FICHA RESUMEN
TRABAJO DE GRADO – MAESTRÍA EN CIENCIA DE DATOS

TITULO: "GESTIÓN DEL RIESGO Y COBRANZA MEDIANTE MODELOS PREDICTIVOS EN CIENCIA DE DATOS EN EDUCACIÓN SUPERIOR"

1. ÉNFASIS: Ciencia de Datos aplicada a la gestión del riesgo financiero en una Institución de Educación Superior.
2. TIPO DE PROYECTO: Aplicado.
3. ÁREA DE TRABAJO: Educación superior y gestión del riesgo financiero institucional.
4. ESTUDIANTE (S): Edgar Esteban Grajales Castaño, Paula Andrea Tovar Rios.
5. CORREO ELECTRÓNICO: eegrajales@javerianacali.edu.co, paolat87@javerianacali.edu.co.
6. DIRECCIÓN Y TELÉFONO: 16 calle 8-49 casalini, zona 10, ciudad de Guatemala cel. 44723352 Calle 15 No 67-51 cel. 3166910873
7. DIRECTOR: Isabel Cristina García Arboleda
8. VINCULACIÓN DEL DIRECTOR: Docente Externo
9. CORREO ELECTRÓNICO DEL DIRECTOR: icga03@gmail.com
10. CO-DIRECTOR(ES) (Si aplica): No aplica
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Universidad Autónoma de Occidente
12. OTROS GRUPOS O EMPRESAS: No aplica
13. PALABRAS CLAVE (al menos 5): Gestión del Riesgo y Cobranza, Educación Superior, Modelos Predictivos, Regresión Logística, Desbalanceo de Clases.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Educación de calidad.
15. FECHA DE INICIO (Desarrollo del proyecto): 01/05/2025 al 25/05/2026
16. RESUMEN (máximo 400 palabras). Este proyecto tiene como objetivo aplicar técnicas de Ciencia de Datos para optimizar la gestión del riesgo y la cobranza de créditos en la educación superior, abordando una problemática crítica en la Universidad Autónoma de Occidente (UAO), la necesidad de garantizar la sostenibilidad financiera y la permanencia estudiantil mediante créditos educativos más eficientes. Actualmente, los modelos tradicionales no permiten identificar con precisión los perfiles de estudiantes con mayor riesgo de incumplimiento, lo que afecta la toma de decisiones estratégicas y limita la capacidad de respuesta ante situaciones de morosidad. A través del desarrollo de modelos predictivos y análisis de datos históricos, se espera mejorar la identificación de perfiles de riesgo, reducir la morosidad y fortalecer la eficiencia en la gestión de cobranza. Además, se busca implementar herramientas de visualización que permitan a la Dirección Financiera interpretar los resultados de manera clara y tomar decisiones informadas, con potencial de replicabilidad en otras instituciones de educación superior. El proyecto se basa en datos reales y en un enfoque contextualizado, lo que garantiza su relevancia práctica y la posibilidad de generar soluciones sostenibles y ajustadas a las necesidades.

GESTIÓN DEL RIESGO Y COBRANZA MEDIANTE MODELOS PREDICTIVOS EN CIENCIA DE DATOS EN EDUCACIÓN SUPERIOR

Edgar Esteban Grajales Castaño Código 9026277,
Paula Andrea Tovar Rios Código 8928119

Proyecto Aplicado para optar al título de
Magíster en Ciencia de Datos

Dra. Isabel Cristina García Arboleda
Directora

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Las instituciones de educación superior que otorgan créditos educativos con recursos propios enfrentan un reto importante para sostener su operación financiera, pues tienen que gestionar oportunamente el riesgo de incumplimiento de pago. Las entidades financieras reguladas trabajan con modelos sólidos de evaluación, seguimiento y cobranza apoyados en datos. Muchas universidades, en cambio, administran este riesgo con herramientas limitadas y enfoques principalmente reactivos. Bajo esas condiciones, la mora deja de ser un indicador financiero aislado y pasa a afectar la planeación, la recuperación de cartera y la gestión de alternativas de financiación que apoyan las estrategias institucionales de permanencia estudiantil.

En la Universidad Autónoma de Occidente (UAO), la problemática se hizo más notoria por el deterioro registrado en la cartera de créditos educativos durante los últimos años. El aumento de la tasa de mora, sobre todo a partir de la pandemia por COVID-19, mostró la necesidad de fortalecer la gestión del riesgo con un enfoque preventivo. De ahí surge el presente proyecto, orientado a identificar patrones de incumplimiento y a construir herramientas analíticas que permitan anticipar escenarios de riesgo y apoyar la toma de decisiones en el Departamento de Apoyo Financiero a Estudiantes.

