UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTAD DE INGENIERÍAS

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA:

SEGMENTACIÓN DE CLIENTES EN MORA PARA OPTIMIZAR ESTRATEGIAS DE COBRANZA MEDIANTE K-MEANS

Trabajo de Titulación previo a la obtención del título de Magíster en Big Data y Ciencia de Datos.

Autor(a)

Ing. Víctor Eduardo Gordillo Montesdeoca

Tutor(a)

Ing. Pastora Fernanda Martínez Tatamues, Mg.

AMBATO – ECUADOR

2025

AUTORIZACIÓN POR PARTE DEL AUTOR PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL, Y PUBLICACIÓN ELECTRÓNICA DEL TRABAJO DE TITULACIÓN

Yo, Víctor Eduardo Gordillo Montesdeoca, declaro ser autor del Trabajo Titulación con el nombre "SEGMENTACIÓN DE CLIENTES EN MORA PARA OPTIMIZAR ESTRATEGIAS DE COBRANZA MEDIANTE K-MEANS", como requisito para optar al grado de Magister en Big Data y Ciencia de Datos, y autorizo al Sistema de Bibliotecas de la Universidad Indoamérica, para que con fines netamente académicos divulgue esta obra a través del Repositorio Digital Institucional (RDI-UTI).

Los usuarios del RDI-UTI podrán consultar el contenido de este trabajo en las redes de información del país y del exterior, con las cuales la Universidad tenga convenios. La Universidad Indoamérica no se hace responsable por el plagio o copia del contenido parcial o total de este trabajo.

Del mismo modo, acepto que los Derechos de Autor, Morales y Patrimoniales, sobre esta obra, serán compartidos entre mi persona y la Universidad Indoamérica, y que no tramitaré la publicación de esta obra en ningún otro medio, sin autorización expresa de la misma. En caso de que exista el potencial de generación de beneficios económicos o patentes, producto de este trabajo, acepto que se deberán firmar convenios específicos adicionales, donde se acuerden los términos de adjudicación de dichos beneficios.

Para constancia de esta autorización, en la ciudad de Ambato, a los 23 días del mes de septiembre de 2025, firmo conforme:

Autor: Víctor Eduardo Gordillo Montesdeoca

Firma: 

Número de Cédula: 0105477194

Dirección: Pichincha, Quito, Chimbacalle, Pio XII.

Correo Electrónico: vgordillo@indoamerica.edu.ec

Teléfono: 0994618946

APROBACIÓN DEL DIRECTOR

En mi calidad de Director del Trabajo de Titulación "SEGMENTACIÓN DE CLIENTES DE TARJETAS DE CRÉDITO EN MORA PARA OPTIMIZAR ESTRATEGIAS DE COBRANZA MEDIANTE K-MEANS" presentado por Víctor Eduardo Gordillo Montesdeoca, para optar por el Título de Magister en Big Data y Ciencia de Datos.

CERTIFICO

Que dicho Trabajo de Titulación ha sido revisado en todas sus partes y considero que reúne los requisitos y méritos suficientes para ser sometido a la presentación pública y evaluación por parte de los Examinadores que se designe.

Ambato, 19 de septiembre de 2025

martinezpf
Firmado digitalmente por martinezpf
Fecha: 2025.09.23 14:10:38 -05'00'

……

Ing. Pastora Fernanda Martínez Tatamues, Mg.

DIRECTOR

DECLARACIÓN DE AUTENTICIDAD

Quien suscribe, declaro que los contenidos y los resultados obtenidos en el presente Trabajo de Titulación, como requerimiento previo para la obtención del Título de Magister en Big Data y Ciencia de Datos, son absolutamente originales, auténticos y personales y de exclusiva responsabilidad legal y académica del autor

Ambato, 23 de septiembre de 2025

…….

Víctor Eduardo Gordillo Montesdeoca

0105477194

APROBACIÓN DE EXAMINADORES

El Trabajo Titulación ha sido revisado, aprobado y autorizada su impresión y empastado, sobre el Tema: SEGMENTACIÓN DE CLIENTES EN MORA PARA OPTIMIZAR ESTRATEGIAS DE COBRANZA MEDIANTE K-MEANS, previo a la obtención del Título de Magister en Big Data y Ciencia de Datos, reúne los requisitos de fondo y forma para que el estudiante pueda presentarse a la sustentación del Trabajo Titulación.

Ambato, 23 de septiembre de 2025

……..

PhD. Andrés Xavier Rubio Proaño

EXAMINADOR

……..

Ing. Patricio Rodrigo Arellano Vargas.

EXAMINADOR

DEDICATORIA

A Dios.

A mis padres y mi hermano.

AGRADECIMIENTO

A Dios por brindarme la oportunidad para culminar esta etapa.

A mis padres y mi hermano, por su constante apoyo y respaldo a lo largo de este proceso.

A mi tutora de tesis por su acompañamiento, paciencia y apoyo para culminar este trabajo.

ÍNDICE DE CONTENIDOS

PORTADA i

AUTORIZACIÓN DE REPOSITORIO DIGITAL ii

APROBACIÓN DEL TUTOR iii

CAPÍTULO I

INTRODUCCIÓN 1

Planteamiento del Problema 1

Antecedentes de la investigación 3

Desarrollo teórico del objeto y campo 4

Hipótesis o idea que se defiende 8

OBJETIVO GENERAL 9

OBJETIVOS ESPECÍFICOS 9

CAPÍTULO II

METODOLOGÍA 10

Diseño de Investigación 10

Población y Muestra 12

Instrumentos de Recolección 13

Proceso Metodológico 13

Limpieza y preprocesamiento de datos: 14

Descripción de las variables 16

Estadísticas descriptivas de las variables 17

CAPÍTULO III

DESARROLLO DE LA INVESTIGACIÓN 31

Aplicación del algoritmo K-means 31

Análisis de Datos 32

Métricas de Evaluación utilizadas 32

Interpretación de Resultados 33

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN 34

Resultados del Clustering: 34

Visualización de Clusters 36

Discusión de Resultados 43

Limitaciones del estudio 45

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES 46

REFERENCIAS BIBLIOGRÁFICAS 48

ANEXOS 51

Anexo 1: Detalles para la Reproducibilidad del Análisis: 51

ÍNDICE DE TABLAS

Tabla 1: Buckets días mora. 3
Tabla 2: Descripción variables. 16
Tabla 3: Variable Monto. 17
Tabla 4: Variable Días Mora. 19
Tabla 5: Variable Edad. 20
Tabla 6: Variable Cupo. 22
Tabla 7: Variable Ingreso Mensual. 24
Tabla 8: Variable Pagos Puntuales. 25
Tabla 9: Variable Pagos Retrasados. 26
Tabla 10: Variable Género. 28
Tabla 11: Variable Estado Civil. 29
Tabla 12: Segmentación Clusters. 35
Tabla 13: Evaluación de la calidad de la segmentación. 44
Tabla 14: Paquetes de R y versión. 51

ÍNDICE DE GRÁFICOS

Gráfico No. 1 Flujo de trabajo CRISP-DM. 11
Gráfico No. 2 Pipeline resumido 14
Gráfico No. 3 Comparación de la distribución del Monto de Crédito antes y después del tratamiento de outliers 15
Gráfico No. 4 Distribución Monto 18
Gráfico No. 5 Distribución Días Mora 20
Gráfico No. 6 Distribución Edad 21
Gráfico No. 7 Distribución Cupo 23
Gráfico No. 8 Distribución Ingreso Mensual 24
Gráfico No. 9 Distribución Pagos Puntuales 26
Gráfico No. 10 Distribución Pagos Retrasados 27
Gráfico No. 11 Distribución Género 28
Gráfico No. 12 Distribución Estado Civil 30
Gráfico No. 13 Número óptimo de cluster 31
Gráfico No. 14 Segmentación de Clientes (K-means) 34
Gráfico No. 15 Segmentación de Clientes Ingreso Mensual vs Días Mora 37
Gráfico No. 16 Segmentación de Clientes Monto vs Días Mora 38
Gráfico No. 17 Distribución del Monto por Cluster 39
Gráfico No. 18 Distribución de Edad por Cluster 40
Gráfico No. 19 Distribución de Pagos Puntuales por Cluster 41
Gráfico No. 20 BoxPlot Monto por Cluster 42
Gráfico No. 21 BoxPlot Días Mora por Cluster 43

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA
FACULTAD DE INGENIERÍAS
MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

TEMA: SEGMENTACIÓN DE CLIENTES EN MORA PARA OPTIMIZAR ESTRATEGIAS DE COBRANZA MEDIANTE K-MEANS.

AUTOR: Ing. Víctor Eduardo Gordillo Montesdeoca
TUTORA: Ing. Pastora Fernanda Martínez Tatamues

RESUMEN EJECUTIVO

La gestión de la mora en tarjetas de crédito representa un reto significativo para las instituciones financieras, dado su impacto en el riesgo crediticio y la rentabilidad. Este trabajo plantea un modelo para clasificar a los clientes en mora utilizando el algoritmo K-means, con el propósito de identificar patrones de comportamiento que faciliten la implementación de estrategias más eficaces para la recuperación de deudas. La investigación se basó en la metodología CRISP-DM, abarcando la recolección y evaluación de la información histórica de los clientes, el procesamiento de variables como la edad, el monto adeudado y los días de mora, y la aplicación del clustering para agrupar a los clientes en categorías con características similares. La selección del número adecuado de grupos se realizó el método Elbow (codo), y la calidad del modelo se validó a través del coeficiente de Silhouette.

Los hallazgos identificaron tres grupos principales de clientes morosos, diferenciados por sus niveles de deuda, el tiempo en mora y sus hábitos de pago. El modelo demostró ser efectivo para clasificar a los clientes según su riesgo, permitiendo identificar estrategias focalizadas de cobranza.

DESCRIPTORES: cobranza, K-means, morosidad, segmentación.

UNIVERSIDAD TECNOLÓGICA INDOAMÉRICA

FACULTY OF ENGINEERING

MASTER'S IN BIG DATA AND DATA SCIENCE

AUTHOR: GORDILLO MONTESDEOCA VICTOR

TUTOR: null MARTINEZ TATAMUES PASTORA

ABSTRACT

Customer segmentation of delinquent accounts to optimize collection strategies using K-means.

Credit card delinquency management poses a significant challenge for financial institutions, given its impact on credit risk and profitability. This study proposes a model for classifying delinquent customers using the K-means algorithm, with the aim of identifying behavioral patterns that facilitate the implementation of more effective debt recovery strategies. The research was based on the CRISP-DM methodology, encompassing the collection and evaluation of historical customer information, the processing of variables such as age, amount owed, and days past due, and the application of clustering to group customers into categories with similar characteristics. The appropriate number of clusters was selected using the Elbow method, and the model's quality was validated using the Silhouette coefficient. The findings identified three main groups of delinquent customers, differentiated by their debt levels, delinquency duration, and payment habits. The model proved effective in classifying customers by risk, enabling the identification of targeted collection strategies.

KEYWORDS:

Debt collection, delinquency, K-means, segmentation.

CAPÍTULO I

INTRODUCCIÓN

Planteamiento del Problema

La gestión de la morosidad relacionada con el uso de tarjetas de crédito constituye actualmente un desafío importante para las entidades financieras, debido a su impacto directo en la rentabilidad y en la estabilidad de sus operaciones. El incumplimiento de pagos por parte de los clientes genera costos adicionales asociados a la gestión de cobranza, reduce el valor y la solidez de la cartera crediticia, incrementando el nivel de riesgo financiero de la entidad (Han, Pei & Kamber, 2011). A pesar de estos desafíos, muchas instituciones financieras continúan utilizando estrategias de recuperación de deudas que son genéricas y poco adaptadas a las características específicas de cada cliente. Estas estrategias tradicionales, que no consideran las diferencias en el comportamiento y perfil de los clientes, tienden a ser ineficaces y costosas (Jain, Murty, & Flynn, 1999).

La falta de una segmentación adecuada de los clientes en mora es una de las principales limitaciones de las estrategias de cobranza actuales. Al tratar a todos los clientes de la misma manera, las instituciones no logran identificar con precisión a aquellos que presentan un mayor riesgo de incumplimiento o que podrían regularizar su situación mediante acciones específicas (Ng, Jordan, & Weiss, 2002). Esto no solo incrementa los costos de recuperación, sino que también deteriora la relación con los clientes, afectando la imagen institucional, y limitando las oportunidades de la retención y fidelización.

Según Bholowalia y Kumar (2014), la personalización de las estrategias de recuperación de deudas mediante técnicas de segmentación mejora considerablemente las tasas de recuperación y reduce los costos operativos. Sin embargo, muchas instituciones financieras aún no han adoptado herramientas de análisis avanzadas que les permitan identificar y entender los patrones de comportamiento de sus clientes en mora. La creciente disponibilidad de datos y el desarrollo de técnicas de análisis de datos basadas en las técnicas

de inteligencia artificial y el uso del aprendizaje automático brindan posibilidades innovadoras para abordar este problema de manera más efectiva.

En el contexto ecuatoriano, el nivel de morosidad de las tarjetas de crédito ha presentado fluctuaciones importantes. Según la Superintendencia de Bancos del Ecuador, la tasa de morosidad de tarjetas de crédito superó el 9% en el año 2023, reflejando un deterioro del comportamiento de pago de los clientes en este segmento específico del portafolio (Superintendencia de Bancos, 2023). Esta situación se agrava en periodos de contracción económica, donde el desempleo y la informalidad dificultan la recuperación de cartera vencida. Frente a esta problemática, algunas instituciones han implementado modelos de aprendizaje automático para predecir el riesgo crediticio y segmentar a los clientes, como el estudio de Qi et al. (2020), que aplicó algoritmos de clustering para clasificar a los deudores según su propensión al incumplimiento, logrando mejorar en un 17% la efectividad de sus estrategias de cobranza. Estos antecedentes muestran que el uso de técnicas de machine learning no solo es viable, si no también eficaz para optimizar la recuperación de cartera y reducir la exposición al riesgo crediticio.

Para abordar la problemática de la morosidad en tarjetas de crédito, se propone la aplicación de métodos de aprendizaje automático no supervisado, en particular el algoritmo K-means, el cual ha probado su eficacia para detectar patrones no evidentes dentro de extensos conjuntos de datos (Jain et al., 1999). El clustering permite clasificar a los clientes en grupos homogéneos según rasgos compartidos, lo que permite desarrollar estrategias de cobranza adaptadas y con mayor efectividad (Han et al., 2011).

En el contexto crediticio la morosidad se refiere al incumplimiento en el pago de las obligaciones dentro del periodo establecido contractualmente. En la institución financiera objeto de este estudio, un cliente se considera en mora a partir del día 31 de atraso desde la fecha de vencimiento del pago mínimo. Este umbral coincide con lo que se conoce en la literatura como morosidad temprana, y marca un punto de inflexión clave para la activación de procesos de cobranza preventiva o intensiva.

A nivel general, las entidades financieras agrupan los niveles de mora en buckets de días vencidos, como se muestra a continuación:

Tabla 1: Buckets días mora.

Bucket	Clasificación
1 – 30 días	Cliente al día / atraso leve
31 – 59 días	Morosidad temprana
60 – 89 días	Morosidad intermedia
Mayor o igual a 90 días	Morosidad severa

Elaborado por: Gordillo, Víctor (2025).

Estos buckets permiten una segmentación inicial del riesgo, pero no son suficientes por sí solos para caracterizar la probabilidad de recuperación. Por ello se recurre a la segmentación por clustering que considera otros factores asociados al comportamiento financiero del cliente.

