57

Ranking	Variable	Importancia RF (%)
9	promedio	2.90
10	dias_sin_pago	1.49
11	_1_30_dias	1.31

Nota. Importancia calculada mediante la impureza de Gini en el modelo Random Forest. Los valores se reportan como referencia comparativa debido a que esta métrica puede sobreestimar variables con múltiples valores únicos.

Figura 13

Importancia Relativa de Variables en Random Forest y XGBoost



Nota. La coincidencia entre ambos modelos en las variables más relevantes evidencia consistencia en la identificación de los principales factores asociados al riesgo financiero. La alta relevancia de las variables del tramo mayor a 90 días debe interpretarse considerando su proximidad conceptual con la variable objetivo.

La Figura 13 compara la importancia relativa de las variables predictoras en los modelos Random Forest y XGBoost. En ambos algoritmos, las variables asociadas al tramo crítico de

mora (mas_90_dias y ratio_mora_90) presentaron los mayores niveles de importancia relativa, seguidas por variables relacionadas con persistencia de mora y distribución de deuda. Esta coincidencia entre modelos evidencia consistencia en la identificación de los principales factores asociados al riesgo financiero.

No obstante, la alta relevancia de variables directamente relacionadas con el tramo superior a 90 días debe interpretarse considerando su proximidad conceptual con la definición de la variable objetivo, aspecto que representa una posible fuente de fuga parcial de información y que fue considerado durante la evaluación metodológica del modelo.

El saldo correspondiente al tramo superior a 90 días (mas_90_dias) presentó la mayor relevancia (27.35%), seguido por la proporción de deuda en mora crítica (ratio_mora_90, 23.46%) y el número de tramos con mora activa (num_tramos_mora, 16.58%). La antigüedad promedio de mora (promedio) ocupó el noveno puesto con 2.90%, lo que indica que la magnitud y distribución de la deuda en tramos críticos tiene mayor capacidad discriminante que la medida temporal de la mora por sí sola.

Si bien la variable mas_90_dias presentó la mayor importancia relativa dentro del modelo, este resultado debe interpretarse considerando su estrecha relación conceptual con la definición de la variable objetivo target_riesgo_mora. Esta situación representa una limitación metodológica potencial asociada al riesgo de fuga parcial de información, aspecto que fue considerado durante la interpretación de los resultados y la evaluación del desempeño del modelo.

En conjunto, los resultados sugieren que el riesgo financiero está asociado principalmente con la persistencia de la deuda, su antigüedad y su distribución en diferentes tramos de mora, más que únicamente con el valor total adeudado. Asimismo, variables como prop_interes,

cuotas_adm y ratio_mora_90 aportan información complementaria para diferenciar unidades con comportamiento financiero crítico.

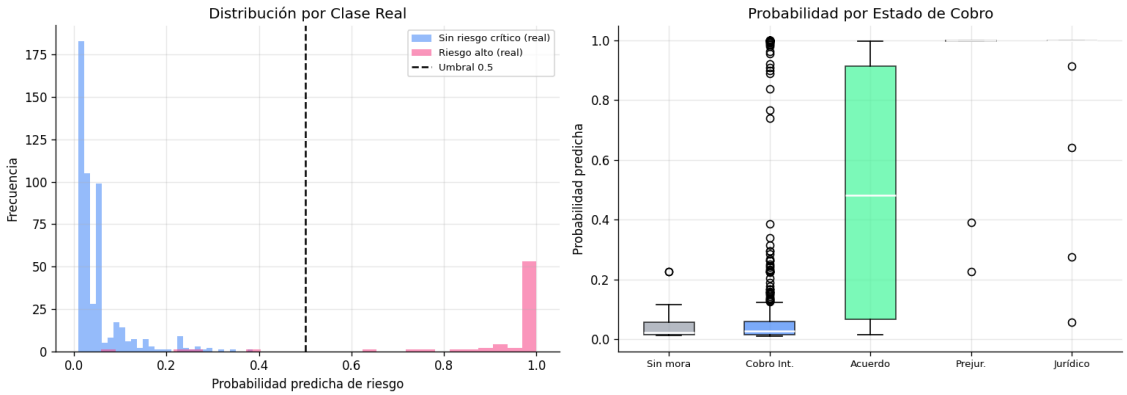
Estos hallazgos son coherentes con los patrones identificados durante el análisis exploratorio de datos y evidencian consistencia entre la etapa de ingeniería de características y el comportamiento observado en el modelado predictivo.