Para trabajar el problema se usó la información histórica de solicitudes de crédito junto con la base de cartera vigente de la institución. Sobre esas fuentes se aplicó la metodología *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que organiza el análisis desde la comprensión del problema hasta la evaluación de los resultados. Durante el recorrido se construyeron variables con enfoque histórico, se adoptaron medidas para evitar sesgos en la información utilizada y se evaluaron distintos modelos de clasificación orientados a estimar la probabilidad de incumplimiento.

Conviene precisar una decisión metodológica tomada durante la ejecución del proyecto. El anteproyecto delimitó inicialmente el análisis a los créditos educativos de corto plazo. Sin embargo, la fase de comprensión de los datos del ciclo CRISP-DM (iterativo por diseño) reveló una diferencia estructural en el comportamiento de pago entre ambos plazos: la tasa de mora alcanzó el 1.8% en los créditos de corto plazo frente al 25% en los de largo plazo. Esa evidencia empírica llevó a ampliar el alcance del modelado al portafolio crediticio completo y a conservar el tipo de plazo como variable explicativa. Con ese ajuste se ganó representatividad y utilidad institucional del modelo.

Los resultados mostraron que sí es posible predecir el riesgo de mora con un desempeño satisfactorio y, a la vez, generar herramientas útiles para la operación institucional. Se identificaron variables con buena capacidad explicativa, se seleccionaron modelos con un balance adecuado entre interpretabilidad y desempeño predictivo, y se validó la estabilidad temporal de

los resultados. A partir de eso se propusieron instrumentos aplicables a la gestión de cartera: un esquema de segmentación por zonas de riesgo, criterios de priorización para acciones de cobranza preventiva y un análisis del retorno esperado de la intervención temprana.

Este trabajo ofrece una referencia metodológica para el análisis del riesgo crediticio en instituciones de educación superior que cuentan con mecanismos de financiación propia. Lo que muestra el estudio es que, incluso fuera del sector financiero tradicional, se pueden desarrollar herramientas de analítica orientadas a apoyar la gestión del riesgo y la toma de decisiones. El estudio aporta entonces al fortalecimiento de la gestión interna y, además, puede servir de base para futuras iniciativas que incorporen enfoques predictivos en la administración de la cartera educativa.

El documento se organiza en ocho capítulos, además de las conclusiones, las referencias bibliográficas y los anexos. La estructura comienza con la definición del problema, los objetivos y el marco teórico que orienta el estudio. Luego se desarrolla la metodología aplicada y se exponen los resultados del proyecto, desde la caracterización de la población y la construcción del conjunto de datos hasta la evaluación de los modelos predictivos, la segmentación de perfiles de riesgo y el diseño de herramientas de apoyo para la gestión institucional.

Nota de confidencialidad y restricción de publicación

El presente documento contiene análisis, resultados, tableros y conclusiones derivados de información institucional real de la Universidad Autónoma de Occidente, utilizada exclusivamente con fines académicos en el marco del Proyecto Aplicado de la Maestría en Ciencia de Datos. Aunque la información se presenta de forma agregada y no incluye identificación directa de estudiantes, los datos, resultados y hallazgos corresponden a procesos internos de gestión financiera, crédito educativo y cobranza de la institución. Por esta razón, el documento, sus anexos, códigos, tableros y archivos asociados no deberán ser publicados, distribuidos, reproducidos ni compartidos de manera abierta sin autorización expresa de la Universidad Autónoma de Occidente. Se solicita que la consulta, almacenamiento o eventual incorporación de este trabajo en repositorios académicos se realice bajo modalidad de acceso restringido, reserva institucional o embargo, según las políticas aplicables y previa autorización de la institución titular de la información.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

Las universidades privadas en Colombia tienen desafíos importantes asociados a la permanencia estudiantil y a la gestión del riesgo crediticio. La deserción estudiantil en la educación superior es un problema crítico que afecta la continuidad académica de los estudiantes y la sostenibilidad financiera de las instituciones [1]. La capacidad de los estudiantes para mantenerse en sus programas está directamente ligada al acceso a financiamiento educativo y al cumplimiento de las obligaciones financieras que adquieren.

La financiación de la educación superior es un eje clave para el acceso, la permanencia y la graduación de miles de estudiantes en Colombia. Aunque muchas universidades asumieron un rol activo en la oferta de créditos institucionales a corto y largo plazo, esas alternativas internas no siempre cuentan con el respaldo financiero, normativo y tecnológico que tienen entidades especializadas como el Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior (ICETEX) o el sistema bancario. Esas entidades disponen de modelos avanzados de credit scoring, herramientas de validación automatizada y políticas de cobranza basadas en datos; con eso mitigan el riesgo de morosidad, protegen el historial crediticio de los estudiantes y aseguran la sostenibilidad financiera de las operaciones.