Antecedentes de la investigación

La morosidad en el sector financiero ha sido objeto de múltiples estudios debido a su impacto directo en la rentabilidad y la estabilidad de las instituciones financieras. El cumplimiento de pagos por parte de los clientes genera costos adicionales asociados a la gestión de cobranza, afecta la calidad de la cartera crediticia y aumenta el riesgo financiero de las instituciones (Han, Pei, & Kamber, 2011). Diversas investigaciones han abordado este problema destacando la necesidad de adoptar enfoques más sofisticados para la gestión de morosidad. Jain, Murty y Flynn (1999) subrayan que la clasificación de clientes a través de métodos de clustering se ha consolidado como una herramienta efectiva para identificar grupos con

comportamientos similares, lo que facilita la implementación de estrategia de cobranza personalizadas. A través de la identificación de patrones de comportamiento en los clientes en mora, las instituciones pueden diseñar acciones específicas que incrementen la efectividad en la recuperación de deudas.

En investigaciones de Bholowalia y Kumar (2014), demostraron que el uso conjunto del algoritmo K-means y la técnica del codo para definir la cantidad óptima de clusters mejora la precisión en la segmentación de clientes. Este enfoque ha sido aplicado exitosamente en el análisis de morosidad, permitiendo identificar aquellos clientes con mayor probabilidad de regularizar sus pagos y aquellos que representan un mayor riesgo de incumplimiento.

Ng, Jordan y Weiss (2002) también enfatizan la importancia del preprocesamiento de datos y la correcta selección de variables para mejorar la efectividad de los modelos de clustering. Variables como el monto de la deuda, los días en mora, la edad del cliente y el historial de pagos han sido identificadas como factores clave en la segmentación de clientes morosos.

En el contexto latinoamericano, estudios realizados en Perú y Chile han mostrado resultados prometedores en el uso de métodos de agrupamiento (clustering) para la gestión de morosidad en instituciones financieras. Sin embargo, existe una necesidad de investigaciones adicionales que adapten estas metodologías a las particularidades del mercado ecuatoriano y otros mercados emergentes.

Desarrollo teórico del objeto y campo

Morosidad en Tarjetas de crédito: Concepto y Desafíos

Si bien ya se ha introducido la problemática general de la morosidad, es pertinente profundizar en su definición técnica. En el contexto de tarjetas de crédito, la morosidad se refiere al incumplimiento de los pagos mínimos requeridos después de la fecha de vencimiento. Este fenómeno puede estar asociado a diversos factores, como el sobreendeudamiento, la falta de educación financiera o situaciones económicas adversas que afectan la capacidad de pago del cliente (Rose & Hudgins, 2013).

El impacto de la morosidad en las instituciones financieras es significativo. No solo afecta la liquidez y la rentabilidad, sino que también puede deteriorar la relación con los clientes y dañar la reputación de la institución. La gestión de la morosidad, por lo tanto, no se limita a

la recuperación de deudas, sino que también implica la implementación de políticas que permitan prevenir el incumplimiento y mitigar el riesgo crediticio (Han et al., 2011).

Segmentación de clientes: Herramienta clave para la Gestión de Morosidad

La segmentación de clientes es una técnica de marketing y análisis que permite clasificar un conjunto heterogéneo de individuos en grupos más pequeños y homogéneos en función de características compartidas. En el ámbito financiero, esta clasificación permite identificar patrones de comportamiento en los clientes, lo que facilita la personalización de productos y servicios y estrategias de cobranza (Kotler & Keller, 2012).

Es importante distinguir la segmentación tradicional de técnicas modernas como el clustering. Mientras la segmentación puede basarse en reglas predefinidas o criterios demográficos y comerciales, el clustering es una técnica de aprendizaje automático no supervisado que permite descubrir grupos ocultos dentro de los datos, sin necesidad de criterios iniciales. Es decir, segmentar es el objetivo, mientras que el clustering es una de las herramientas más efectivas para lograrlo.

Según Bester y Rosman (2024), la segmentación basada en datos transaccionales ha permitido optimizar la recuperación de cartera vencida en instituciones financieras latinoamericanas.

Existen varios criterios de segmentación:

1. **Demográficos:** Edad, Género, relación dependencia, estado civil, nivel de estudios.
2. **Geográficos:** Ubicación geográfica a diferentes niveles (región, provincia, ciudad).
3. **Psicográficos:** Estilo de vida, valores, actitudes hacia el crédito.
4. **Comportamentales:** Historial de pagos, frecuencia de uso de la tarjeta, días en mora, monto de deuda, proporción de uso de tarjeta respecto al cupo.

La segmentación basada en comportamientos financieros es particularmente relevante para la gestión de morosidad, ya que permite a las instituciones financieras identificar a los clientes con mayor riesgo de incumplimiento y diseñar planes de acción específicos para cada grupo (Jain et al., 1999).

Aprendizaje Automático y Clustering en el Ámbito Financiero

El aprendizaje automático (machine learning) constituye una rama de la inteligencia artificial enfocada en desarrollar algoritmos capaces de aprender a partir de los datos y optimizar su desempeño con el tiempo, sin requerir una programación explícita para cada tarea específica (Bishop, 2006). Este tipo de aprendizaje se clasifica en:

1. **Aprendizaje supervisado:** El modelo se entrena utilizando datos que ya cuentan con etiquetas y para los cuales se conoce la respuesta correcta.
2. **Aprendizaje no supervisado:** El modelo analiza datos sin etiquetar, con el objetivo de detectar patrones o estructuras ocultas. Este último enfoque es el que se emplea en las técnicas de clustering.

El uso de machine learning en morosidad ha cobrado relevancia en investigaciones recientes en América Latina, destacando su utilidad para mejorar modelos predictivos de riesgo. En este campo, el clustering facilita la segmentación de clientes, detección de anomalías y priorización de carteras (Morales-Vargas et al., 2023).

Entre los algoritmos de clustering más utilizados se encuentran:

1. **K-means:** Método que agrupa datos en *k* clusters, minimizando la variación interna de cada grupo.
2. **DBSCAN:** Algoritmo que se basa en la densidad para identificar clusters con formas arbitrarias y detectar puntos atípicos (Ester et al., 1996)
3. **Gaussian Mixture Models (GMM) y clustering jerárquico:** Estrategia que construye una jerarquía de clusters mediante un enfoque aglomerativo o divisivo.

Aunque este estudio se centra en K-means, se reconoce que otros algoritmos pueden ofrecer ventajas metodológicas. Investigaciones recientes en Ecuador han comparado K-means con DBSCAN y GMM, señalando que la elección depende del tipo y calidad de los datos (Sánchez-Morán & Vargas-Aguiar, 2023).

Las aplicaciones del aprendizaje automático en el sector financiero han avanzado considerablemente en los últimos años. En 2023, Jabeen et al. desarrollaron un modelo basado en Light Gradient Boosting Machine (LightGBM) para predecir impago de tarjetas de crédito, demostrando mejoras en la precisión predictiva frente a modelos tradicionales.

Simultáneamente, García (2023) identificó las variables más relevantes en la predicción de morosidad utilizando técnicas como Categorical Boosting (CatBoost) y Shapley Additive exPlanations (SHAP) para explicar su importancia. Otros estudios recientes abordaron la predicción del riesgo de crédito mediante Synthetic Minority Over-sampling Technique (SMOTE) y LightGBM (Naik, 2021), mientras que Pestana (2025) aplicó redes neuronales y Extreme Gradient Boosting (XGBoost) para automatizar ajustes de límites de crédito, optimizando decisiones bancarias con validación cruzada. Estos estudios refuerzan la importancia del uso de técnicas avanzadas de machine learning más allá de K-means en la gestión de riesgo crediticio.

Justificación del uso del algoritmo K-means

El algoritmo K-means ha sido ampliamente utilizado por su eficiencia computacional, facilidad de implementación e interpretación. Su desempeño ha sido validado en aplicaciones reales de entidades financieras de la región (Lozano, García, & Carrillo, 2020; Ramírez & Torres, 2021). Aunque no se aplicaron comparaciones formales en esta investigación, se reconoce la necesidad de futuras comparaciones con modelos más complejos como GMM o clustering jerárquico para reforzar la solidez metodológica y explorar escenarios con datos menos estructurados.

Funcionamiento del Algoritmo K-means:

1. **Inicialización:** Se selecciona el número k de clusters y se eligen aleatoriamente k centroides iniciales.
2. **Asignación:** Cada observación se asocia al cluster cuyo centroide esté más próximo.
3. **Actualización:** Se recalculan los centroides de cada cluster promediando los puntos asignados a él.
4. **Iteración:** Los pasos de asignación y actualización se repiten hasta que los centroides se estabilizan o se alcanza un límite predefinido de iteraciones.

La determinación del número óptimo de clusters es un paso crucial en el uso de K-means. El método del codo (Elbow Method) es una técnica comúnmente utilizada para este propósito. Consiste en graficar la suma de los errores cuadrados (Sum of Squared Errors, SSE) en función del número de clusters y buscar el punto en el que la disminución de la variación

interna comienza a disminuir, lo que indica el número óptimo de clusters (Bholowalia & Kumar, 2014).

Aplicación de K-means en la Gestión de Morosidad

La aplicación del algoritmo K-means en la gestión de morosidad posibilita reconocer patrones de comportamiento entre los clientes en mora, facilitando la implementación de estrategias de cobranza personalizadas. Esta técnica ofrece varias ventajas:

1. **Identificación de patrones críticos:** El clustering permite detectar grupos de clientes con características similares, como aquellos con morosidad reciente pero un historial de pagos positivo, y aquellos con morosidad prolongada y antecedentes de incumplimiento (Ng et al., 2002).
2. **Optimización de recursos:** La segmentación facilita la asignación de recursos de cobranza de manera eficiente, enfocándose en los clientes que representan un mayor riesgo de incumplimiento (Bholowalia & Kumar, 2014).
3. **Personalización de estrategias de cobranza:** La segmentación permite diseñar estrategias adaptadas a las características de cada grupo de clientes, lo que incrementa la probabilidad de éxito en la recuperación de deudas (Ester et al., 1996).
4. **Mejora en la toma de decisiones:** El análisis de los resultados del clustering proporciona datos relevantes que apoyan la toma de decisiones estratégicas en la gestión de morosidad, permitiendo a las instituciones financieras identificar oportunidades de mejora en sus procesos (Han et al., 2011).

La implementación del algoritmo K-means en el software R, facilita la automatización del proceso de segmentación y análisis de datos. Herramientas como dplyr para la manipulación de datos, ggplot2 para la visualización de resultados y factorextra para la interpretación de los clusters son fundamentales para garantizar la efectividad del análisis (Wickham, 2016).

Hipótesis o idea que se defiende

La idea principal que sustenta este proyecto plantea que clasificar a los clientes con tarjetas de crédito en mora mediante el uso del algoritmo K-means mejorará de forma notable la eficiencia de las estrategias de cobro. Esta propuesta parte de la idea de que examinar grandes

cantidades de datos con técnicas de aprendizaje automático facilita descubrir patrones de conducta que resultan difíciles de detectar con métodos convencionales (Ng et al., 2002).

OBJETIVO GENERAL

Elaborar un modelo de segmentación basado en el algoritmo K-means para identificar patrones de comportamiento de pago en clientes de tarjeta de crédito, con el fin de optimizar las estrategias de cobranza y reducir el riesgo crediticio en la institución financiera.

OBJETIVOS ESPECÍFICOS

- Recolectar y preprocesar datos históricos de clientes en mora con el objetivo de asegurar la precisión y uniformidad de la información empleada durante el análisis.
- Identificar y seleccionar variables clave como sociodemográficas, monto de la deuda, días mora e historial de pagos, que afectan las conductas de pago de los clientes objeto de estudio.
- Aplicar algoritmo K-means para agrupar a los clientes en segmentos uniformes según sus rasgos financieros y conductuales.
- Interpretar los resultados de segmentación para identificar patrones de comportamiento entre los diferentes grupos de clientes.

CAPÍTULO II

METODOLOGÍA

La metodología de esta investigación está orientada a desarrollar un modelo de segmentación para identificar patrones de comportamiento en clientes de tarjeta de crédito que presentan días de morosidad, utilizando técnicas de aprendizaje automatizado. La investigación se apoya en un enfoque cuantitativo, con la aplicación del algoritmo K-means para agrupar a los clientes en segmentos homogéneos y facilitar la toma de decisiones estratégicas en la gestión de la morosidad.

Diseño de Investigación

Este estudio adopta un diseño no experimental, ya que se limita a observar y analizar datos históricos de clientes de tarjetas de crédito en mora, sin manipulación directa de variables.

La investigación es descriptiva y cuantitativa, dado que busca identificar y describir patrones de comportamiento mediante el análisis de datos numéricos. Además, es un estudio transversal, ya que se analiza un conjunto de datos en un momento específico, sin realizar seguimiento longitudinal a lo largo del tiempo.

El enfoque metodológico se estructura siguiendo el ciclo de vida CRISP-DM (Cross-Industry Standard Process for Data Mining), que contempla las siguientes etapas:

1. Comprensión del negocio: Análisis del impacto de la morosidad en la institución financiera.
2. Comprensión de los datos: Evaluación de las variables disponible en la base de datos.
3. Preparación de los datos: Procesos de limpieza, transformación y creación de nuevas variables para el análisis.
4. Modelado: Implementación del algoritmo K-means para la segmentación.
5. Evaluación: Validación de la calidad de los clusters mediante métricas.
6. Implementación: Interpretación y recomendaciones para efectuar nuevas estrategias de cobranza.

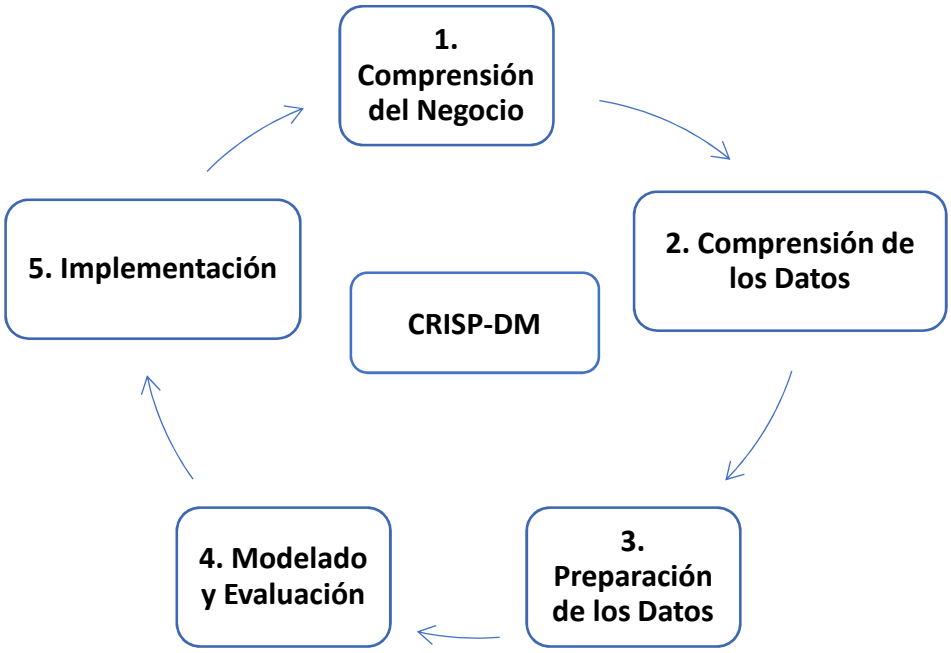

Gráfico No. 1 Flujo de trabajo CRISP-DM.

Elaborado por: Gordillo, Víctor (2025).