Recomendaciones Para la Gestión de Recaudo

Los resultados del modelado predictivo y la segmentación de cartera permiten formular recomendaciones orientadas a fortalecer la gestión de recaudo del conjunto residencial. Estas recomendaciones se derivan directamente de los hallazgos analíticos obtenidos y buscan apoyar la toma de decisiones basada en evidencia.

Figura 14

Distribución de Probabilidades Predichas de Riesgo de Mora Crítica



Nota. El análisis se realizó sobre el conjunto de datos completo (n = 578). El panel izquierdo muestra el histograma de probabilidades por clase real y el panel derecho presenta la distribución por estado de cobro.

La Figura 14 muestra la distribución de las probabilidades predichas de riesgo de mora crítica sobre el dataset completo. Se observa una adecuada separación entre las distribuciones de las dos clases, con la mayoría de las unidades sin riesgo crítico concentradas en valores de probabilidad bajos, mientras las unidades de riesgo alto presentan probabilidades notablemente superiores. Este resultado respalda la utilidad operativa del modelo para la priorización del recaudo.

Priorización por Segmento y Probabilidad Predicha

La combinación de la segmentación mediante K-Means y las probabilidades estimadas por el modelo predictivo permite identificar de manera más precisa las unidades con mayor criticidad financiera. En este sentido, se recomienda priorizar las acciones de cobro formal sobre el segmento clasificado como Riesgo Crítico, debido a su alta concentración de deuda y prolongada antigüedad de mora.

Intervención Preventiva en el Segmento 0

El modelo también permitió identificar unidades con señales tempranas de deterioro financiero dentro del segmento de bajo riesgo. Para estos casos, se recomienda implementar estrategias preventivas de recaudo, como recordatorios oportunos, acuerdos de pago y seguimiento temprano, con el propósito de evitar la evolución hacia estados jurídicos o prejurídicos.

Variables de Alerta Temprana

El análisis de importancia de variables evidenció que la acumulación de deuda en múltiples tramos de mora y el incremento de intereses asociados constituyen señales relevantes de deterioro de cartera. Por esta razón, se recomienda monitorear periódicamente estos indicadores como mecanismo de alerta temprana para apoyar la gestión preventiva del recaudo.

Actualización Periódica del Modelo

Debido a que el comportamiento financiero de los propietarios puede variar con el tiempo, se recomienda actualizar periódicamente el modelo predictivo incorporando nuevos cortes de cartera y reevaluando la relevancia de las variables utilizadas. Esta práctica permite mantener la estabilidad y pertinencia del sistema analítico como herramienta de apoyo a la toma de decisiones administrativas.

Síntesis del Capítulo

El desarrollo del capítulo de modelado predictivo permitió avanzar de manera estructurada en los tres objetivos analíticos centrales del proyecto. La segmentación mediante K-Means reveló una partición clara de la cartera en dos grupos con comportamientos financieros marcadamente diferenciados: un segmento mayoritario de bajo riesgo y un segmento reducido de riesgo crítico que concentra la totalidad de los casos jurídicos y prejurídicos. Esta clasificación no supervisada aporta una comprensión estructural de la cartera que complementa el análisis descriptivo desarrollado en el capítulo anterior.

Los modelos supervisados de clasificación demostraron capacidad predictiva sólida. Random Forest, seleccionado como modelo final, alcanzó un F1-score de 0.880 y un ROC-AUC de 0.901 sobre el conjunto de prueba. La validación cruzada estratificada confirmó la estabilidad del modelo con un F1-score promedio de 0.969 en cinco pliegues independientes, reforzando su capacidad de generalización. El análisis de importancia Gini identificó mas_90_dias (27.35%), ratio_mora_90 (23.46%) y num_tramos_mora (16.58%) como las variables con mayor capacidad predictiva. Adicionalmente, la convergencia entre el Segmento 1 de K-Means (100% riesgo alto) y la clase de riesgo alto del modelo supervisado valida la coherencia interna del análisis: dos

técnicas independientes identifican exactamente el mismo subconjunto de unidades como prioritarias para la gestión de recaudo.

En conjunto, los resultados demuestran la pertinencia de la ciencia de datos como herramienta de apoyo a la gestión administrativa en el contexto de la propiedad horizontal, transformando registros históricos de cartera en conocimiento útil para la toma de decisiones basada en evidencia cuantitativa.

Interpretación e Impacto Organizacional

Los resultados obtenidos mediante el análisis exploratorio, la segmentación de cartera y el modelado predictivo permiten evidenciar que la aplicación de técnicas de ciencia de datos puede aportar valor práctico a la gestión administrativa y financiera de la propiedad horizontal. Más allá del desempeño estadístico de los modelos desarrollados, el proyecto demuestra que la información histórica de cartera puede transformarse en herramientas de apoyo para la toma de decisiones orientadas al recaudo y la sostenibilidad financiera del conjunto residencial.