La aplicación de modelos predictivos basados en Ciencia de Datos ofrece una alternativa viable para mejorar la gestión de créditos educativos. Estos modelos facilitan la identificación de perfiles de riesgo, la anticipación de posibles incumplimientos y el diseño de estrategias de cobranza personalizadas [2].

Las instituciones de educación superior no han sido ajenas a esta problemática. Aunque cuenta con un sistema de financiamiento estudiantil que ofrece créditos educativos a corto y largo plazo, todavía enfrenta retos importantes en la evaluación del riesgo crediticio y en la gestión de la cobranza. Si bien se han implementado procesos de automatización para la solicitud de créditos, persisten limitaciones, sobre todo en la predicción de incumplimientos y en la definición de estrategias de cobranza más efectivas. La ausencia de modelos analíticos que permitan evaluar el perfil del solicitante, segmentar la cartera por nivel de riesgo y activar alertas tempranas ha producido procesos operativos poco eficientes, ha aumentado la carga manual y ha afectado la percepción de los estudiantes frente al sistema.

A partir de esa problemática, se aplicaron técnicas de Ciencia de Datos para construir un sistema orientado a la identificación temprana del riesgo crediticio, a la segmentación de la cartera según patrones de comportamiento y a la priorización de acciones de cobranza con base en modelos predictivos. El sistema integró algoritmos de aprendizaje supervisado, análisis multivariado y visualización de indicadores clave. Eso permitió tomar decisiones con datos reales y anticiparse a escenarios de incumplimiento.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo contribuye la Ciencia de Datos a la identificación temprana del riesgo de incumplimiento en créditos educativos otorgados por instituciones de educación superior?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo predictivo basado en técnicas de Ciencia de Datos que permita identificar tempranamente el riesgo de incumplimiento en créditos educativos otorgados por instituciones de educación superior, facilitando la toma de decisiones estratégicas en la gestión del riesgo crediticio.

2.2 OBJETIVOS ESPECÍFICOS

- Analizar las variables académicas, socioeconómicas y de comportamiento de los estudiantes beneficiarios de créditos educativos, identificando aquellas asociadas al riesgo de incumplimiento.
- Preparar, transformar y depurar los datos históricos del sistema de financiación institucional para su uso en modelos de aprendizaje supervisado.
- Implementar y evaluar modelos de clasificación para anticipar el riesgo de incumplimiento, seleccionando el más adecuado según métricas de desempeño.
- Aplicar técnicas de segmentación y diseñar un tablero de control interactivo que facilite la visualización de perfiles de riesgo y apoye la toma de decisiones institucionales.

3 MARCO TEÓRICO Y ANTECEDENTES

Este marco teórico sustenta teórica y metodológicamente el proyecto Gestión del riesgo y cobranza mediante modelos predictivos en Ciencia de Datos en educación superior. El capítulo articula los fundamentos teóricos con las decisiones metodológicas del estudio. Para cada concepto se explica por qué es relevante para el problema específico de la UAO y cómo condiciona el diseño del modelo predictivo.

3.1 MARCO TEÓRICO

3.1.1 Gestión del riesgo crediticio

La gestión del riesgo crediticio es un componente esencial en las organizaciones que otorgan financiación, sobre todo en contextos como la educación superior, donde los créditos educativos son un mecanismo clave para garantizar el acceso y la permanencia estudiantil. El riesgo crediticio se define como la probabilidad de que un prestatario incumpla total o parcialmente sus obligaciones financieras y genere pérdidas a la entidad prestamista [2]. Este riesgo es especialmente importante para organizaciones que dependen de flujos de caja confiables para mantener su solvencia y liquidez.

La Superintendencia Financiera de Colombia define el riesgo crediticio como la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos por el incumplimiento de un deudor o contraparte, ya sea porque falle en el cumplimiento oportuno o cumpla imperfectamente los términos acordados en los contratos de crédito [3]. En el ámbito educativo, la gestión de créditos implica identificar y mitigar este riesgo mediante el análisis de la capacidad de pago, el historial crediticio y las garantías ofrecidas [4].

Los modelos tradicionales para evaluar el riesgo crediticio incluyen herramientas estadísticas y financieras como los modelos de *scoring* crediticio, el análisis de ratios financieros y los modelos econométricos [2]. Con esas herramientas se ha podido clasificar a los clientes según su probabilidad de incumplimiento. Sin embargo, estos marcos regulatorios y herramientas fueron diseñados para el sector financiero formal (bancos, cooperativas y entidades vigiladas por la Superintendencia Financiera), que cuenta con acceso a centrales de riesgo, modelos de *scoring* estandarizados y equipos especializados en análisis de cartera. Las instituciones de educación superior que otorgan créditos internos operan en un contexto distinto, lo que obliga a adaptar tanto los conceptos como las herramientas. La siguiente sección analiza esa adaptación.