El proceso de análisis incluye las siguientes etapas:

1. Recolección y Preprocesamiento de Datos: La primera fase implica la recopilación de datos históricos de clientes en mora, incluyendo variables clave como la edad del cliente, el monto de la deuda, los días en mora, el historial de pagos. Estos datos serán limpiados y normalizados para garantizar su calidad y precisión en el modelado (Bishop, 2006).
2. Análisis Exploratorio de Datos (EDA): En esta etapa se utilizarán técnicas de visualización y análisis estadístico exploratorio para identificar tendencias y patrones iniciales en los datos. Esto incluye la elaboración de histogramas, gráficos de dispersión y estudios de correlación para comprender mejor las relaciones entre las variables (Hastie, Tibshirani & Friedman, 2009).
3. Modelado y Evaluación: Se empleará el algoritmo K-means con el fin de clasificar a los clientes en segmentos homogéneos dentro de sí mismo, pero heterogéneos entre

ellos. La calidad de los grupos obtenidos se verificará utilizando indicadores como el coeficiente de Silhouette, que mide la coherencia interna de los clusters.

4. Interpretación y Aplicación de Resultados: Una vez que los clientes hayan sido segmentados, se procederá a interpretar las características de cada grupo para diseñar estrategias de cobranza específicas. Esto permitirá enfocar los recursos de la institución en los segmentos que presentan un mayor riesgo de morosidad, mientras se ofrecen soluciones personalizadas a los clientes con mayor probabilidad de regularizar sus pagos (Ng et al., 2002).

Población y Muestra

La población objetivo de esta investigación está integrada por clientes de una institución financiera que poseen tarjetas de crédito y presentan morosidad en sus pagos. Para fines de la investigación, se ha utilizado una muestra de 25,004 registros de clientes que, dentro del periodo de análisis (enero a diciembre 2024), hayan presentado al menos un mes con más de 30 días de mora. Este umbral se eligió porque representa una morosidad significativa desde el punto de vista operativo de la institución financiera.

La variable días en mora utilizada en los análisis refleja el número de días en mora al momento del corte del mes más reciente disponible, y no necesariamente representa el máximo valor histórico del cliente. Por tanto, si bien algunos registros pueden mostrar 0 días en el mes final, estos clientes si cumplieron el criterio de haber presentado más de 30 días de mora en algún mes dentro del periodo anual.

Se descartó la exigencia de mora en todos los meses, ya que no refleja el comportamiento real de los clientes.

La muestra proviene de una sola institución financiera del sistema financiero ecuatoriano. Aunque representa únicamente a esta entidad, se considera adecuada por su volumen, cobertura nacional y la diversidad de perfiles crediticios incluidos. No obstante, los resultados deben interpretarse en función de este contexto institucional, recomendándose su validación futura con datos de otras entidades para evaluar su generalización.

Criterios de inclusión:

1. Se considera únicamente personas naturales, ya que los criterios de evaluación y manejo de crédito en personas jurídicas difieren considerablemente.
2. Registros completos con información relevante sobre variables clave como edad, ingreso mensual, monto de deuda, cupo de la tarjeta, días en mora e historial de pagos.

Criterios de exclusión:

1. Clientes cuya deuda haya sido cancelada.
2. Valores atípicos en las variables.

Instrumentos de Recolección

Fuentes de Datos:

La información empleada en este estudio proviene de bases de datos internas que contienen información histórica, donde incluye detalles sobre el perfil demográfico y financiero de los clientes, así como su historial de pagos y comportamiento de días mora.

Herramientas Tecnológicas usadas:

El análisis de datos y la implementación del modelo de clustering se realizarán utilizando el software R, una herramienta ampliamente reconocida en el ámbito de la ciencia de datos por su capacidad para el análisis estadístico, ya que mantiene una gran cantidad de funciones estadísticas que facilitan dicho análisis y la visualización de datos.

Paquetes de R utilizados:

1. dplyr, tidyr: para la manipulación y limpieza de datos.
2. ggplot2: Para la visualización gráfica de los resultados del análisis
3. cluster y factoextra: Para la aplicación y evaluación del algoritmo K-means.
4. Nbclust: Para determinar el número óptimo de clusters.

Proceso Metodológico

El proceso metodológico se basa en las fases CRISP-DM, adaptadas a las necesidades específicas del estudio. A continuación, se describe el pipeline seguido:

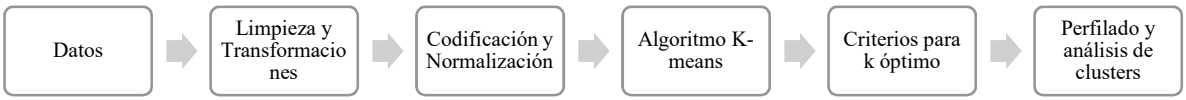

Gráfico No. 2 Pipeline resumido.

Elaborado por: Gordillo, Víctor (2025).

Limpieza y preprocesamiento de datos:

Antes de aplicar el algoritmo K-means, es necesario preparar la información para asegurar que cumpla con los requisitos de calidad y sea adecuada para el análisis. Este procedimiento incluye:

1. **Calidad y completitud de datos:**

   La base de datos proviene directamente de los sistemas internos de la institución financiera, por lo cual no se eliminaron registros ni se realizaron imputaciones, al haberse validado previamente su integridad. La completitud y consistencia de los campos fue verificada de manera general, sin identificarse errores o inconsistencias que afecten el análisis.

2. **Manejo de valores atípicos:**

   En esta investigación, no se encontraron valores atípicos significativos tras aplicar métodos de detección como el rango intercuartílico (IQR). Esta ausencia puede deberse un proceso riguroso de depuración y control de calidad en los datos por parte de la institución financiera, motivo por el cual no fue necesario aplicar transformaciones adicionales ni eliminar registros. Esto garantiza la integridad de la muestra sin comprometer la robustez del análisis. No obstante, con el fin de verificar gráficamente el impacto de los valores extremos, se realizó una comparación visual entre la distribución del monto de crédito antes y después de aplicar una técnica de winsorización moderada, que consiste en limitar los valores extremos sin eliminarlos. En el Gráfico No. 3 se puede observar que, si bien existen observaciones con montos elevados, su frecuencia es baja y no distorsiona significativamente la forma general

de la distribución. Por tanto, se optó por conservar los datos originales, evitando así comprometer la integridad del conjunto.

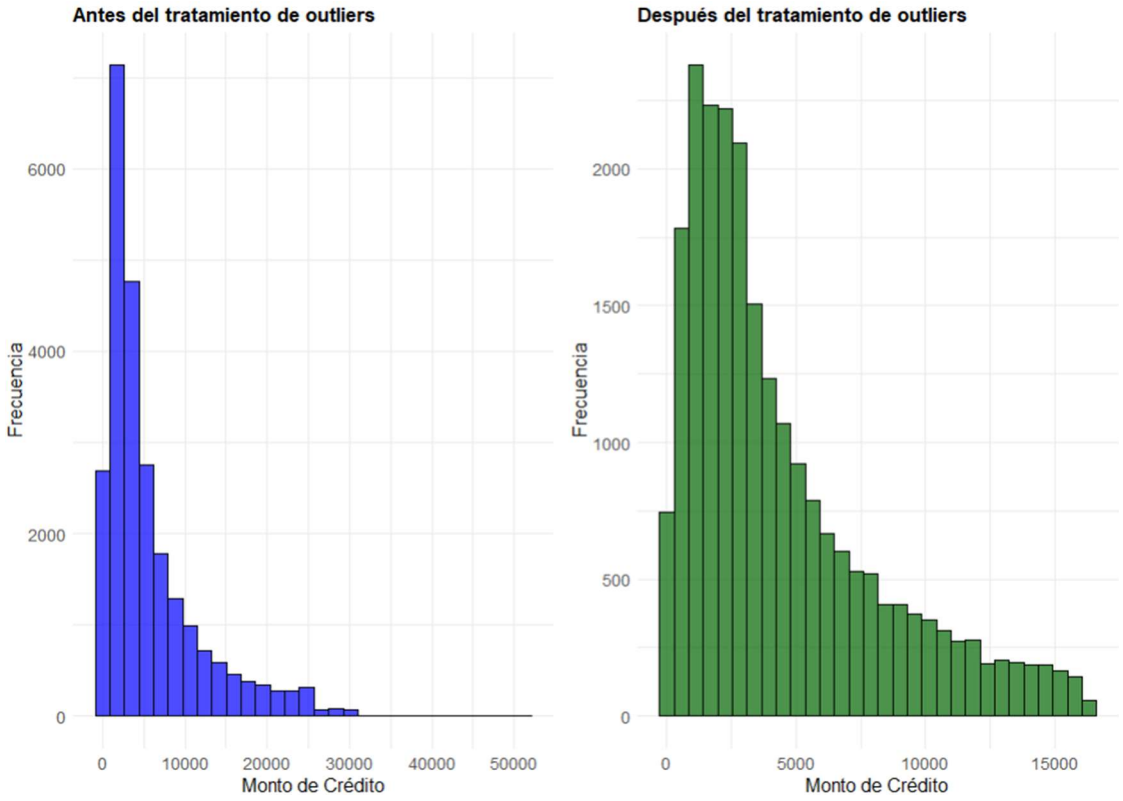

Gráfico No. 3 Comparación de la distribución del Monto de Crédito antes y después del tratamiento de outliers.

Elaborado por: Gordillo, Víctor (2025).

3. **Codificación de variables categóricas:**

   Las variables categóricas (por ejemplo, género, estado civil) fueron transformadas mediante codificación One-Hot. Si bien K-means no es idóneo para variables categóricas debido a su dependencia de la distancia euclidiana, se optó por esta codificación por dos razones:

   - Permite transformar las categorías en una forma numérica interpretable para el algoritmo.

- Facilita la posterior comparación con otros métodos como k-prototypes o Partitioning Around Medoids (PAM) con distancia de Gower, lo que se sugiere como trabajo futuro.

4. **Normalización de variables numéricas:**

   Las variables numéricas como edad, monto de deuda, cupo, ingreso mensual y días mora se normalizarán utilizando la técnica min-max scalling, con el fin de mantener una escala uniforme para el análisis. Se eligió la técnica de min-max scalling sobre otros métodos como la estandarización (z-score), debido a que mantiene los valores originales dentro de un rango entre 0 y 1, lo cual resulta más adecuado para algoritmos de clustering como K-means que son sensibles a la escala de las variables. Además, esta técnica evita la influencia de unidades de medidas distintas, garantizando una mayor comparabilidad entre variables como monto, cupo, edad, días en mora.

Descripción de las variables

Tabla 2: Descripción variables.

Variable	Descripción	Tipo de Dato
Monto	Valor de la deuda del cliente	Numérica
Días en mora	Días que el cliente ha estado en mora	Numérica
Edad	Edad del cliente en años	Numérica
Cupo	Cupo del cliente	Numérica
Ingreso Mensual	Ingreso Mensual del cliente	Numérica

Pagos Puntuales	Cantidad de pagos realizados a tiempo en los últimos 12 meses	Numérica
Pagos Retrasados	Cantidad de pagos realizados con retraso en los últimos 12 meses	Numérica
Género	Género del cliente	Categórica
Estado Civil	Estado civil actual del cliente	Categórica

Elaborado por: Gordillo, Víctor (2025).

Las variables numéricas fueron escaladas utilizando la técnica min-man scalling para garantizar homogeneidad en la escala. Todas las variables están expresadas en sus unidades originales: edad en años, ingresos y montos en USD. Se evaluó la correlación entre variables para evitar posibles problemas de colinealidad o filtración de información (feature leakage). En particular, las variables relacionadas con pagos fueron registradas hasta el mes previo al evento de mora, asegurando que no se utilizará información futura en la segmentación.

Estadísticas descriptivas de las variables

Variable Monto:

Tabla 3: Variable Monto.

Mínimo	1.03
Q1	1,745.49
Mediana	3,478.23
Media	5,768.35
Q3	7,591.36
Máximo	51,358.24

Desviación Estándar	6,014.76

Elaborado por: Gordillo, Víctor (2025).

La Tabla 3 presenta las estadísticas descriptivas de la variable Monto de crédito. Se observa que el valor mínimo es de $1.03, mientras que el valor máximo alcanza los $51,358.24, evidenciando una gran dispersión entre los clientes. La mediana es de $3,478.23 y la media de $5,768.35, lo cual indica que la distribución es asimétrica positiva, influenciada por valores atípicos altos. El primer cuartil (Q1) se ubica en $1,745.49 y el tercer cuartil (Q3) en $7,591.36, lo que sugiere que el 50% central de los montos se encuentra dentro de ese rango. La desviación estándar de $6,014.76 confirma la alta variabilidad en los valores de crédito otorgado. Este análisis permite evidenciar la necesidad de diferenciar estrategias de cobranza según el nivel de deuda de los clientes.

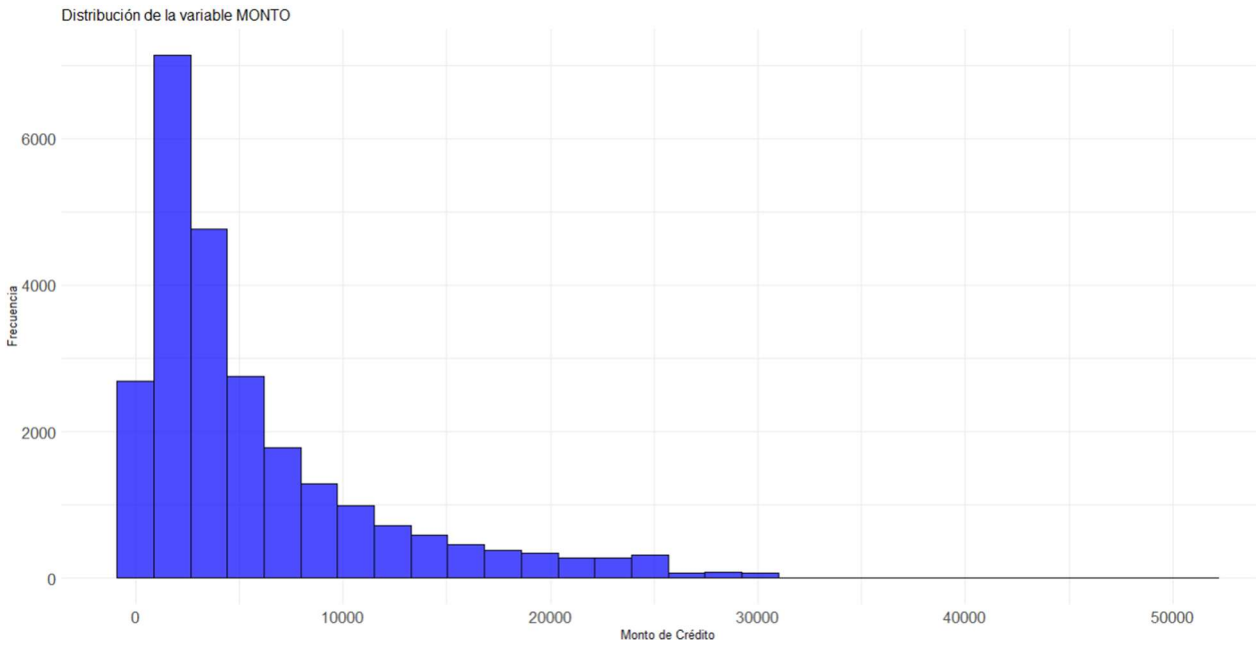

Gráfico No. 4 Distribución Monto.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 4 se observa claramente que la mayoría de los clientes mantiene deudas por debajo de los $10,000, concentrándose especialmente entre $1,000 y $5,000, lo cual

representa una alta frecuencia en rangos bajos. La distribución es asimétrica positiva, con una cola extendida hacia la derecha debido a un pequeño grupo de clientes con montos de deuda muy elevados, incluso superiores a $ 50,000. Esta estructura de la distribución permite inferir que el portafolio de crédito está compuesto principalmente por montos moderados, pero también existen casos extremos. Por ello se recomienda diseñar políticas de cobranza segmentadas: una para la mayoría de deudores con montos bajos o medios, y otra para los casos de alta exposición que podrían representar un mayor riesgo financiero.