En primer lugar, la segmentación obtenida mediante K-Means facilita la identificación de grupos de unidades con comportamientos financieros diferenciados, permitiendo priorizar estrategias de recaudo de acuerdo con el nivel de criticidad de la cartera. Mientras el segmento de bajo riesgo puede gestionarse mediante acciones preventivas y seguimiento periódico, el segmento de riesgo crítico requiere procesos de intervención más focalizados, debido a la alta concentración de deuda y antigüedad de mora identificada en estas unidades.

De igual manera, el modelo predictivo desarrollado permite identificar señales tempranas de deterioro financiero antes de que las obligaciones evolucionen hacia estados jurídicos o prejurídicos. Variables como la mora superior a 90 días, la antigüedad promedio de la deuda y la acumulación de intereses demostraron ser factores relevantes para estimar el riesgo de

incumplimiento. Esto representa una oportunidad para fortalecer las estrategias preventivas de recaudo y reducir el escalamiento de casos hacia procesos jurídicos de mayor complejidad y costo administrativo.

Asimismo, la implementación de modelos analíticos puede contribuir a optimizar los procesos internos de la administración, disminuyendo la dependencia de revisiones manuales y permitiendo una gestión de cartera basada en evidencia cuantitativa. La priorización de unidades con mayor riesgo facilita una asignación más eficiente de los recursos administrativos y fortalece el seguimiento sobre los casos de mayor impacto financiero para el conjunto residencial.

Desde una perspectiva financiera, el proyecto también evidencia que la mora prolongada puede afectar la sostenibilidad presupuestal y operativa de la propiedad horizontal, debido a su impacto sobre el flujo de caja y la disponibilidad de recursos para mantenimiento y operación. En este contexto, el uso de herramientas predictivas puede apoyar la toma de decisiones relacionadas con planeación financiera, control de cartera y estrategias de recaudo preventivo.

Finalmente, los resultados obtenidos abren la posibilidad de integrar futuros desarrollos orientados a la automatización y monitoreo continuo de la cartera, mediante herramientas de inteligencia de negocios como Power BI o Microsoft Fabric. La incorporación de dashboards y sistemas de alerta temprana permitiría fortalecer la capacidad analítica de la administración y avanzar hacia una gestión de cartera más proactiva, objetiva y basada en datos.

Conclusiones

En primer lugar, el análisis exploratorio de datos permitió identificar que la cartera del conjunto residencial presenta una alta concentración del valor adeudado en un grupo reducido de unidades con mora crítica, especialmente en obligaciones superiores a 90 días. Aunque únicamente el 43.8% de las unidades registra mora activa, la mora superior a 90 días concentra más del 60% del valor total adeudado, evidenciando que el principal impacto financiero y jurídico de la cartera se encuentra asociado a casos de deuda crónica y prolongada. Asimismo, se identificaron unidades con antigüedades de mora de hasta 42 meses, lo que confirma que la severidad de la cartera depende más de la persistencia de la deuda que de la cantidad total de unidades morosas.

Adicionalmente, la aplicación de técnicas de segmentación mediante K-Means permitió identificar dos grupos claramente diferenciados de comportamiento financiero. El Segmento 0, conformado por 551 unidades (95.3%), agrupa propietarios con bajo nivel de riesgo, una deuda promedio cercana a $0.10 millones COP y una antigüedad promedio de mora de 0.5 meses. Por su parte, el Segmento 1, compuesto por 27 unidades (4.7%), concentra el 100% de los casos clasificados como riesgo crítico, con una deuda promedio aproximada de $5.44 millones COP y una antigüedad promedio de mora de 20.4 meses. Estos resultados evidencian que la segmentación constituye una herramienta útil para priorizar estrategias de recaudo y orientar acciones preventivas y correctivas de manera más eficiente.

De igual manera, los modelos supervisados evaluados demostraron que es posible clasificar el riesgo de mora utilizando información histórica de cartera disponible en la administración del conjunto residencial. Entre los algoritmos analizados, Random Forest presentó el mejor desempeño predictivo, alcanzando un F1-score de 0.880 y un ROC-AUC de

0.901 sobre el conjunto de prueba, además de un F1-score promedio de 0.969 en la validación cruzada estratificada. Los resultados evidenciaron que variables como mas_90_dias, ratio_mora_90 y num_tramos_mora fueron las de mayor capacidad predictiva, confirmando que la magnitud y distribución de la deuda en tramos críticos representan factores más determinantes que el valor absoluto adeudado para anticipar comportamientos de incumplimiento.