3.1.2 La gestión del riesgo crediticio

La Universidad cuenta con programas de financiación interna pensados para facilitar el acceso a la educación superior mediante créditos educativos de corto y largo plazo. Estos programas requieren seguimiento continuo del comportamiento de pago de los estudiantes, ya que la renovación del crédito depende del cumplimiento en los pagos anteriores. Durante la pandemia

de COVID-19, la institución aplicó medidas de flexibilización en plazos y criterios de pago, lo que afectó la capacidad de predicción del riesgo y aumentó la heterogeneidad de la cartera. Por eso es necesario contar con modelos adaptativos que permitan analizar la evolución del comportamiento financiero estudiantil y proyectar escenarios de riesgo con base en información histórica. El presente proyecto se orienta a desarrollar modelos predictivos que integren variables financieras, académicas y socioeconómicas, con el fin de anticipar situaciones de incumplimiento y proponer estrategias de cobranza proactiva.

3.1.3 Ciencia de Datos y *Machine Learning*

La Ciencia de Datos es un campo interdisciplinario dedicado a la extracción de conocimiento a partir de grandes volúmenes de datos, que combina estadística, informática, matemáticas e ingeniería. Se centra en los tipos de analítica predictiva (que usa datos históricos para anticipar comportamientos futuros) y prescriptiva (que recomienda acciones específicas) [2].

En el contexto del riesgo crediticio, la Ciencia de Datos aporta un conjunto de herramientas y metodologías que transforman datos complejos en conocimiento útil para la toma de decisiones. Permite analizar grandes volúmenes de información, identificar patrones ocultos y generar modelos predictivos que sirven de apoyo a esa toma de decisiones. Incluye técnicas de análisis exploratorio, reducción de dimensionalidad y modelos predictivos como la regresión, la clasificación y el *clustering* [5].

Los modelos de *scoring* crediticio tradicionales se han apoyado en reglas estáticas o en análisis financiero clásico, propios del sector bancario [6]. Sin embargo, esos esquemas no contemplan particularidades del crédito educativo en instituciones de educación superior, como su carácter condicionado y su evaluación continua. La financiación interna está sujeta al buen comportamiento de pago dentro del semestre: solo con ese cumplimiento el estudiante puede renovar su crédito para el periodo siguiente. De lo contrario, el caso pasa a los procesos de cobranza establecidos por la División Financiera. Esta dinámica exige modelos más flexibles y adaptativos, como los que ofrece la Ciencia de Datos.

El *Machine Learning*, como rama de la inteligencia artificial, busca construir algoritmos capaces de aprender patrones y realizar predicciones precisas. En este proyecto se emplean dos enfoques: el aprendizaje supervisado para tareas de clasificación con datos etiquetados (como regresión logística y *Random Forest*), y el aprendizaje no supervisado, que busca descubrir estructuras ocultas sin datos etiquetados, como en el *clustering* aplicado a la segmentación de perfiles de riesgo [5].

Una situación frecuente en los problemas de *scoring* es el desbalanceo de clases. La mayoría de los estudiantes cumplen con sus pagos, lo que plantea retos adicionales en la construcción de modelos confiables [7]. En el caso del Departamento de Apoyo Financiero de la Universidad, ese desbalance se acentúa porque una proporción importante de los créditos son aprobados como parte de una política institucional orientada a facilitar el acceso y la permanencia en la educación superior. La estrategia, esencial desde una perspectiva de inclusión, exige desarrollar modelos predictivos capaces de identificar con precisión los pocos casos de incumplimiento dentro de una

población altamente positiva. Para afrontar ese reto se utilizan técnicas como el *Synthetic Minority Oversampling Technique* (SMOTE) [8] y métricas específicas como la matriz de confusión y las curvas *Receiver Operating Characteristic* (ROC), que validan la precisión y la capacidad predictiva de los modelos generados.

Estas herramientas evitan que los modelos caigan en sesgos derivados de la desproporción de clases y mejoran su efectividad. Además, la selección de modelos y su evaluación se apoyan en métricas como la precisión, que ayudan a descubrir relaciones entre variables socioeconómicas, académicas y comportamentales y a adaptarse a las características particulares de este tipo de financiación.