Variable Días en Mora:

Tabla 4*:* Variable Días Mora.

Mínimo	0
Q1	0
Mediana	15
Media	43.64
Q3	77
Máximo	169
Desviación Estándar	51.02

Elaborado por: Gordillo, Víctor (2025).

La Tabla 4 expone las estadísticas descriptivas de la variable Días en Mora. El valor mínimo y el primer cuartil (Q1) son ambos 0, lo cual indica que al menos el 25% de los clientes no tiene morosidad. La mediana es de 15 días, mientras que la media es significativamente mayor, con 43.64 días, lo que sugiere nuevamente una asimetría positiva en la distribución. El tercer cuartil (Q3) es 77 días y el valor máximo registrado asciende a 169 días en mora. La desviación estándar de 51.02 refleja una considerable dispersión, lo cual implica que existe un grupo importante de clientes con moras extensas, a pesar de que muchos no presentan atraso alguno. Esta tabla evidencia la heterogeneidad en los comportamientos de pago, aspecto crucial para segmentar clientes por nivel de riesgo y diseñar políticas de recuperación más adecuadas.

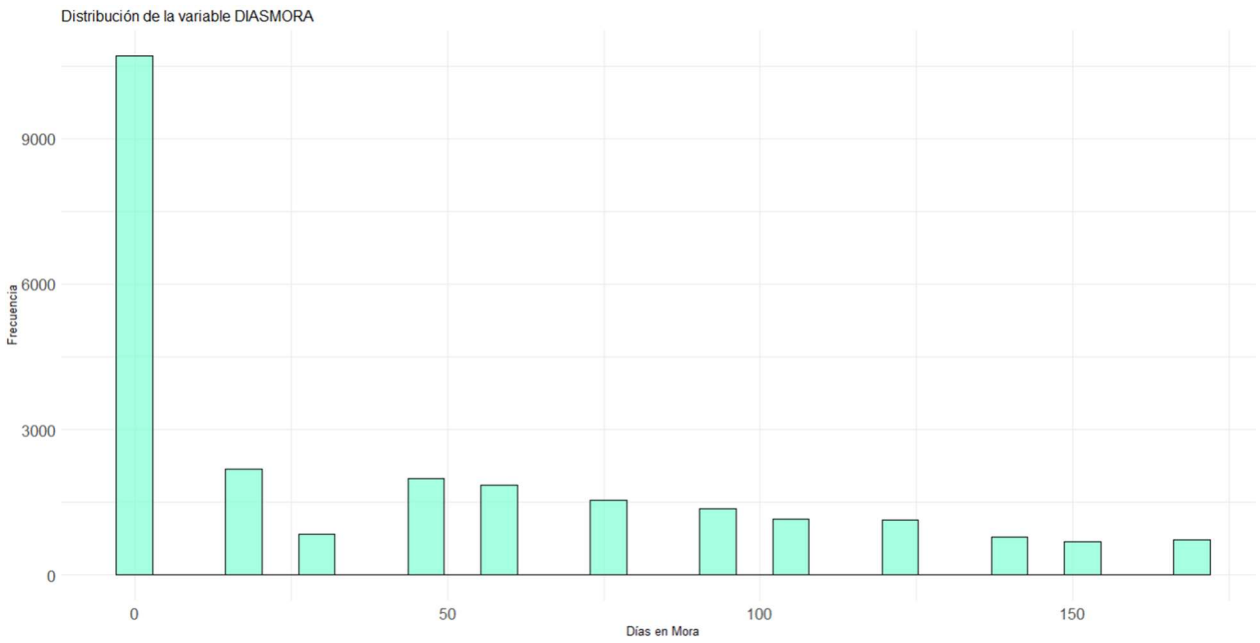

Gráfico No. 5 Distribución Días Mora.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 5 se muestra la distribución de la variable Días en Mora. Se observa que una proporción considerable de clientes (mayor a 9,000 casos) no tiene días de mora, lo que constituye el grupo más frecuente. Le siguen grupos con morosidades de alrededor de 15 a 77 días, conforme lo reflejan también la mediana y el tercer cuartil. La distribución presenta un sesgo positivo, es decir, hay pocos clientes con moras prolongadas, pero estos representan un riesgo importante por la acumulación de deuda y el tiempo transcurrido sin pago.

Esta visualización permite identificar la necesidad de implementar acciones preventivas en los primeros días de atraso, así como estrategias correctivas para los casos crónicos.

Variable Edad:

Tabla 5*:* Variable Edad.

Mínimo	20
Q1	34
Mediana	40
Media	41.17
Q3	47

Máximo	88
Desviación Estándar	10.13

Elaborado por: Gordillo, Víctor (2025).

La Tabla 5 muestra que la edad mínima de los clientes es de 20 años, mientras que la edad máxima alcanza los 88 años, evidenciando una amplia diversidad etaria. La media (41.17 años) y la mediana (40 años) están bastante próximas, lo que sugiere una distribución relativamente simétrica con una ligera inclinación hacia valores mayores. El rango intercuartílico (entre Q1 = 34 y Q3 = 47) indica que el 50% de los clientes se concentra entre estas edades. La desviación estándar de 10.13 años refleja una dispersión moderada respecto a la media.

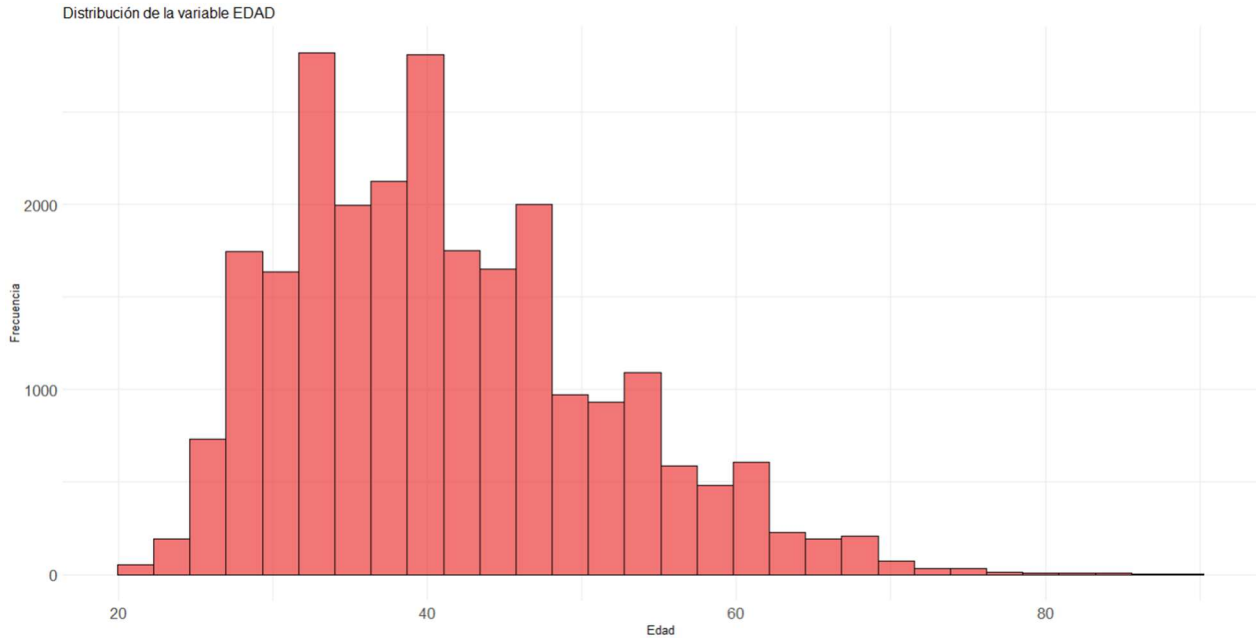

Gráfico No. 6 Distribución Edad.

Elaborado por: Gordillo, Víctor (2025).

Como se aprecia en el Gráfico No. 6, la distribución de la edad de los clientes morosos presenta una mayor concentración entre los 30 y 50 años, siendo este el grupo de mayor presencia dentro de la cartera analizada. Este hallazgo coincide con los valores de mediana y Q1, Q3 observados en la tabla anterior. La forma del histograma presenta una leve asimetría

positiva, mostrando que, aunque hay una alta densidad de clientes en edades intermedias, también existen casos con edades más avanzadas (superiores a 60 años), aunque con menor frecuencia. Esto podría estar asociado a clientes de mayor edad con deudas significativas o con patrones de pago distintos. Este análisis respalda la necesidad de considerar la edad como un factor clave al momento de diseñar estrategias diferenciadas de recuperación de cartera, enfocando esfuerzos en grupos etarios con mayor representatividad o vulnerabilidad.

Variable Cupo:

Tabla 6*:* Variable Cupo.

Mínimo	41.18
Q1	2,300
Mediana	4,200
Media	6,644.84
Q3	8,771.54
Máximo	51,358.24
Desviación Estándar	6,410.41

Elaborado por: Gordillo, Víctor (2025).

La Tabla 6 presenta los estadísticos descriptivos de la variable Cupo, correspondiente al monto disponible para los clientes en sus tarjetas de crédito. Se evidencia que el valor mínimo es de $41.18, mientras que el máximo alcanza los $51,358.24, reflejando una amplia dispersión en los montos asignados. La mediana se ubica en $4,200, lo que indica que el 50% de los clientes tiene cupos iguales o inferiores a este valor. La media, por su parte, es de $6,644.84, superior a la mediana, lo que sugiere la presencia de valores extremos que elevan el promedio. Esto se refuerza con la desviación estándar de $6,410.41, que señala una considerable variabilidad en los datos. Los valores del primer cuartil (Q1 = $2,300) y del tercer cuartil (Q3 = $8,771.54) muestran que el 50% central de los clientes posee cupos entre esos dos valores. Esto confirma una alta concentración de cupos en rangos bajos, mientras que una minoría accede a límites de crédito mucho mayores.

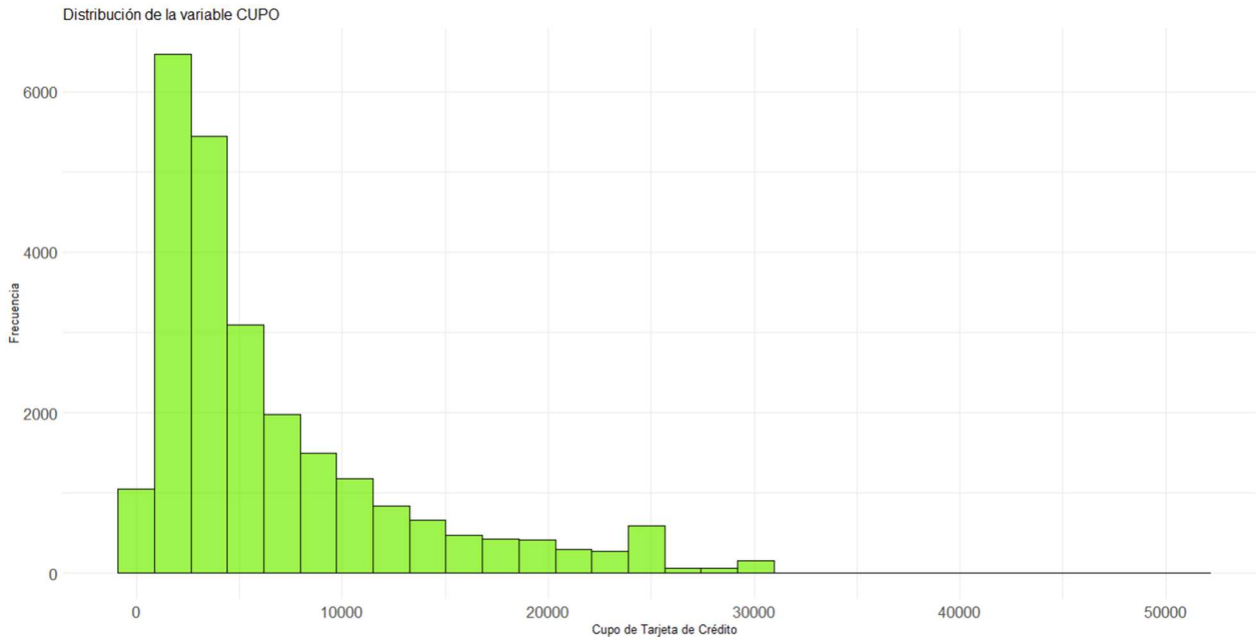

Gráfico No. 7 Distribución Cupo.

Elaborado por: Gordillo, Víctor (2025).

Como se observa en el Gráfico No. 7, la distribución del cupo de tarjeta de crédito entre los clientes muestra un claro sesgo positivo o asimetría hacia la derecha. La mayor parte de los clientes concentra sus cupos en rangos bajos, principalmente entre $0 y $10,000, lo cual se evidencia en las primeras barras del histograma, con frecuencias significativamente mayores. A medida que el valor del cupo aumenta, se observa una disminución progresiva en la frecuencia, lo que indica que solo una pequeña proporción de los clientes accede a montos elevados. Específicamente, se identifican pocos casos que superan los $30,000, lo que corrobora la presencia de valores extremos en la muestra, que afectan la media y aumentan la dispersión general de la variable. Este tipo de distribución es típica en productos financieros de consumo, donde existe una marcada segmentación del perfil crediticio, permitiendo a ciertos clientes acceder a líneas de crédito más elevadas en función de su historial, ingresos o perfil de riesgo. La visualización presentada en el gráfico complementa y refuerza la interpretación de los estadísticos de la Tabla 6, permitiendo una comprensión más clara de la naturaleza y comportamiento de esta variable en la cartera analizada.

Variable Ingreso Mensual:

Tabla 7*:* Variable Ingreso Mensual.

Mínimo	0
Q1	605
Mediana	1,000
Media	1,407.51
Q3	1,610
Máximo	7,000
Desviación Estándar	1,295.66

Elaborado por: Gordillo, Víctor (2025).

La Tabla 7 muestra un ingreso mensual mínimo de $0 y un ingreso máximo de $7,000. El primer cuartil (Q1) se encuentra en $605, lo que indica que el 25% de los clientes tienen ingresos iguales o inferiores a ese valor. La mediana es de $1,000, es decir, la mitad de los clientes ganan menos de esa cantidad. La media es de $1,407.51, ligeramente mayor a la mediana, lo que sugiere una ligera asimetría a la derecha. El tercer cuartil (Q3) se encuentra en $1,610, por lo que el 75% de los clientes ganan hasta ese valor. La desviación estándar es de $1,295.66, lo cual refleja una alta dispersión en los ingresos mensuales.