Por otra parte, el desarrollo del proyecto evidenció la utilidad de la ciencia de datos y el aprendizaje automático como herramientas de apoyo a la gestión administrativa en escenarios de propiedad horizontal. La integración de análisis exploratorio, segmentación y modelado predictivo permitió transformar registros operativos en información analítica útil para la toma de decisiones, facilitando procesos de priorización de cobro, seguimiento preventivo de cartera y focalización de esfuerzos administrativos sobre las unidades de mayor criticidad financiera.

Finalmente, se concluye que la metodología CRISP-DM proporcionó una estructura adecuada para el desarrollo del proyecto, permitiendo organizar de manera sistemática las etapas de comprensión del problema, preparación de datos, modelado y evaluación. Los resultados obtenidos demuestran que la aplicación de técnicas de ciencia de datos en propiedad horizontal puede fortalecer la gestión de cartera y apoyar la toma de decisiones administrativas basadas en evidencia, aportando herramientas analíticas para mejorar la sostenibilidad financiera y la eficiencia en los procesos de recaudo.

Recomendaciones

Se recomienda a la administración del conjunto residencial implementar estrategias de seguimiento preventivo enfocadas en las unidades que presenten incrementos progresivos en los indicadores de mora superior a 90 días, debido a que estas variables demostraron ser los principales factores asociados al riesgo financiero. La identificación temprana de estos casos permitiría intervenir oportunamente antes de que las obligaciones evolucionen hacia estados jurídicos o prejurídicos.

Asimismo, se recomienda utilizar los resultados de la segmentación obtenida mediante K-Means como herramienta de apoyo para la priorización de procesos de recaudo. La clasificación de unidades según niveles de riesgo puede facilitar la asignación eficiente de recursos administrativos y la definición de estrategias diferenciadas de cobranza, evitando aplicar las mismas acciones a todos los propietarios independientemente de su comportamiento financiero.

De igual manera, se considera pertinente avanzar hacia la integración de herramientas analíticas y tableros de control que permitan automatizar el monitoreo de la cartera. La incorporación de plataformas como Power BI o soluciones basadas en Microsoft Fabric podría fortalecer la visualización de indicadores críticos, facilitar el seguimiento en tiempo real y apoyar la toma de decisiones basada en datos dentro de la administración de propiedad horizontal.

Por otra parte, se recomienda que futuros análisis incorporen variables adicionales relacionadas con el comportamiento histórico de pago, reincidencia en mora y factores socioeconómicos, con el propósito de enriquecer la capacidad predictiva de los modelos y reducir posibles sesgos derivados de la información disponible actualmente.

Finalmente, se sugiere realizar evaluaciones periódicas del desempeño de los modelos predictivos implementados, debido a que el comportamiento financiero de los propietarios puede

variar con el tiempo. La actualización continua de los datos y la recalibración de los modelos permite mantener niveles adecuados de precisión y utilidad práctica en los procesos de gestión de cartera.

Limitaciones del Estudio

Alcance de un Único Caso de Estudio

El presente estudio se desarrolló a partir de información histórica de cartera correspondiente a un único conjunto residencial de propiedad horizontal ubicado en la ciudad de Bogotá, conformado por 578 unidades. Por esta razón, los resultados obtenidos deben interpretarse dentro de ese contexto específico y no pueden generalizarse directamente a otros conjuntos residenciales sin validación previa. Las diferencias en la estructura socioeconómica de los propietarios, el modelo de administración, la antigüedad del conjunto, la normativa interna aplicada y las dinámicas locales de pago pueden producir distribuciones de cartera marcadamente distintas, lo que afectaría tanto la relevancia de las variables predictoras identificadas como el desempeño de los modelos construidos. En la literatura de aprendizaje automático aplicado al riesgo financiero, la dependencia del contexto institucional es reconocida como una fuente primaria de sesgo de selección que limita la validez externa de los modelos (Thomas et al., 2002).