3.1.4 *Clustering* y su relevancia en el sector financiero

El *clustering* o agrupamiento de datos permite identificar grupos homogéneos dentro de la cartera crediticia y facilita la segmentación de clientes según su nivel de riesgo y comportamiento de pago [5]. La literatura especializada identifica varios enfoques de *clustering* (jerárquico, por partición, basado en densidad, entre otros) que ofrecen flexibilidad metodológica según la estructura de los datos y los objetivos del análisis [9]. El algoritmo *K-means*, muy usado en estudios financieros, ayuda a diferenciar estrategias de cobranza y a diseñar intervenciones específicas según el perfil del grupo [2].

El agrupamiento se vuelve una herramienta esencial para entender la diversidad de perfiles estudiantiles y para priorizar recursos en la gestión del crédito educativo.

3.1.5 Interpretabilidad, equidad y visualización

Cuando las predicciones de un modelo afectan decisiones sobre personas reales, la interpretabilidad y la equidad son requisitos éticos y operativos. La interpretabilidad se refiere a la capacidad de explicar por qué el modelo asignó un determinado *score* a un estudiante específico, lo que permite al asesor financiero validar la predicción y tomar decisiones informadas. Existen técnicas que entregan explicaciones locales para cada predicción individual, con las que es posible entender qué variables influyeron en el score asignado a cada estudiante [10].

La equidad algorítmica plantea la pregunta de si el modelo trata de manera justa a diferentes grupos demográficos [11]. En el crédito educativo, el uso de variables como el estrato socioeconómico o el sexo como predictores podría generar discriminación sistémica si el modelo penalizara desproporcionadamente a un grupo protegido. Este proyecto no implementó un sistema formal de monitoreo de equidad, pero reconoce de manera explícita la necesidad de evaluar las métricas de desempeño por subgrupos para detectar sesgos.

La visualización de datos cumple un rol importante en el despliegue de modelos predictivos en organizaciones donde los usuarios finales no son científicos de datos. Los principios de visualización adoptados en este proyecto consideran criterios de expresividad, efectividad y accesibilidad [12]. En este proyecto la visualización no se limitó a las figuras del análisis

exploratorio; se extendió al diseño de un tablero de control interactivo como herramienta de despliegue operativo para el Departamento de Apoyo Financiero.

3.1.6 Credit scoring y métricas de evaluación en riesgo crediticio

El credit scoring es una herramienta analítica utilizada para estimar la probabilidad de incumplimiento de un solicitante o de una obligación financiera. Su propósito no es reemplazar la decisión institucional, sino aportar una medida cuantitativa que apoye la evaluación, priorización y seguimiento del riesgo. En los sistemas tradicionales de crédito, el scoring permite ordenar a los deudores según su probabilidad estimada de incumplimiento y facilita la definición de políticas diferenciadas de otorgamiento, seguimiento y cobranza.

En el contexto de este proyecto, el concepto de credit scoring se adapta al crédito educativo institucional. A diferencia del sector financiero tradicional, donde el análisis suele apoyarse en información externa proveniente de centrales de riesgo, capacidad de endeudamiento, ingresos formales y garantías, el modelo se construye principalmente con información interna del estudiante, del crédito y de su comportamiento histórico con la institución. Por tanto, el score no debe interpretarse como una calificación financiera general del estudiante, sino como una estimación del riesgo de incumplimiento de una obligación educativa dentro del contexto institucional analizado.

La evaluación de modelos de scoring requiere métricas que permitan medir su capacidad de discriminación y su utilidad operativa. La discriminación hace referencia a la capacidad del modelo para asignar mayores probabilidades de riesgo a las obligaciones que efectivamente presentan incumplimiento y menores probabilidades a aquellas que no lo presentan. En este sentido, el AUC-ROC es una métrica ampliamente utilizada porque resume la capacidad del modelo para separar las clases positiva y negativa a lo largo de diferentes umbrales de clasificación. Un AUC cercano a 0,5 indica que el modelo discrimina de forma similar al azar, mientras que valores superiores reflejan mayor capacidad predictiva.

Además del AUC-ROC, en problemas de riesgo crediticio es común utilizar el coeficiente de Gini y el estadístico KS. El coeficiente de Gini se deriva directamente del AUC mediante la expresión:

$$\text{Gini} = 2 \times \text{AUC} - 1$$

Esta métrica permite expresar la capacidad discriminante del modelo en una escala donde valores cercanos a cero indican baja separación entre buenos y malos pagadores, mientras que valores más altos indican mayor poder de clasificación. Por su parte, el estadístico Kolmogorov-Smirnov, conocido como KS, mide la máxima distancia entre la distribución acumulada de los scores de la clase positiva y la distribución acumulada de los scores de la clase negativa. En modelos de riesgo, un mayor KS indica que el modelo logra separar mejor los casos con incumplimiento de los casos sin incumplimiento.