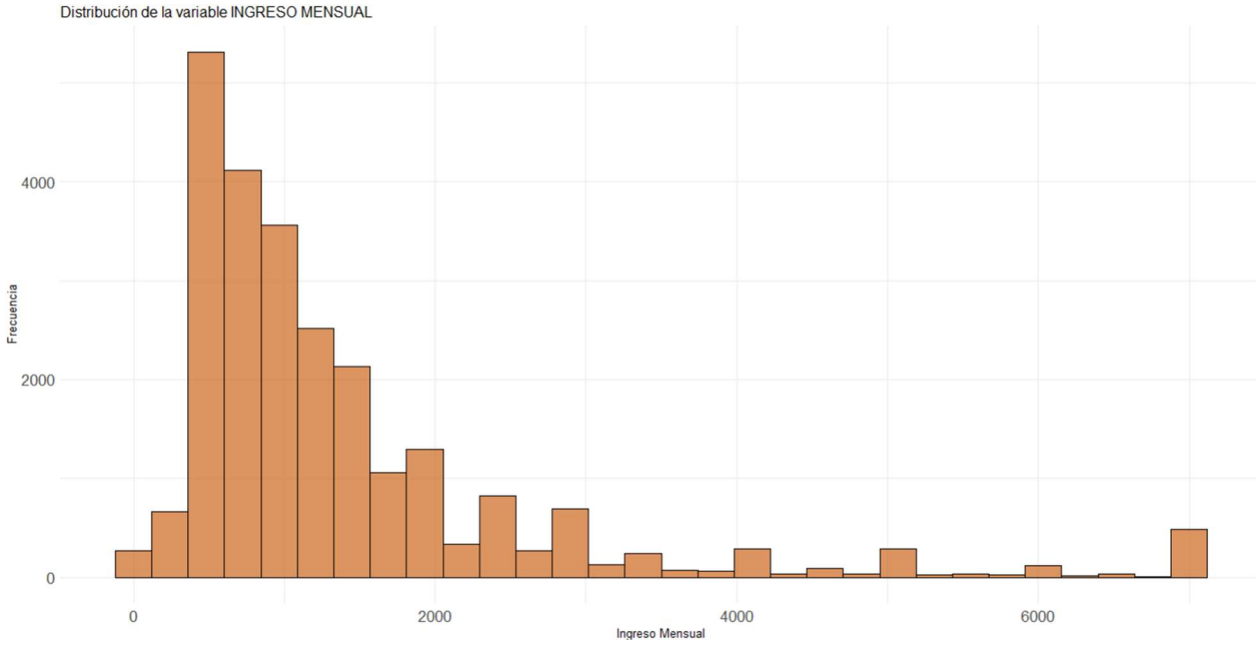

Gráfico No. 8 Distribución Ingreso Mensual.

Elaborado por: Gordillo, Víctor (2025).

Como se observa en el Gráfico No. 8, la distribución del ingreso mensual presenta una marcada asimetría positiva (sesgada hacia la derecha). Existe una alta concentración de clientes con ingresos por debajo de los $2,000, particularmente en el rango entre $0 y $1,000. A medida que aumentan los ingresos, la frecuencia de clientes disminuye de manera significativa. Se identifica una larga cola derecha en la distribución, lo que indica que solo una pequeña proporción de clientes alcanza ingresos mensuales elevados, por encima de los $4,000. Este comportamiento es típico en poblaciones con desigualdad de ingresos, donde los valores atípicos influyen sobre la media y la dispersión general de los datos.

Variable Pagos Puntuales:

Tabla 8*:* Variable Pagos Puntuales.

Mínimo	0
Q1	5
Mediana	7
Media	6.62
Q3	8
Máximo	11
Desviación Estándar	2.32

Elaborado por: Gordillo, Víctor (2025).

Según la Tabla 8, el número de pagos puntuales realizados por los clientes en los últimos 12 meses presenta una media de 6.62 y una mediana de 7, lo que indica un cumplimiento moderado en los pagos. El valor mínimo registrado es de 0 pagos, mientras que el máximo llega hasta 11 pagos puntuales. Los cuartiles muestran que el 25% de los clientes realizaron 5 o menos pagos puntuales (Q1), el 50% realizó 7 o menos pagos (mediana), y el 75% no superó los 8 pagos (Q3), con una desviación estándar de 2.32, lo que refleja una variabilidad moderada en la conducta de pago regular.

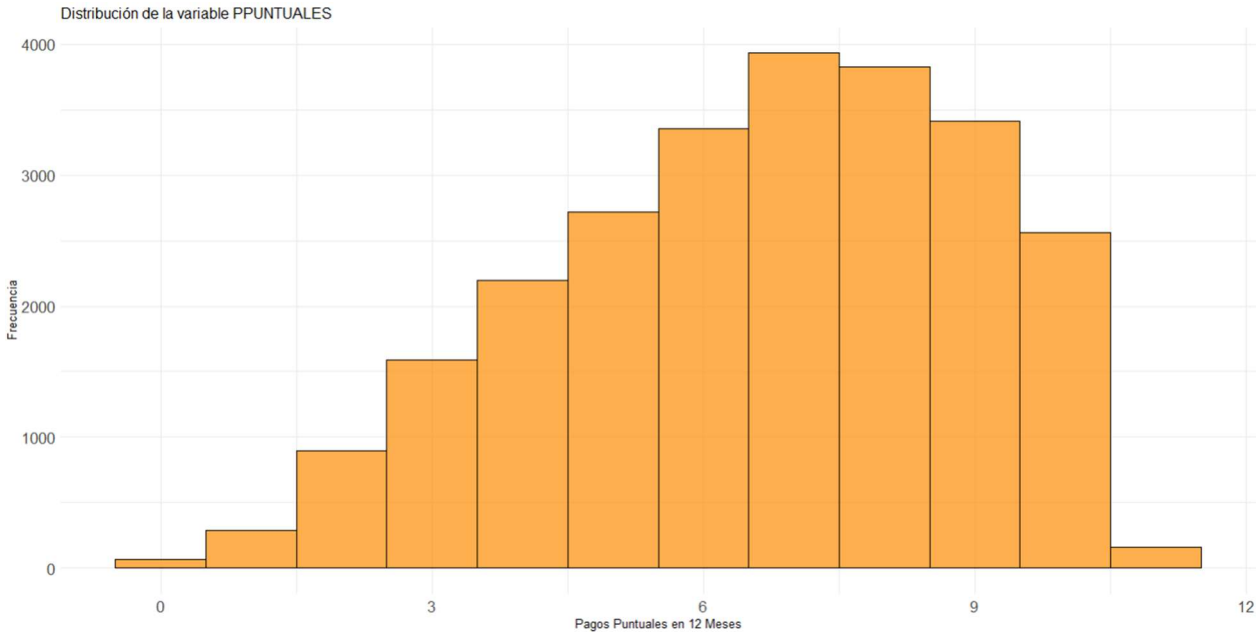

Gráfico No. 9 Distribución Pagos Puntuales.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 9, la distribución de los pagos puntuales es aproximadamente simétrica, con mayor concentración de clientes que realizaron entre 6 y 8 pagos durante el periodo analizado. Esto sugiere un comportamiento intermedio de cumplimiento, donde la mayoría no alcanza el total de 12 pagos (uno por mes), pero tampoco incumple completamente. La baja frecuencia de clientes con 0 o 11 pagos indica que los extremos son menos comunes, reforzando la idea de un patrón de cumplimiento parcial y constante dentro del grupo analizado.

Variable Pagos Retrasados:

Tabla 9*:* Variable Pagos Retrasados.

Mínimo	1
Q1	4
Mediana	5
Media	5.38
Q3	7
Máximo	12
Desviación Estándar	2.32

Elaborado por: Gordillo, Víctor (2025).

La Tabla 9 presenta un resumen estadístico de la variable Pagos Retrasados, evidenciando un comportamiento relativamente homogéneo dentro de la muestra analizada. El valor mínimo observado es 1 pago retrasado y el máximo alcanza los 12 pagos, lo que corresponde al número total de meses del año. La mediana se sitúa en 5 pagos retrasados, mientras que la media es ligeramente menor, con 5.38, lo que sugiere una distribución ligeramente sesgada a la derecha. Los valores del primer cuartil (Q1) y tercer cuartil (Q3) son 4 y 7 respectivamente, indicando que el 50% de los clientes se concentra entre 4 y 7 pagos morosos. La desviación estándar de 2.32 refuerza que existe cierta dispersión en el comportamiento, aunque contenida dentro de rangos moderados.

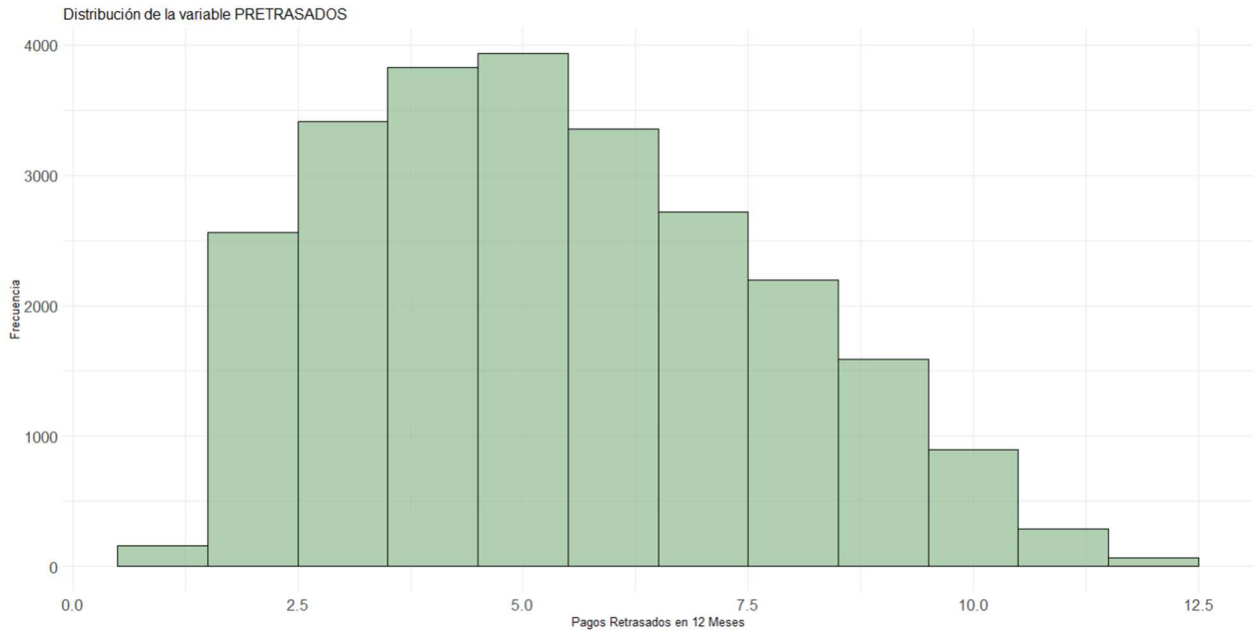

Gráfico No. 10 Distribución Pagos Retrasados.

Elaborado por: Gordillo, Víctor (2025).

El Gráfico No. 10, correspondiente a la distribución de la variable Pagos Retrasados, muestra una forma ligeramente asimétrica hacia la derecha. Se observa que la mayor densidad de clientes se encuentra entre los 4 y 6 pagos atrasados, lo cual concuerda con los valores de la

mediana y cuartiles reportados en la tabla. Esta distribución sugiere que una proporción significativa de los clientes presenta una morosidad leve a moderada, ya que más del 75% de los registros se mantiene dentro del rango de 1 a 7 pagos morosos en los últimos 12 meses. Esta información resulta crítica para la segmentación del riesgo crediticio y la definición de estrategias diferenciadas de cobranza.

Variable Género:

Tabla 10*:* Variable Género.

Género	Frecuencia	Proporción Porcentaje
Femenino	10,656	42.62%
Masculino	14,348	57.38%

Elaborado por: Gordillo, Víctor (2025).

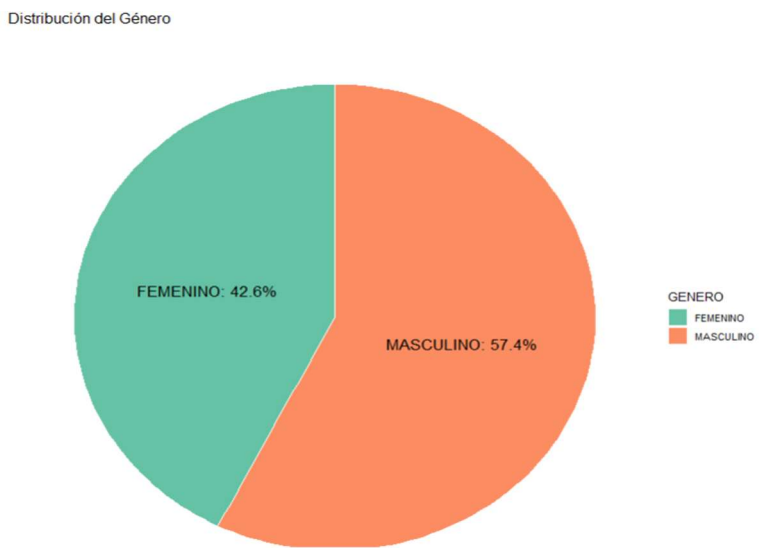

Gráfico No. 11 Distribución Género.

Elaborado por: Gordillo, Víctor (2025).

En la Tabla 10, se observa que el género masculino representa la mayor proporción de los clientes con tarjetas de crédito en mora, con un 57.38% (14.348 casos), mientras que el 42.62% (10.656 casos) corresponde al género femenino. Esta distribución indica una ligera predominancia masculina en la muestra analizada. Por su parte, en el Gráfico No. 11, se visualiza de forma clara esta proporción mediante un gráfico circular que refuerza la mayor

presencia del género masculino entre los clientes morosos. Si bien la diferencia no es extremadamente amplia, resulta relevante en el contexto del análisis crediticio, ya que podría estar vinculada a patrones de comportamiento financiero diferenciados por género, lo cual podría considerarse en estrategias de segmentación y cobranza.

Variable Estado Civil:

Tabla 11*:* Variable Estado Civil.

Estado Civil	Frecuencia	Proporción Porcentaje
Casado	11,885	47.54%
Divorciado	2,914	11.65%
Soltero	9,647	38.58%
Unión Libre	314	1.26%
Viudo	244	0.98%

Elaborado por: Gordillo, Víctor (2025).

La Tabla 11 presenta la distribución de los estados civiles entre los clientes con tarjetas de crédito en mora. Se observa que el grupo mayoritario corresponde a personas casadas (47.54%), seguido de solteros (38.58%). En menor proporción se encuentran los divorciados (11.65%), y con valores significativamente bajos los estados de unión libre (1.26%) y viudos (0.98%). Esta distribución sugiere que el estado civil predominante entre los clientes en situación de mora es el casado, lo que podría estar asociado con compromisos financieros familiares o responsabilidades compartidas.