Tamaño Muestral Reducido y Variabilidad Estadística de las Métricas

El dataset analizado comprende 578 registros, de los cuales únicamente 70 corresponden a la clase de riesgo alto (12.1%). Con una partición de evaluación del 20%, el conjunto de prueba contiene 14 positivos, lo que implica que cada error de clasificación individual altera el Recall en aproximadamente ±7 puntos porcentuales y el F1-score en ±4–5 puntos. Esta alta variabilidad estadística inherente al tamaño muestral limita la precisión con la que pueden compararse los modelos entre sí y con resultados de la literatura especializada. Adicionalmente, el tamaño del dataset restringe la aplicación de técnicas de remuestreo más robustas como SMOTE o variantes de ensemble específicas para datos desbalanceados, cuya eficacia aumenta con conjuntos de

mayor tamaño (Géron, 2022). La validación cruzada estratificada de cinco pliegues mitiga parcialmente esta limitación al promediar el desempeño sobre múltiples particiones, pero no elimina la incertidumbre asociada a la clase minoritaria.

Corte Transversal Único: Ausencia de Dimensión Temporal

Los datos utilizados corresponden a un único corte de cartera con fecha de referencia del 31 de marzo de 2026, lo que implica que el análisis captura el estado financiero de las unidades en un momento específico pero no refleja la evolución temporal del comportamiento de pago. Esta naturaleza transversal del dataset impide distinguir entre unidades con mora reciente y deterioro progresivo de unidades con mora histórica crónica que eventualmente se estabilizó, así como identificar patrones estacionales o tendencias de largo plazo que pudieran afectar la predictibilidad del riesgo. Los modelos de clasificación construidos sobre cortes únicos tienen validez operativa para la priorización al momento del corte, pero su capacidad predictiva prospectiva —es decir, para anticipar qué unidades en buen estado financiero actual evolucionarán hacia mora crítica en períodos futuros— no puede evaluarse sin datos longitudinales (Hyndman & Athanasopoulos, 2021). Esta limitación representa la principal restricción para el uso del modelo como herramienta de alerta temprana.

Ausencia de Variables Socioeconómicas y Comportamentales Externas

El análisis se construyó exclusivamente con información financiera y administrativa proveniente de los registros de cartera del conjunto residencial, sin incorporar variables relacionadas con la capacidad de pago de los propietarios, su nivel socioeconómico, reincidencia histórica detallada, cambios en el núcleo familiar, eventos laborales o factores macroeconómicos como variaciones en tasas de inflación o empleo. En la literatura de scoring crediticio, la inclusión de variables socioeconómicas y comportamentales ha demostrado incrementar

significativamente la capacidad discriminante de los modelos predictivos (Thomas et al., 2002; López & Serán, 2021). La ausencia de estas variables no invalida el modelo desarrollado, que opera exclusivamente sobre información disponible en el sistema administrativo del conjunto, pero sí constituye una fuente de varianza no explicada que podría reducirse en versiones futuras del modelo con acceso a fuentes complementarias de información.

Desbalance de Clases y Sensibilidad de Métricas de Evaluación

La variable objetivo target_riesgo_mora presenta un desbalance de aproximadamente 7.3:1 entre la clase sin riesgo crítico (508 unidades) y la clase de riesgo alto (70 unidades). Este desequilibrio es inherente a la naturaleza del problema —la mora crítica afecta una minoría de la cartera— pero introduce retos metodológicos en la construcción y evaluación de modelos supervisados. Aunque se aplicaron estrategias de mitigación como el ajuste de pesos de clase (class_weight='balanced') y el uso de métricas orientadas a la clase minoritaria (F1-score, Recall, ROC-AUC), el desbalance puede influir en la estabilidad del umbral de clasificación óptimo, especialmente cuando el tamaño absoluto de la clase positiva es reducido. En conjuntos con mayor número de positivos, técnicas como SMOTE o variantes de sobremuestreo informado podrían ofrecer mejoras adicionales en la sensibilidad del modelo (Géron, 2022).

Proximidad Conceptual Entre Predictores y Variable Objetivo

Durante el desarrollo metodológico se identificaron dos niveles de riesgo de fuga de información que requirieron tratamiento diferenciado. En el primer nivel, la variable binaria tiene_mora_90 fue excluida del conjunto de predictores por ser componente directo de la definición de target_riesgo_mora, evitando así un data leakage explícito que habría invalidado las métricas del modelo. En el segundo nivel, las variables continuas mas_90_dias y ratio_mora_90 fueron conservadas por aportar información sobre la magnitud y proporción de la

deuda crítica; sin embargo, mantienen una proximidad conceptual con el target: cuando mas_90_dias > 0, la primera condición del target es verdadera por construcción. Esta condición, inherente a la definición operacional del target adoptada, explica en parte los altos valores de desempeño observados y debe reconocerse explícitamente al interpretar las métricas del modelo. En consecuencia, el modelo tiene validez operativa para la priorización de cartera al corte de marzo de 2026, pero no debe interpretarse como evidencia de capacidad predictiva prospectiva sobre unidades que aún no han materializado mora en el tramo crítico.