No obstante, las métricas de discriminación no son suficientes por sí solas. Dado que el objetivo institucional es anticipar posibles casos de mora y activar acciones preventivas, también se consideran métricas asociadas a la clase minoritaria, como Recall, Precision y F1-score. El Recall permite medir qué proporción de los casos reales de mora fueron identificados por el modelo.

Esta métrica es especialmente importante en el proyecto, porque un falso negativo representa una obligación riesgosa que no sería priorizada para intervención preventiva. El F1-score, por su parte, resume el balance entre Precision y Recall, y resulta útil cuando existe desbalance de clases.

Por esta razón, la evaluación del modelo no debe basarse en una única métrica. El desempeño puede analizarse de manera integral a partir de AUC-ROC, Recall, Precision, F1-score, matriz de confusión, PR-AUC y medidas de calibración como Brier Score. Adicionalmente, en modelos de riesgo crediticio resulta recomendable incorporar métricas complementarias como Gini y KS, especialmente en futuras iteraciones del modelo, con el fin de fortalecer la comparación con prácticas habituales de credit scoring.

3.1.7 Desbalanceo de clases y técnicas de remuestreo

El desbalanceo de clases es una condición frecuente en los problemas de clasificación asociados al riesgo crediticio. En este tipo de problemas, la proporción de casos que incumplen suele ser considerablemente menor que la proporción de casos que cumplen con sus obligaciones. Esta característica genera un reto metodológico, porque un modelo puede alcanzar altos niveles de exactitud clasificando correctamente la clase mayoritaria, pero fallar en la identificación de los casos de mayor interés: aquellos que presentan incumplimiento.

En el caso del crédito educativo de la Universidad, la clase minoritaria corresponde a las obligaciones clasificadas como mora o cartera, mientras que la clase mayoritaria corresponde a las obligaciones canceladas dentro de los plazos definidos. Este desbalance es consistente con la naturaleza institucional del crédito educativo, cuyo propósito principal es apoyar el acceso y la permanencia estudiantil. Sin embargo, desde el punto de vista predictivo, el bajo volumen relativo de casos en mora puede afectar el aprendizaje del modelo, especialmente si el algoritmo tiende a privilegiar la clase mayoritaria.

Para abordar esta situación se evaluó el uso de técnicas de remuestreo, particularmente SMOTE, Synthetic Minority Oversampling Technique. Esta técnica genera observaciones sintéticas de la clase minoritaria a partir de la interpolación entre casos existentes, con el fin de ofrecer al algoritmo una representación más equilibrada de los patrones asociados al incumplimiento. A diferencia de una simple duplicación de registros, SMOTE construye nuevos ejemplos sintéticos dentro del espacio de características de la clase minoritaria.

El uso de SMOTE debe realizarse con precaución. Si se aplica antes de dividir los datos o sobre los conjuntos de prueba, puede producir contaminación metodológica y sobreestimar el desempeño del modelo. Por esta razón, en el presente proyecto SMOTE se aplica únicamente sobre el

conjunto de entrenamiento, después de la división temporal de los datos y antes del ajuste del modelo. Los conjuntos de validación, holdout temporal y validación out-of-time se conservan con la distribución real de clases, con el propósito de evaluar el desempeño en condiciones similares a las que enfrentaría la institución en operación.

Adicionalmente, dado que el remuestreo puede alterar la prevalencia observada durante el entrenamiento y afectar la interpretación de las probabilidades predichas, su efecto debe ser contrastado empíricamente en futuras iteraciones del modelo. En este proyecto, SMOTE se utiliza como una estrategia metodológica para enfrentar el desbalanceo de clases, aplicada únicamente sobre el conjunto de entrenamiento. No obstante, se reconoce que una comparación formal entre modelos entrenados con y sin SMOTE permitiría determinar si esta técnica aporta mejoras reales en la identificación de la clase minoritaria o si introduce sesgos que afecten la calibración del modelo.

3.1.8 Validación temporal, calibración de probabilidades y selección de umbrales

En modelos predictivos aplicados a riesgo crediticio, la validación debe considerar la dimensión temporal del problema. En un contexto operativo real, el modelo se entrena con información histórica y se utiliza para predecir el comportamiento de obligaciones futuras. Por esta razón, una división aleatoria de los datos puede generar una estimación optimista del desempeño, al mezclar registros de periodos recientes en el entrenamiento y registros antiguos en la prueba. Para evitar este problema, es recomendable utilizar esquemas de validación temporal.