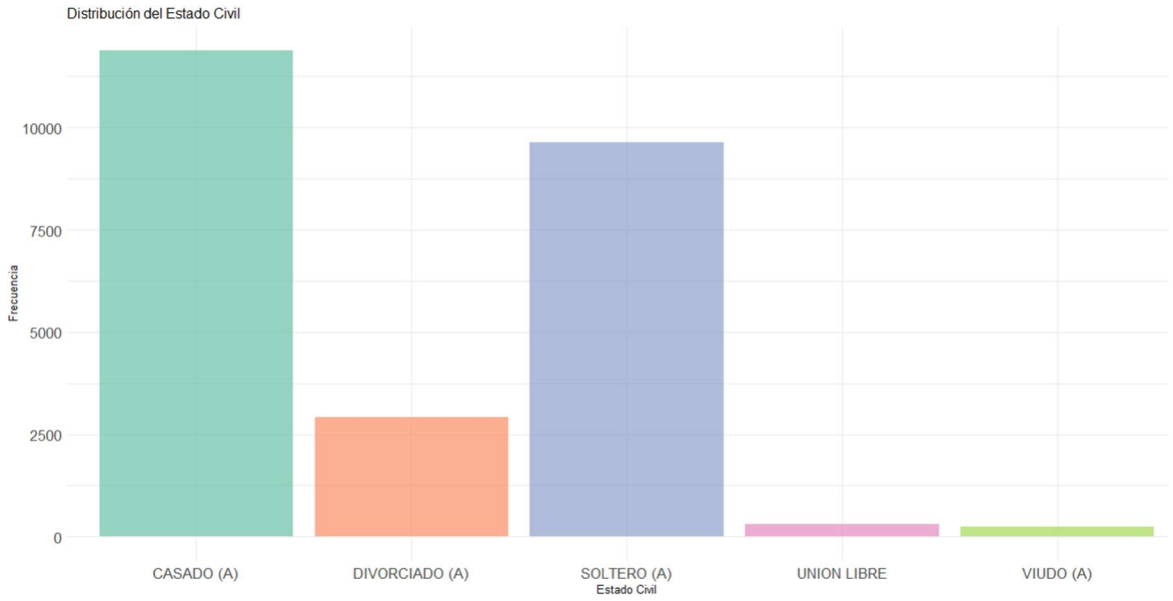

Gráfico No. 12 Distribución Estado Civil.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 12, se visualiza esta distribución mediante un gráfico de barras que confirma la predominancia de los grupos casado y soltero. Estos dos grupos superan ampliamente a las demás categorías. Las barras correspondientes a unión libre y viudo son notablemente más bajas, lo cual indica que representan una pequeña fracción de la población total analizada. Este patrón podría ser relevante al momento de segmentar a los clientes para el diseño de estrategias de cobranza diferenciadas por estado civil.

CAPÍTULO III

DESARROLLO DE LA INVESTIGACIÓN

Aplicación del algoritmo K-means

Una vez que los datos han sido limpiados y preprocesados, se procede a implementar el algoritmo K-means con el objetivo de segmentar a los clientes.

1. **Determinación del número óptimo de clusters:**

   Para identificar el número óptimo de clusters se emplearon dos métodos, el método del codo (Elbow Method) y el coeficiente de Silhouette.

   Se evaluaron diferentes valores de k utilizando:

   - Método del codo, que sugirió $k = 3$ como un número adecuado.

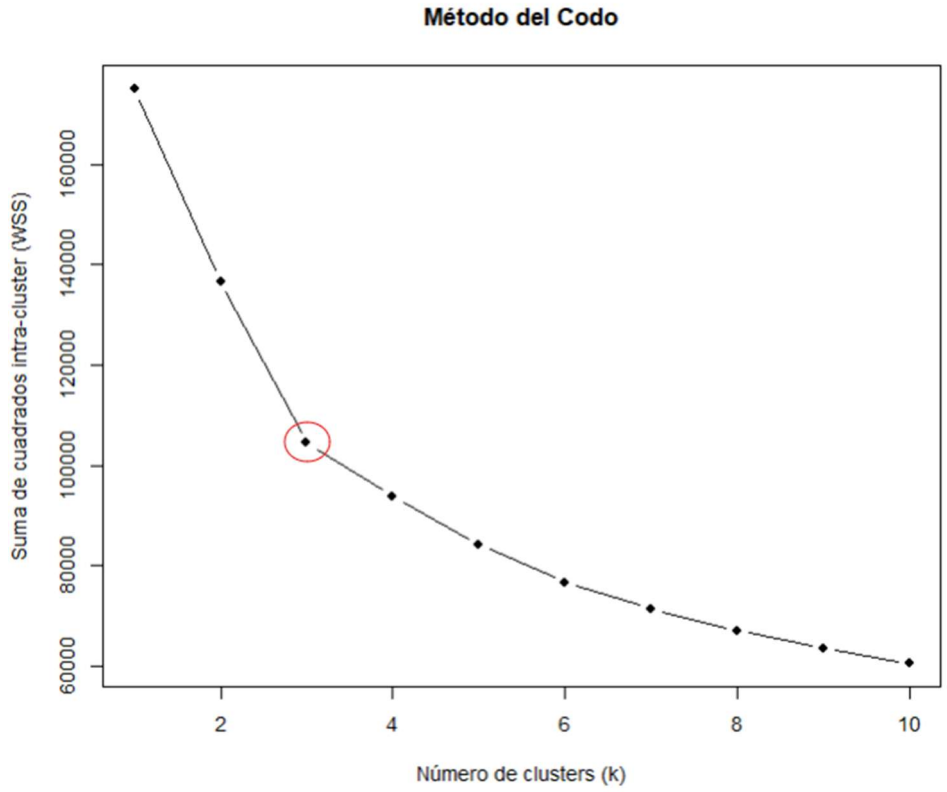

Gráfico No. 13 Número óptimo de cluster.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 13, se representa la variación de la suma de cuadrados intra-cluster frente al número de clusters. Se observa que a partir de k = 3, la reducción de la varianza se vuelve marginal, evidenciando el punto de inflexión del "codo", lo que sugiere que tres agrupaciones logran una adecuada segmentación sin incurrir en sobreajuste ni pérdida de información. Esta elección se respalda también con el valor del coeficiente de Silhouette correspondiente a k=3, que mostró un nivel aceptable, lo cual indica una separación razonable entre los clusters formados. En conjunto, ambos métodos coinciden en que k = 3 es el valor óptimo, asegurando cohesión intra-cluster y una buena diferenciación entre grupos.

2. **Implementación del algoritmo K-means:**
   El algoritmo se aplica utilizando el paquete cluster en R, asignando a cada cliente un cluster en función de la similitud de sus características.
3. **Iteración y ajuste del modelo:**
   Se realizan múltiples iteraciones para asegurar la convergencia del algoritmo y la estabilidad de los clusters formados.

Análisis de Datos

La calidad de los clusters formados mediante el algoritmo K-means será evaluada utilizando diversas métricas estadísticas que permiten medir la coherencia interna de los grupos y la separación entre ellos.

Métricas de Evaluación utilizadas

1. **Inercia (Suma de Errores Cuadrados - SSE):**
   Esta métrica calcula la variación interna de cada cluster, la cual corresponde a la suma de las distancias al cuadrado entre cada punto y el centroide del grupo al que pertenece. Valores más bajos de inercia indican clusters más compactos y con mejor definición.
2. **Coeficiente de Silhouette:**

Mide la calidad de la segmentación verificando el grado de similitud con otros clusters. Sus valores oscilan entre -1 y 1, siendo cercanos a 1 indicativos de una clara separación entre grupos.

3. **Método del codo (Elbow Method)**

   Consiste en representar gráficamente la inercia en función del número de clusters y seleccionar el punto en que la reducción de la inercia empieza a ser mínima, lo que indica el número óptimo de grupos.

Interpretación de Resultados

Una vez evaluada la calidad de los clusters, se procede a interpretar las características de cada segmento de clientes. Esta interpretación permite identificar patrones de comportamiento que facilita la implementación de nuevas políticas o estrategias de cobranza personalizadas, adaptadas a las características específicas de cada grupo de clientes, lo que permite mejorar la eficiencia en la recuperación de deudas.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

En este apartado se exponen los hallazgos obtenidos a partir de la implementación del algoritmo K-means, incluyendo la visualización de los clusters formados, la interpretación de las características principales de cada grupo y la comparación con estudios previos relevantes. Estos resultados permiten identificar patrones de comportamiento en clientes morosos y proporcionan información valiosa para diseñar estrategias de cobranza personalizadas.

Resultados del Clustering:

Los clientes fueron segmentados en tres clusters principales, diferenciados principalmente por sus características financieras y comportamiento de pago.

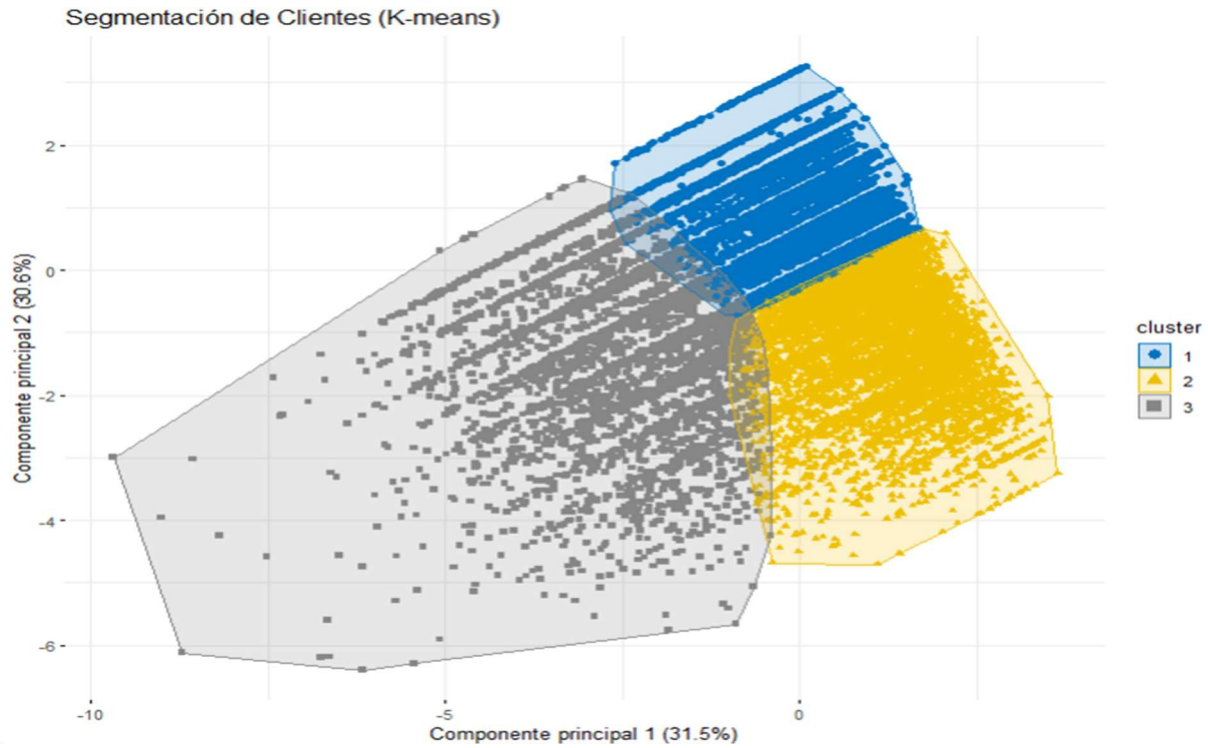

Gráfico No. 14 Segmentación de Clientes (K-means).

Elaborado por: Gordillo, Víctor (2025).

El Gráfico No.14 muestra la segmentación proyectada en dos componentes principales (PCA), permitiendo visualizar la separación entre grupos:

- El Cluster 1 (color azul) agrupa a una gran cantidad de clientes con características intermedias.
- El Cluster 2 (color amarillo) concentra a clientes con mejor comportamiento de pago.
- El Cluster 3 (color gris) representa a un grupo diferenciado por condiciones financieras más específicas, posiblemente asociadas a mayor riesgo de mora o características particulares de uso del crédito.

Esta segmentación evidencia una estructura clara en los datos, que permite entender mejor los perfiles de clientes y ofrece una base sólida para diseñar estrategias diferenciadas de cobranza o retención.

La siguiente tabla muestra un resumen de las características principales de cada clúster:

Tabla 12*:* Segmentación Clusters.

Cluster	Monto Prom. (USD)	Ingreso Mensual Prom. (USD)	Días en Mora Prom.	Edad Prom. (años)	Cupo Prom. (USD)	Pagos Puntuales Prom.	Pagos Retrasados Prom.	N° de Casos
1	17,256	2,462	36.8	46.6	18,975	6.78	5.22	3,796
2	3,630	1,197	24.9	40.1	4,412	8.17	3.83	12,573
3	3,831	1,251	74.0	40.3	4,476	4.29	7.71	8,635

Elaborado por: Gordillo, Víctor (2025).

Análisis por grupo:

1. **Cluster 1: Clientes de alto valor con Mora Moderada**
    - Este grupo se caracteriza por clientes con ingresos y cupos altos con una deuda significativa.
    - Presentan una mora intermedia (36.8 días) y un comportamiento de pago relativamente equilibrado.
    - Son clientes valiosos para la institución financiera y pueden beneficiarse de una reestructuración personalizada o planes de fidelización, ya que una recuperación oportuna puede garantizar su permanencia.
2. **Cluster 2: Clientes con buen comportamiento de pago**
    - Contiene el mayor número de clientes.
    - Tienen los niveles más bajos de deuda y mora (24.9 días), y el promedio más alto de pagos puntuales.
    - Este segmento presenta bajo riesgo crediticio y un perfil confiable, adecuado para mantener con esquemas tradicionales de seguimiento y reforzar su buen historial.
3. **Cluster 3: Clientes con riesgo alto de incobrabilidad**
    - Se diferencia por tener los mayores días en mora (74) y menos pagos puntuales.
    - Aunque los ingresos y el monto de deuda son similares al Cluster 2, su historial de pagos muestra un patrón de riesgo considerable.
    - Se recomienda aplicar estrategias de cobranza intensiva y preventiva, así como revisar su elegibilidad para productos financieros futuros.

Visualización de Clusters

Para comprender mejor la segmentación, se generaron gráficos que ilustran la distribución de las principales variables en los clusters identificados.

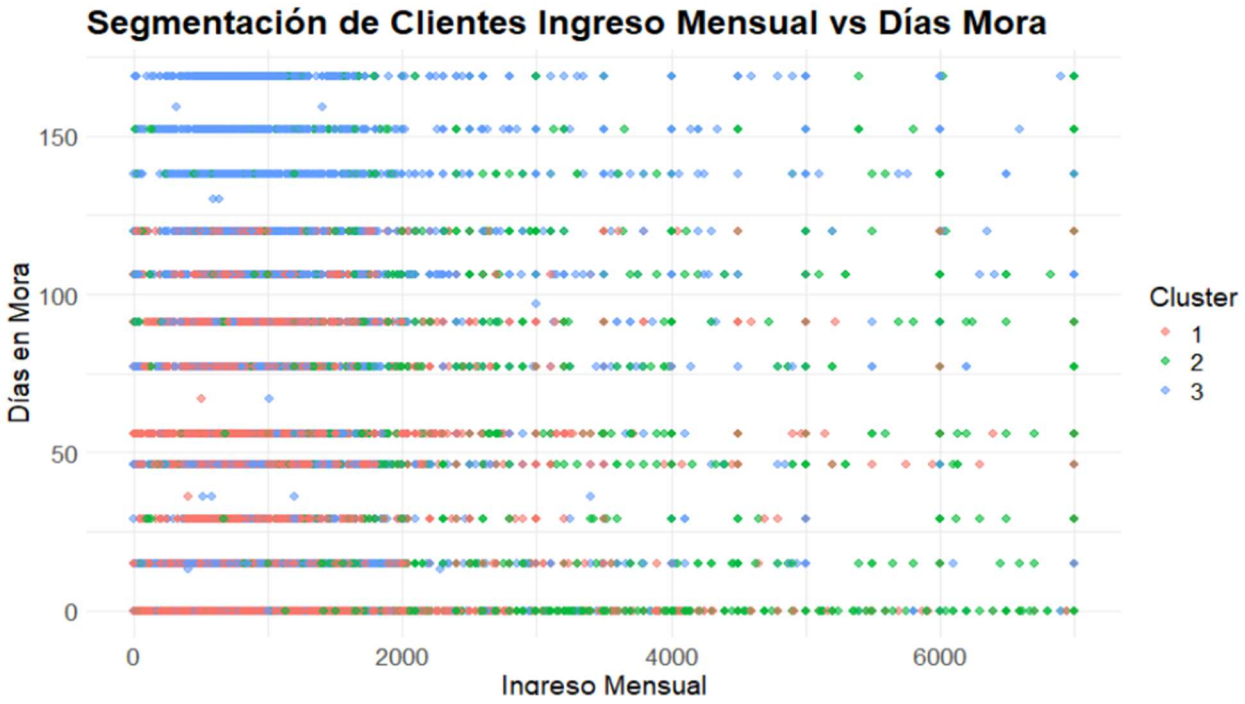

Gráfico No. 15 Segmentación de Clientes Ingreso Mensual vs Días Mora.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 15 se muestra la relación entre ingreso mensual y días en mora para cluster, donde el Cluster 2 concentra clientes con ingresos moderados y menor dispersión en mora, indicando bajo riesgo crediticio, el Cluster 3 muestra mayor dispersión vertical con clientes que, a pesar de tener ingresos similares, presentan moras prolongadas, sugiriendo mayor riesgo, por otro lado, el Cluster 1 presenta ingresos más elevados pero con mora intermedia, lo que sugiere un segmento con oportunidad de gestión personalizada y estrategias diferenciadas de cobranza.