Interpretabilidad Limitada de los Modelos de Ensamble

Si bien el modelo Random Forest demostró el mayor desempeño estadístico sobre el conjunto de prueba (F1 = 0.880, ROC-AUC = 0.901), su naturaleza de ensamble de árboles de decisión limita la interpretabilidad directa de las predicciones individuales. A diferencia de la Regresión Logística, cuyos coeficientes cuantifican el efecto marginal de cada variable sobre la probabilidad de riesgo, la importancia Gini del Random Forest refleja la contribución promedio de cada predictor a la reducción de impureza a lo largo de todos los árboles, lo que no permite asociar directamente una predicción específica con los factores que la determinaron para una unidad en particular. En contextos administrativos donde las decisiones de cobro deben ser justificadas ante los propietarios o ante instancias legales, la capacidad de explicar individualmente por qué una unidad fue clasificada como riesgo alto tiene valor práctico tan relevante como la precisión estadística del modelo. Técnicas de explicabilidad post-hoc como SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-agnostic Explanations) constituyen líneas de trabajo futuro que abordarían directamente esta limitación.

Ausencia de Validación Temporal Externa

Los modelos fueron entrenados y evaluados sobre datos del mismo corte temporal (marzo de 2026), utilizando una partición aleatoria estratificada para la separación entre entrenamiento y prueba. Esta estrategia, estándar en aprendizaje automático, estima la capacidad de generalización del modelo sobre datos no vistos del mismo período, pero no evalúa su desempeño sobre datos de períodos posteriores. La validación temporal —consistente en entrenar el modelo con datos históricos y evaluarlo sobre un corte futuro— constituye el estándar de referencia para medir la utilidad predictiva real de modelos de riesgo financiero (Hyndman & Athanasopoulos, 2021). Sin esta validación, no es posible confirmar que los patrones aprendidos por el modelo se mantengan estables ante cambios en el comportamiento de pago de la comunidad, variaciones en las políticas de cobro de la administración o shocks económicos externos que alteren las dinámicas de mora.

Definición Operacional del Target y su Impacto en la Clasificación

La variable objetivo target_riesgo_mora fue construida combinando dos condiciones de naturaleza distinta: mora activa en el tramo superior a 90 días —criterio temporal— y estado jurídico o prejurídico —criterio administrativo—. Si bien esta definición es operativamente pertinente para el contexto del conjunto residencial, los dos criterios no siempre coinciden: una unidad con mora superior a 90 días no necesariamente ha iniciado proceso jurídico, y una unidad en estado prejurídico puede tener saldos en tramos de mora más recientes. Esta heterogeneidad en la composición de la clase positiva puede introducir ruido en el entrenamiento del modelo, ya que unidades con perfiles financieros similares podrían recibir etiquetas distintas dependiendo de la gestión administrativa específica aplicada a cada caso. Definiciones alternativas del target, basadas exclusivamente en la antigüedad de la mora o en la probabilidad de escalamiento

jurídico, representan variantes metodológicas que estudios futuros podrían explorar para comparar la estabilidad y consistencia de los resultados obtenidos.

Trabajo Futuro

Incorporación de Datos Longitudinales y Modelos de Evolución Temporal

Como línea de continuidad del presente proyecto, se propone ampliar el alcance del estudio mediante la incorporación de información histórica longitudinal de la cartera, con cortes periódicos mensuales o trimestrales que permitan analizar la evolución del comportamiento de pago de las unidades residenciales en el tiempo. Este enfoque facilitaría el tránsito desde un análisis de corte transversal hacia modelos capaces de capturar trayectorias de deterioro financiero, reincidencia en mora y cambios progresivos en el nivel de riesgo. Desde el punto de vista metodológico, la disponibilidad de datos panel habilitaría la aplicación de modelos de análisis de supervivencia (survival analysis), los cuales estiman no solo si una unidad entrará en mora crítica, sino cuándo es probable que lo haga, incorporando la dimensión temporal como variable central del análisis (Hyndman & Athanasopoulos, 2021). Esta aproximación representaría una mejora sustancial respecto al modelo de clasificación binaria actual, cuya capacidad predictiva prospectiva está limitada por la naturaleza transversal del dataset utilizado.