La validación temporal permite evaluar si el modelo conserva su capacidad predictiva cuando se enfrenta a periodos posteriores a aquellos utilizados durante el entrenamiento. Este criterio es especialmente relevante en el presente proyecto, debido a que el comportamiento de pago de los estudiantes se vio afectado por cambios estructurales durante y después de la pandemia por COVID-19. Por tanto, el desempeño del modelo debe analizarse no solo en términos de clasificación, sino también en términos de estabilidad frente a cambios en la distribución de los datos.

Además de la validación temporal, los modelos de scoring deben evaluar la calibración de las probabilidades predichas. Un modelo está bien calibrado cuando las probabilidades que asigna se corresponden con la frecuencia real observada del evento. Por ejemplo, si un grupo de obligaciones recibe una probabilidad promedio de mora del 30%, se esperaría que aproximadamente el 30% de esas obligaciones incurra efectivamente en mora. La calibración es importante porque el proyecto no solo busca clasificar obligaciones como riesgosas o no riesgosas, sino también construir zonas de riesgo y umbrales operativos para orientar la gestión preventiva.

La selección del umbral de clasificación es otra decisión metodológica relevante. El umbral

transforma una probabilidad continua en una clasificación binaria. Aunque el umbral de 0,5 es común en problemas generales de clasificación, no necesariamente es el más adecuado en escenarios de mora, especialmente cuando existe desbalance de clases y cuando el costo de no identificar un caso riesgoso puede ser mayor que el costo de intervenir preventivamente un caso que finalmente no incumple. Por ello, la selección del umbral debe responder a criterios estadísticos y operativos.

En este proyecto, la definición del umbral se orienta a equilibrar la capacidad de detección de la mora, la precisión de la intervención y la viabilidad operativa del proceso financiero. Por tanto, el umbral óptimo se analiza considerando métricas como Recall, Precision, F1-score, matriz de confusión, curva Precision-Recall y distribución de scores por zonas de riesgo.

3.2 ANTECEDENTES

Para cualquier organización, un objetivo estratégico ha sido asegurar su sostenibilidad a largo plazo, generar flujos de efectivo que permitan mantener un capital de trabajo adecuado, atender los requerimientos de inversión en activos, servir la deuda y mantener políticas financieras coherentes [13]. El crédito puede entenderse como una inversión realizada en un cliente, vinculada a la venta de un producto o servicio, cuyo propósito es incrementar las ventas y garantizar la recuperación de recursos [14].

Desde la década de 1930, varios investigadores desarrollaron modelos para evaluar la capacidad de pago y el riesgo de incumplimiento de los clientes. En ese periodo se propuso un modelo pionero que usaba información contable financiera como insumo para valorar el riesgo de crédito [15]. Durante la década de 1960 aparecieron aportes relevantes que aplicaron técnicas de análisis discriminante para la clasificación de riesgo [15], [16], [17]. Más adelante se avanzó hacia modelos con componentes estocásticos, mediante algoritmos de máxima verosimilitud y redes neuronales, que ampliaron la capacidad predictiva en escenarios de incertidumbre [18].

En el ámbito financiero y educativo, la pandemia del COVID-19 iniciada en 2019 profundizó las crisis económicas a nivel mundial. El confinamiento, la pérdida de empleo y la reducción de ingresos familiares afectaron la capacidad de pago de los estudiantes y generaron mayores presiones sobre los esquemas de financiamiento de la educación superior [19]. Hoy, la cartera de créditos educativos es uno de los principales indicadores de riesgo que las instituciones financieras y educativas deben monitorear. Una de las metodologías más utilizadas para evaluar el riesgo ha sido el método de las 5C’s, ampliamente adoptado por instituciones financieras para determinar la viabilidad de otorgar un crédito a un cliente potencial. Ese método evalúa factores como carácter, capital, capacidad, colateral y condiciones, que son criterios clave para el análisis y la mitigación del riesgo [20]. Una colocación inadecuada del crédito, sobre todo en el ámbito educativo, puede afectar directamente la liquidez de las instituciones privadas y limitar su capacidad de invertir en calidad académica e investigación [4].

El ICETEX, creado mediante el Decreto 2586 de 1950 y reorganizado después, se destaca como la principal entidad financiera de apoyo educativo en Colombia. Ofrece créditos con tasas preferenciales, sistemas de amortización flexibles y condonaciones parciales en casos específicos [21]. Su sistema de administración del riesgo crediticio (SARC) asegura el cumplimiento normativo y la gestión técnica del riesgo en la financiación educativa.

En Colombia, algunas universidades han empezado a aplicar técnicas de Ciencia de Datos para mejorar la gestión de becas y alternativas de financiación. Las experiencias documentadas en la predicción de riesgo de incumplimiento en créditos educativos son limitadas, debido a restricciones financieras, normativas y estructurales.