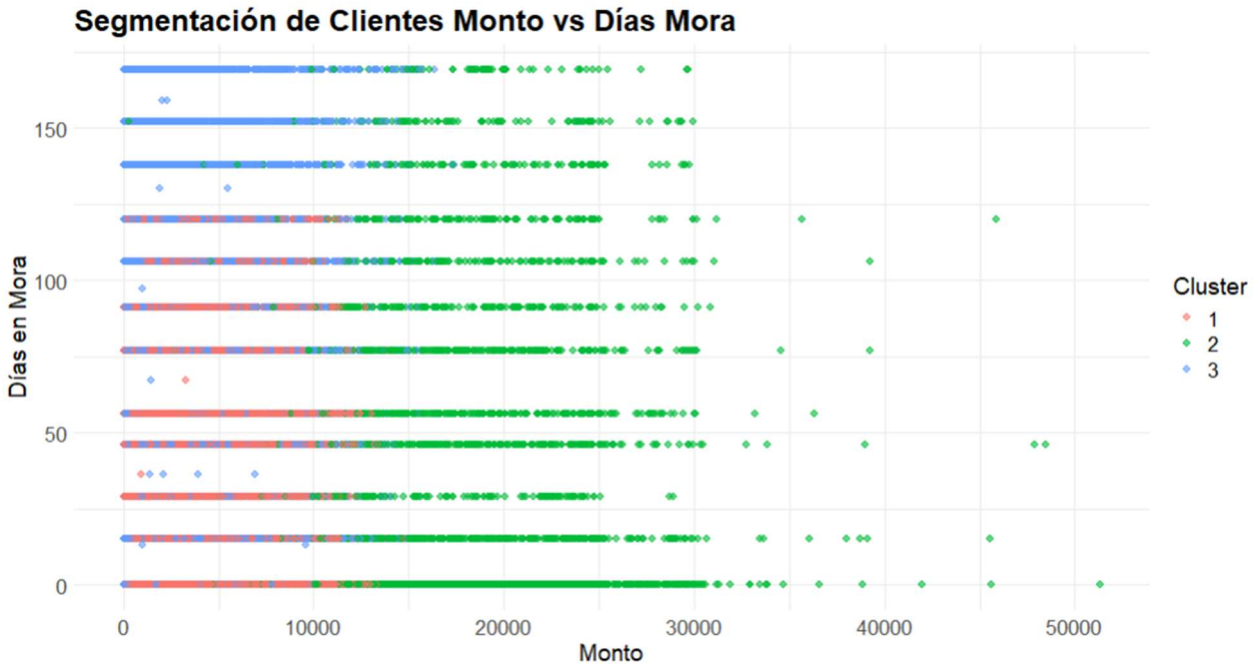

Gráfico No. 16 Segmentación de Clientes Monto vs Días Mora.
Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 16 se observa cómo se distribuyen los clusters en función del monto de la deuda y los días en mora, el Cluster 2 destaca por mayor concentración en montos intermedios y menor dispersión vertical, indicando menor riesgo de mora, mientras el Cluster 3 presenta alta dispersión en días mora, reflejando clientes con atrasos prolongados aunque montos de deuda similares, y el Cluster 1 se ubica con montos de deuda más altos y mora moderada, sugiriendo la necesidad de estrategias diferenciadas de cobranza para cada segmento.

Para analizar la distribución de los datos dentro de cada cluster, se generaron histogramas de las siguientes variables:

- **Monto de la deuda:** Indica la distribución del saldo de los clientes.
- **Edad:** Permite entender el perfil demográfico de los clientes.
- **Pagos puntuales:** Refleja la tendencia de los clientes a cumplir con sus obligaciones financieras.

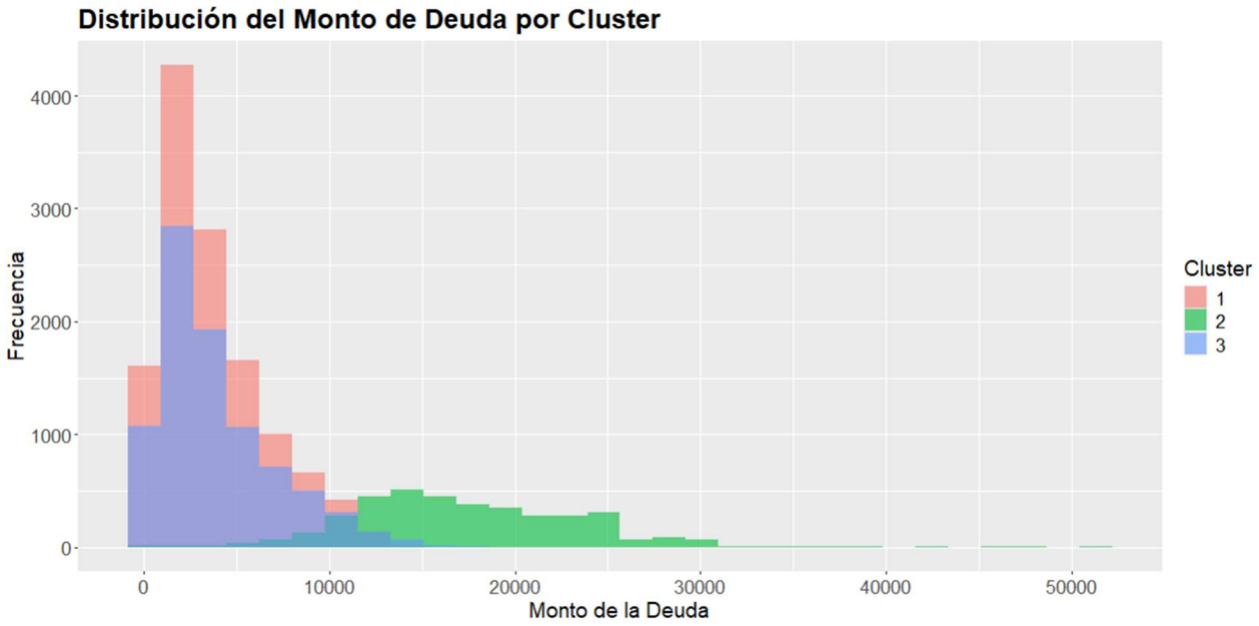

Gráfico No. 17 Distribución del Monto por Cluster.

Elaborado por: Gordillo, Víctor (2025).

El Gráfico No.17 evidencia diferencias claras entre los clusters formados. El Cluster 1 (color rojo) agrupa principalmente a clientes con deudas bajas, concentradas por debajo de los $5,000, lo que sugiere un perfil de riesgo relativamente menor o una menor exposición crediticia. En contraste, el Cluster 2 (color verde) muestra una distribución más dispersa y extensa hacia montos elevados, superando los $30,000 en algunos casos, lo que podría reflejar clientes con un alto nivel de endeudamiento o con líneas de crédito más amplias. Por su parte, el Cluster 3 (color azul) se caracteriza por una distribución intermedia, aunque también muestra valores considerables hacia la derecha, indicando la presencia de algunos clientes con deudas elevadas. Esta diferenciación en los montos refleja el grado de heterogeneidad financiera entre los segmentos y justifica la segmentación realizada mediante K-means.

Gráfico No. 18 Distribución de Edad por Cluster.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 18, se observa la distribución etaria de los clientes segmentados por clúster. El Cluster 2 (color azul) concentra una proporción significativa de clientes entre los 35 y 55 años, reflejando un perfil de edad intermedia. El Cluster 1 (color rojo) muestra una alta concentración de individuos jóvenes, particularmente entre los 25 y 40 años, lo que podría asociarse a un menor nivel de estabilidad financiera o menor experiencia crediticia. Por su parte, el Cluster 3 (color verde) presenta una distribución más homogénea y extendida, incluyendo tanto jóvenes como adultos mayores, aunque con menor frecuencia. Esta segmentación etaria permite entender el perfil demográfico de los clientes en mora y su posible relación con los hábitos de pago y niveles de endeudamiento, aportando elementos relevantes para el diseño de estrategias de cobranza diferenciadas por edad.

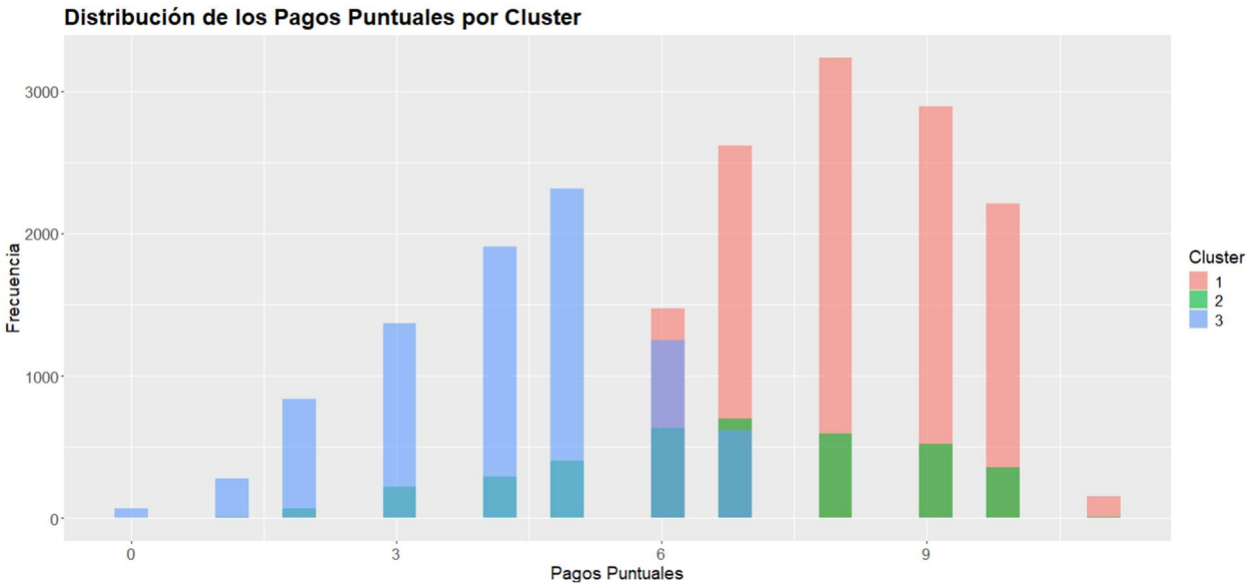

Gráfico No. 19 Distribución de Pagos Puntuales por Cluster.

Elaborado por: Gordillo, Víctor (2025).

En el Gráfico No. 19 el Cluster 1 (rojo) se destaca por tener la mayor proporción de clientes cumplidos, con entre 8 y 10 pagos realizados puntualmente. El Cluster 2 (azul) muestra una distribución más heterogénea, con frecuencias distribuidas entre los rangos de 3 a 6 pagos. Por su parte, el Cluster 3 (verde) evidencia bajos niveles de cumplimiento, con frecuencias elevadas en los pagos entre 0 y 4, lo que representa una señal de mayor riesgo de morosidad.

Los siguientes diagramas de caja presentan variabilidad dentro de cada cluster para las variables **Monto de deuda, días en mora.**

Monto de deuda:

En la siguiente gráfica el cluster 2 presenta montos significativamente más altos y una mayor dispersión, indicando clientes con deudas más elevadas y variabilidad considerable, mientras que los Clusters 1 y 3 muestran montos menores y rango intercuartílicos más estrechos, sugiriendo perfiles más homogéneos que podrían facilitar estrategias de cobranza diferenciadas según el nivel de deuda y el riesgo asociado.

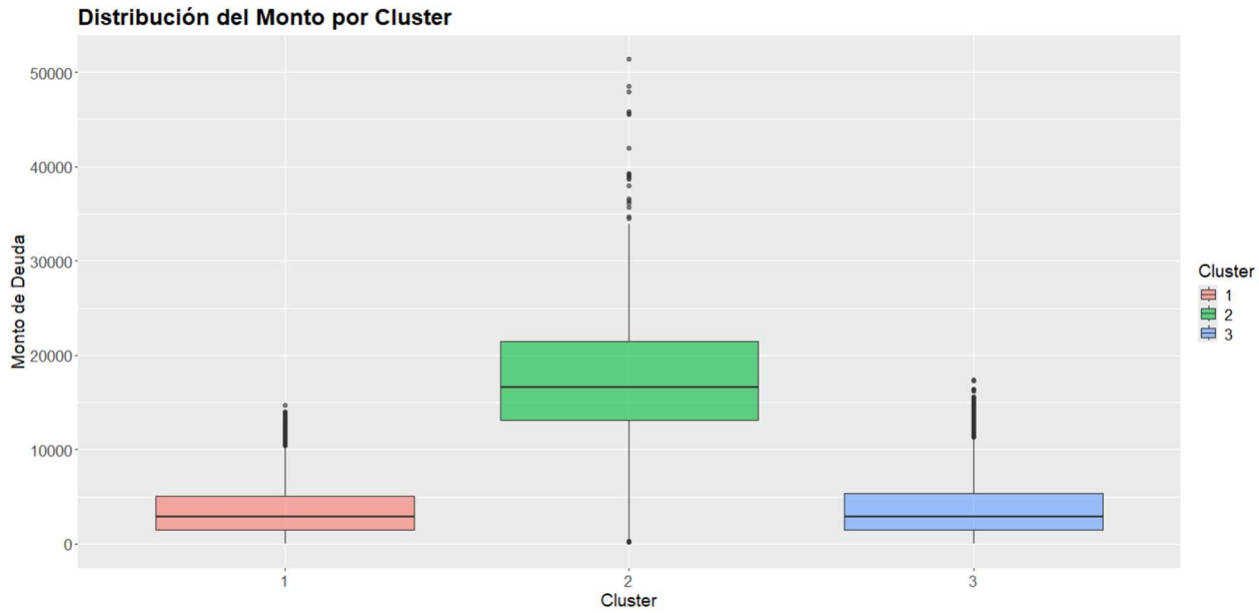

Gráfico No. 20 BoxPlot Monto por Cluster.

Elaborado por: Gordillo, Víctor (2025).

Días en mora:

En la siguiente gráfica se evidencia diferencias claras entre los clusters, el Cluster 1 presenta una distribución más concentrada y menor mediana, lo que indica clientes con moras más controladas, el Cluster 2 tiene dispersión moderada y algunos valores atípicos, sugiriendo un riesgo intermedio, mientras que el Cluster 3 exhibe la mayor dispersión y una mediana

elevada, evidenciando clientes con moras más prolongadas y mayor riesgo de incobrabilidad, lo que justifica estrategias de cobranza más intensivas para este segmento.