Enriquecimiento de Variables Predictoras

Futuros trabajos podrían incorporar nuevas variables relacionadas con el comportamiento financiero y administrativo de los propietarios, tales como historial de pagos parciales, frecuencia de acuerdos de pago, reincidencia en mora, cumplimiento de compromisos previos y variaciones en el saldo adeudado entre períodos. La inclusión de estas variables de comportamiento histórico permitiría capturar patrones dinámicos de pago que el modelo actual, construido sobre un corte único, no puede representar. De igual manera, cuando sea ética y legalmente viable, podrían considerarse variables contextuales o socioeconómicas anonimizadas, tales como estrato socioeconómico, años de residencia en el conjunto o antigüedad de la deuda

hipotecaria, que permitan enriquecer la capacidad explicativa del modelo sin afectar la privacidad de los residentes (López & Serán, 2021).

Validación Temporal y Externa en Múltiples Conjuntos

Desde el punto de vista metodológico, se recomienda evaluar los modelos desarrollados mediante validación temporal y validación externa en diferentes conjuntos residenciales. La validación temporal consistiría en entrenar el modelo con cortes históricos anteriores y evaluarlo sobre períodos subsiguientes, midiendo la degradación del desempeño a lo largo del tiempo (model drift) para determinar la frecuencia óptima de reentrenamiento. La validación externa implicaría replicar la metodología en conjuntos residenciales con características distintas —en términos de tamaño, estrato, ubicación o composición de cartera— para evaluar si los patrones identificados se mantienen o requieren ajustes según el contexto. Esta doble validación es el estándar metodológico de referencia en la literatura de scoring crediticio para establecer la validez externa de los modelos (Thomas et al., 2002) y constituye el paso necesario para posicionar el enfoque desarrollado como una metodología replicable en el sector de propiedad horizontal.

Optimización del Umbral de Clasificación Según Costo Asimétrico de Errores

El presente modelo utiliza el umbral estándar de 0.5 para la clasificación binaria. Sin embargo, en la gestión de cartera los costos asociados a los dos tipos de error son inherentemente asimétricos: clasificar una unidad de riesgo alto como sin riesgo (falso negativo) implica no intervenir oportunamente sobre un caso crítico, mientras que clasificar una unidad sin riesgo como riesgo alto (falso positivo) implica una intervención innecesaria de menor costo administrativo. Un trabajo futuro relevante consistiría en construir una matriz de costos explícita que cuantifique el costo diferencial de cada tipo de error —en términos de pérdida financiera

esperada, costo de gestión jurídica y costo de intervención preventiva— y optimizar el umbral de clasificación para minimizar el costo total esperado en lugar del error de clasificación global (James et al., 2023). Este enfoque, denominado clasificación sensible al costo (cost-sensitive classification), está bien documentado en la literatura de credit scoring y puede mejorar significativamente la utilidad operativa del modelo sin requerir cambios en su arquitectura.

Explicabilidad e Interpretabilidad Post-Hoc

Se plantea como línea futura profundizar en técnicas de explicabilidad e interpretabilidad de modelos, como SHAP (SHapley Additive exPlanations) y LIME (Local Interpretable Model-agnostic Explanations), con el fin de comprender con mayor detalle la contribución de cada variable en la clasificación del riesgo de mora a nivel individual. SHAP, fundamentado en la teoría de juegos cooperativos, asigna a cada variable un valor de contribución marginal para cada predicción específica, permitiendo responder preguntas como "¿cuánto aumentó la probabilidad de riesgo de la unidad X por su saldo en el tramo >90 días?" Este nivel de explicabilidad resulta especialmente relevante en contextos administrativos donde las decisiones de cobro deben ser justificables ante los propietarios y ante instancias legales, y donde la adopción de herramientas analíticas requiere confianza por parte de usuarios no especializados. Asimismo, se recomienda explorar el modelo predictivo prospectivo descrito en la sección de limitaciones, entrenado exclusivamente con variables de tramos tempranos de mora (1–90 días) para estimar qué unidades en mora incipiente evolucionarán hacia mora crítica.

Análisis de Equidad Algorítmica

Una dimensión frecuentemente omitida en estudios de scoring aplicados a contextos residenciales es el análisis de equidad algorítmica (algorithmic fairness), que evalúa si el modelo produce predicciones sesgadas hacia subgrupos específicos de la población. En el contexto del

conjunto residencial, esto implicaría verificar si las tasas de falsos positivos o falsos negativos difieren sistemáticamente según la torre de residencia, la antigüedad del propietario en el conjunto, el tipo de unidad (apartamento vs. parqueadero) u otras características administrativas disponibles. Un sesgo sistemático en las predicciones podría generar intervenciones de cobro inequitativas, con implicaciones legales y reputacionales para la administración. Futuros trabajos deberían incorporar métricas de equidad como paridad demográfica, igualdad de oportunidades o calibración por subgrupo, y evaluar el compromiso entre equidad y desempeño predictivo (fairness-accuracy tradeoff) dentro del contexto específico de la propiedad horizontal.

Integración con Sistemas de Gestión y Automatización del Scoring

Futuros desarrollos podrían avanzar hacia la integración del modelo predictivo con los sistemas de gestión administrativa del conjunto residencial, habilitando la automatización periódica del scoring de riesgo por unidad. Esta integración implicaría la construcción de un pipeline de datos que actualice automáticamente las predicciones con cada nuevo corte de cartera, conectando los resultados del modelo con tableros de control en herramientas de inteligencia de negocios como Power BI o Microsoft Fabric. Desde el punto de vista académico, esta línea representa el tránsito desde la fase de modelado (CRISP-DM Fase 4) hacia la fase de despliegue (CRISP-DM Fase 6), cerrando el ciclo metodológico completo. La implementación de alertas tempranas basadas en el modelo permitiría a la administración pasar de una gestión reactiva de cartera —actuar sobre mora ya consolidada— hacia una gestión preventiva y basada en evidencia (Chapman et al., 2000).

Medición del Impacto Organizacional

Finalmente, se recomienda evaluar el impacto organizacional de la aplicación del modelo mediante indicadores cuantificables que vinculen los resultados analíticos con la sostenibilidad

financiera del conjunto residencial. Indicadores relevantes incluirían la reducción porcentual de cartera vencida en tramos superiores a 90 días tras la implementación de estrategias de cobro basadas en el modelo, la disminución en el número de casos escalados a cobro jurídico como consecuencia de intervenciones preventivas oportunas, la mejora en la tasa de recaudo mensual y la optimización del tiempo promedio de gestión por caso. Esta medición de impacto requeriría un diseño cuasi-experimental que compare el desempeño de la cartera en períodos con y sin uso del modelo, controlando por factores externos que pudieran afectar el comportamiento de pago. Este enfoque convertiría el presente proyecto en una línea de investigación aplicada con capacidad de demostrar valor organizacional medible, consolidando la pertinencia de la ciencia de datos en la administración de propiedad horizontal.

Referencias Bibliográficas

Bedoya Ríos, S., & Herrera Arbeláez, D. (2024). Construcción de un modelo para predecir la morosidad de cartera. *Cuaderno Activa, 15*(1). https://doi.org/10.53995/20278101.1229

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/es/CRISP-DM.pdf

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

Congreso de la República de Colombia. (2001). *Ley 675 de 2001, por medio de la cual se expide el régimen de propiedad horizontal*. Diario Oficial No. 44.509. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=3366

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

IBM. (2025). *Guía de CRISP-DM de IBM SPSS Modeler*. IBM Documentation. https://www.ibm.com/docs/es/spss-modeler/saas?topic=overview-crisp-dm-in-spss-modeler

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in Python*. Springer. https://doi.org/10.1007/978-3-031-38747-0

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470316801

López, C., & Serán, J. (2021). Predicción del riesgo de mora en carteras de crédito mediante técnicas de machine learning: Un estudio comparativo. *Revista Finanzas y Política Económica, 13*(2), 341–368. https://doi.org/10.14718/revfinanzpolitecon.v13.n2.2021.3369

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

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

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718317

Velásquez, J. D., & Ramírez, A. (2021). Modelos de predicción de incumplimiento de pago en carteras de vivienda: Evidencia para Colombia. *Cuadernos de Administración, 34*(62). https://doi.org/10.11144/Javeriana.cao34-62.mpip

Apéndices

Apéndice A

Soporte Técnico y Evidencia Reproducible del Desarrollo del Proyecto

Como soporte técnico y evidencia reproducible del desarrollo del proyecto, se creó el repositorio PROYECTO_DE_GRADO_II en GitHub, el cual contiene el código fuente, los datos anonimizados, el notebook de modelado y la documentación del trabajo de grado, organizados para garantizar la trazabilidad del proceso analítico y la reproducibilidad de los resultados.

https://github.com/augustolism90-code/PROYECTO_DE_GRADO_II

Nota. El repositorio documenta el proyecto: Segmentación de cartera y modelo predictivo de mora en propiedad horizontal, UNAD 2026.

Apéndice B

Presentación Audiovisual

Presentación audiovisual en la que se exponen el problema de investigación, la metodología aplicada, los principales resultados, las conclusiones y las recomendaciones del estudio.

https://youtu.be/uYZHsYrNJK0

Nota. El video corresponde a la sustentación del trabajo de grado ante el jurado evaluador, UNAD 2026.