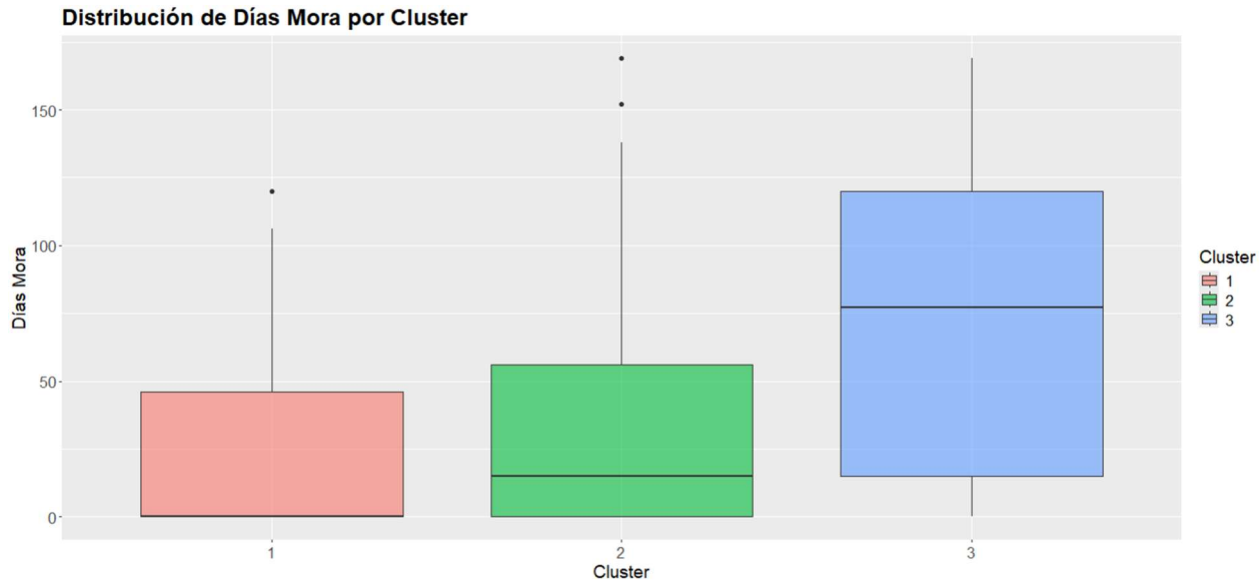

Gráfico No. 21 BoxPlot Días Mora por Cluster.
Elaborado por: Gordillo, Víctor (2025).

Discusión de Resultados

Evaluación de la segmentación y métricas de validación

Para evaluar la calidad del modelo de segmentación basado en K-means, se utilizaron las siguientes métricas:

- **Inercia:** Este valor representa la compacidad de los grupos formados, un valor menor indica una mejor agrupación. En el análisis realizado, el modelo obtuvo un valor de 104,482, lo que evidencia que la segmentación fue adecuada, considerando la variabilidad propia de datos financieros. Según Jain (2010), la inercia es una métrica fundamental para evaluar la compacidad en modelos de clustering como K-means, siendo ampliamente utilizada para seleccionar el número óptimo de grupos mediante el método del codo.

- **Coeficiente de Silhouette:** Mide la separabilidad de los clusters, cuando los valores son cercanos a 1 indican una mejor segmentación. El coeficiente promedio obtenido fue de 0.26, indicando que los clusters tienen una separación aceptable. Según Kaufman y Rousseeuw (2005), un valor superior a 0.25 puede considerarse como una segmentación útil en contexto reales con datos heterogéneos.

  Si bien el coeficiente de Silhouette de 0.26 supera el umbral mínimo considerado útil en contextos reales, su valor relativamente bajo sugiere un cierto grado de solapamiento entre los clusters, lo cual puede deberse a similitudes en los perfiles financieros de los clientes o la falta de variables adicionales que ayuden a discriminar mejor los grupos. Este hallazgo plantea una limitación en la capacidad de separación del modelo y debe ser considerado en la implementación de estrategias operativas basadas en esta segmentación.

  Tabla 13*:* Evaluación de la calidad de la segmentación.

Número de Cluster	Inercia	Coeficiente Silhouette
3	104,482	0.26
4	93,811.96	0.23
5	84,175.14	0.24
6	76,592.7	0.22
7	71,347.82	0.22

  Elaborado por: Gordillo, Víctor (2025).

La Tabla 13 presenta los valores de inercia y del coeficiente de Silhouette obtenidos al aplicar el algoritmo K-means con diferentes cantidades de clusters, con el fin de evaluar la calidad de la segmentación. Se observa que al aumentar el número de clusters de 3 a 7, la inercia disminuye progresivamente, lo cual es esperado, ya que la partición mejora al aumentar la cantidad de grupos, reduciendo así la varianza intra-cluster. Sin embargo, esta reducción

presenta rendimientos decrecientes, siendo más pronunciada entre 3 y 5 clusters y menos significativa a partir de 6. Por otro lado, el coeficiente de Silhouette, que mide tanto la cohesión como la separación de los clusters, alcanza su valor máximo (0.26) cuando el número de clusters es 3. A partir de este punto, el coeficiente disminuye, lo cual indica que una mayor cantidad de grupos no mejora la separación entre los mismos, sino que la degrada ligeramente. En conjunto, estos resultados respaldan la elección de k = 3 clusters como el número óptimo para esta segmentación, ya que proporciona un equilibrio adecuado entre compacidad (baja inercia) y separación (alto coeficiente de Silhouette).

En comparación con estudios similares realizados en contextos latinoamericanos, como el de de Souza y Bodin Jr. (2021) en Brasil o el de Zapata-Rodríguez et al. (2023) en Colombia, los resultados obtenidos son consistentes en cuanto a la dificultad de lograr una alta separación entre grupos cuando se trata de segmentar clientes morosos. Estos estudios también reportaron coeficientes de Silhouette bajos a moderados y resaltaron la utilidad práctica de los clusters a pesar de dicha limitación. Este trabajo aporta evidencia local valiosa al aplicar técnicas de aprendizaje no supervisado en datos reales de una institución financiera ecuatoriana, contribuyendo así al escaso cuerpo de literatura sobre gestión de morosidad con machine learning en Ecuador.

Limitaciones del estudio

Esta investigación se desarrolló utilizando datos correspondientes únicamente al periodo enero-diciembre 2024, por lo cual no se consideraron efectos estacionales o variaciones interanuales en el comportamiento de morosidad. Además, la base de datos empleada está compuesta exclusivamente por variables internas de la institución financiera, sin incluir factores macroeconómicos relevantes como tasas de interés, inflación o desempleo, los cuales pueden tener un efecto significativo sobre la capacidad de pago de los clientes. Estas limitaciones abren espacio para investigaciones futuras que integren datos externos y análisis longitudinales.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

1. La segmentación con K-means permitió identificar a los clientes en grupos homogéneos que permiten identificar patrones de pago y riesgo, alineándose con el objetivo de optimizar estrategias de cobranza y reducir riesgo crediticio.
2. Se identificaron perfiles diferenciados: Cluster 2 como clientes confiables con bajo riesgo, Cluster 1 como clientes de alto valor con mora moderada y Cluster 3 como segmento crítico de alto riesgo e incobrabilidad.
3. Las métricas de inercia y silhouette indican una segmentación aceptable para fines estratégicos.
4. Los resultados obtenidos respaldan la hipótesis planteada, evidenciando que el uso del algoritmo K-means permite mejorar de forma notable la eficiencia de las estrategias de cobranza, al facilitar la segmentación de clientes con base a patrones de riesgo y comportamiento de pago.

Recomendaciones:

1. Estrategias específicas por cluster:
    - Cluster 1 – Clientes de alto valor con mora moderada:
      Se recomienda aplicar esquemas personalizados de reestructuración de deuda, enfocados en mantener la relación comercial con estos clientes valiosos. Dado su nivel de ingresos y cupo elevado, podrían ofrecerse incentivos como reducción de tasas, periodos de gracia o refinanciamiento flexible, priorizando su fidelización y evitando su migración hacia la morosidad crítica.
    - Cluster 2 – Clientes con buen comportamiento de pago:
      Este grupo presenta el mejor perfil de riesgo, por lo que se sugiere continuar con políticas de mantenimiento del buen historial crediticio, tales como bonificaciones por pago puntual, aumentos progresivos de cupo y acceso a productos financieros preferenciales.
    - Cluster 3 – Clientes con riesgo alto de incobrabilidad:

Dado su elevado número de días en mora y bajo promedio de pagos puntuales, es necesario implementar estrategias de cobranza intensiva, priorizando canales digitales automatizados (SMS, correo, WhatsApp) para optimizar costos y tiempos. Además, se recomienda establecer alertas tempranas, bloqueos preventivos, e incluso políticas de no renovación de crédito según el historial.

2. Se propone como investigación futura el desarrollo de modelos híbridos, que integren la segmentación no supervisada (clustering) con modelos supervisados de clasificación, como árboles de decisión, XGBoost o regresión logística, que permitan no solo agrupar perfiles, sino también predecir la probabilidad de incumplimiento y actuar proactivamente con base en dicha predicción.
3. Se recomienda replicar el análisis utilizando otros algoritmos como DBSCAN, GMM o clustering jerárquico, que podrían detectar relaciones no lineales entre variables y grupos con formas más complejas que K-means no capta eficientemente.
4. Se aconseja incluir más variables en futuros análisis, como la frecuencia de uso de la tarjeta de crédito, antigüedad del cliente y variables macroeconómicos (tasa de desempleo, inflación, PIB), esto permitirá una visión más contextual y dinámica del comportamiento de pago.
5. Se sugiere replicar este estudio en diferentes sectores financieros para validar la aplicabilidad de la metodología en otras áreas de crédito.

REFERENCIAS BIBLIOGRÁFICAS

Arthur, D. &. (2007). Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms., (págs. 1027-1035).

Bancos, S. d. (2023). *Boletín Estadístico Mensual: Tasa de morosidad en tarjetas de crédito.* Quito, Ecuador: Superintendencia de Bancos. Obtenido de https://www.superbancos.gob.ec/

Bester, T., & Rosman, B. (2024). *Towards financially inclusive credit products through financial time series clustering.* arXiv. doi:https://doi.org/10.48550/arXiv.2402.11066

Bholowalia, P. &. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications, 105*(9), 17-24.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Charrad, M. G. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1-36.

de Souza, D. H. (2021). Ensemble and mixed learning techniques for credit card fraud detection. *arXiv*. doi:https://doi.org/10.48550/arXiv.2112.02627

Ester, M. K. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, (págs. 226-231).

García, M. (2023). *Predicción de Default en RD: un enfoque de Machine Learning para la evaluación del riesgo crediticio.* Santo Domingo: Superintendencia de Bancos de la República Dominicana. Obtenido de https://sb.gob.do/media/aadppshl/prediccion-default-rd-enfoque-machine-learning-riesgo-crediticio.pdf

Han, J. P. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Jabeen, M. e. (2023). Modelo de Machine Learning basado en la Máquina Potenciadora de Gradiente de Luz para predecir la probabilidad de impago en clientes de la cartera de tarjeta de crédito. *Revista de Investigación de Sistemas e Informática, 16*(2), 155-168. doi:https://doi.org/10.15381/risi.v16i2.27140

Jain, A. K. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323.

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

Kaufman, L. &. (2005). *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons.

Kotler, P. &. (2012). *Marketing management* (14th ed.). Pearson.

Kotler, P., & Keller, K. L. (2012). *Dirección de marketing* (14.ª ed. ed.). Educación Pearson.

Lozano, J. A., García, R. A., & Carrillo, M. (2020). Segmentación de clientes morosos utilizando algoritmos de agrupamiento no supervisado. *Revista Colombiana de Computación, 21*(2), 34-47. doi:https://doi.org/10.22201/rcuc.2020.21.2.450

Morales-Vargas, D., González-Albarracín, M. L., & Gómez, M. C. (2023). Modelo híbrido para la predicción del comportamiento de pago en microfinanzas utilizando técnicas de Machine Learning. *Contaduría y Administración, 68*(2), 1-25. doi:https://doi.org/10.22201/fca.24488410e.2023.2751

Naik, K. S. (2021). Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach. *arXiv*. doi:https://doi.org/10.48550/arXiv.2110.02206

Ng, A. Y. (2002). On spectral clustering: Analysis and an algorithm., *14*, págs. 849-856.

Pestana, D. (2025). Automating Credit Card Limit Adjustments Using Machine Learning. *arXiv*. doi:https://doi.org/10.48550/arXiv.2501.10451

Qi, L., Wang, D., Zhang, C., & Xu, L. (2020). A clustering-based approach for credit card customer segmentation using behavioral data. *Journal of Risk and Financial Management, 13*(7), 153. doi:https://doi.org/10.3390/jrfm13070153

Ramírez, D., & Torres, J. (2021). Análisis de la cartera vencida mediante técnicas de machine learning en cooperativas financieras ecuatorianas. *Revista de Ciencias Financieras y Económicas, 11*(1), 58-70. doi:https://doi.org/10.21676/16574923.4140

Rokach, L. &. (2005). Data Mining and Knowledge Discovery Handbook. En O. &. Maimon (Ed.). Springer.

Rose, P. S. (2013). *Bank management & financial services* (9th ed.). McGraw-Hill.

Sánchez-Morán, J., & Vargas-Aguiar, C. (2023). Comparación de algoritmos de clustering para segmentación de clientes en instituciones microfinancieras de Ecuador. *Revista Iberoamericana de Ciencia de Datos, 7*(1), 45-60. doi:https://doi.org/10.32719/abcd.v7.n1.2023.45

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer.

Zapata-Rodríguez, M., Martínez, L. C., & Rodríguez, C. J. (2023). Machine learning clustering model to classify customers in default risk using real banking data from Colombia. *Journal of Applied Economics, 26*(1), 55-70. doi:https://doi.org/10.1080/15140326.2023.2254761

ANEXOS

Anexo 1: Detalles para la Reproducibilidad del Análisis:

Con el objetivo de garantizar la transparencia y la posibilidad de replicar los resultados obtenidos en esta investigación, a continuación, se detalla el entorno computacional, los paquetes utilizados y los principales pasos del análisis realizado.

1. **Entorno Computacional**
    - Versión de R utilizada: R version 4.3.3 (2024-02-29)
    - Sistema operativo: Windows 11 Pro, 64 bits
2. **Paquetes de R y versiones utilizadas**

Tabla 14*:* Paquetes de R y versión.

Paquete	Versión
dplyr	1.1.4
tidyr	1.3.1
ggplot2	3.5.0
factoextra	1.0.7
cluster	2.1.6
NbClust	3.0.1
readr	2.1.5
lubridate	1.9.3
scales	1.3.0

Elaborado por: Gordillo, Víctor (2025).

La obtención de las versiones se realizó a través de la función sessionInfo() de R.

3. **Flujo general del análisis**

   El análisis siguió la siguiente secuencia de pasos, alineada a la metodología CRISP-DM:

   - Carga e inspección inicial del dataset (readr::read_csv)
   - Limpieza y transformación de variables clave (dplyr, tidyr)
   - Análisis exploratorio y visualización (ggplot2, scales)
   - Determinación del número óptimo de clusters (método del codo y NbClust)
   - Aplicación del algoritmo K-means (stats::kmeans, factoextra::fviz_cluster)
   - Interpretación y segmentación según características de riesgo

4. **Observaciones adicionales**

   - Los scripts fueron organizados en archivos .R y ejecutados de forma secuencial.
   - Las visualizaciones fueron generadas con ggplot2 y exportadas como archivos .png.
   - El conjunto de datos utilizado proviene de una fuente institucional con restricciones de acceso, por lo que no se publica en este documento. Sin embargo, el código completo y reproducible puede ser compartido previa solicitud, bajo acuerdo de confidencialidad.