Esos modelos no siempre consideran las particularidades del sector educativo, como la evolución académica o las condiciones específicas de matrícula. Por eso el presente proyecto se diferenció al integrar técnicas de Ciencia de Datos aplicadas a datos institucionales reales. La propuesta es una solución contextualizada con resultados directamente aplicables a la mejora operativa del Departamento de Apoyo Financiero, con potencial de replicabilidad en otras instituciones de educación superior.

En conjunto, estos antecedentes ofrecen un marco de conceptos y estrategias que sustentaron el desarrollo del proyecto. Granda [22] analizó los determinantes del incumplimiento en créditos educativos del ICETEX e identificó variables como historial académico y capacidad de pago, que se consideraron en el modelo aunque adaptadas al contexto institucional. Zapata [23] propuso una guía metodológica para medir riesgo crediticio en entidades financieras; eso inspiró la aplicación de técnicas estadísticas, aunque el proyecto se diferenció al centrarse en créditos educativos internos y no en banca tradicional. Francova [24] exploró el uso de *Machine Learning* para predecir riesgo crediticio, lo que abrió la posibilidad de incorporar algoritmos predictivos avanzados. Por último, estudios como el de Montalván [25] sobre *credit scoring* con redes neuronales mostraron el potencial de modelos complejos, pero el enfoque del proyecto buscó equilibrar precisión y simplicidad para facilitar su implementación operativa.

4. METODOLOGÍA: APLICACIÓN DE CRISP-DM AL SISTEMA DE CRÉDITO EDUCATIVO

La metodología es lo que le da solidez y coherencia al proyecto. No es un trámite previo a lo interesante; es el marco que da validez científica a los resultados, permite que otro investigador los reproduzca, hace que cada decisión sea justificable con argumentos claros y asegura que los hallazgos sirvan para este contexto. Este capítulo presenta esa metodología como un relato continuo. Recorre, de principio a fin, el camino seguido desde la elección del marco de trabajo hasta la traducción del modelo en herramientas operativas para el Departamento de Apoyo Financiero. Pasa por la arquitectura del sistema de datos, la aplicación de cada fase del ciclo CRISP-DM al caso del crédito educativo, el diseño de la validación estadística y los principios de reproducibilidad que rigieron el proceso.

El punto de partida de cualquier proyecto de Ciencia de Datos aplicada es elegir el marco metodológico que estructurará el trabajo. Hay varias opciones. El proceso *Knowledge Discovery in Databases* (KDD) fue uno de los primeros enfoques en formalizar la extracción de conocimiento a partir de bases de datos [26], pero su carácter lineal no contempla las iteraciones entre fases que aparecen en un proyecto real. La metodología *Sample, Explore, Modify, Model, Assess* (SEMMA), propietaria de SAS *Institute*, se centra en el proceso técnico estadístico, pero no incluye la comprensión del negocio como fase explícita y está ligada a una plataforma comercial. El *Team Data Science Process* (TDSP) de Microsoft define roles, cronogramas y artefactos específicos; sin embargo, su dependencia del ecosistema Azure lo hace inadecuado para un proyecto académico que debe ser independiente de plataforma. Los marcos ágiles tipo *Scrum* ofrecen flexibilidad y velocidad de entrega, pero no contemplan la rigurosidad científica de validación de modelos ni la documentación de decisiones estadísticas que exige un proyecto de maestría [27]. La Tabla 4.1 resume la comparación de estos marcos candidatos.

Tabla 4.1 Comparación de marcos metodológicos candidatos para proyectos de Ciencia de Datos aplicada

Marco metodológico	**Origen y enfoque**	**Fortalezas**	**Limitaciones para este proyecto**	**Adoptado**
CRISP-DM	*Cross-Industry Standard Process for Data Mining* con base en [28]. Orientado a proyectos de minería de datos y *Machine Learning* en organizaciones.	Cíclico, iterativo, agnóstico a herramientas. Integra comprensión del negocio con el proceso técnico. Ampliamente validado en el sector financiero.	No contempla criterios para tratar cartera no madurada ni *drift* estructural en la variable objetivo.	✓ **ADOPTADO** Justificado en sección 4.1.1
TDSP	*Team Data Science Process* (*Microsoft*). Orientado a equipos empresariales con Azure.	Define roles, cronogramas y artefactos técnicos específicos.	Atado al ecosistema *Microsoft*. No apropiado para proyecto académico independiente de plataforma.	No adoptado

KDD Process	*Knowledge Discovery in Databases*. Proceso lineal de extracción de conocimiento.	Fundación teórica sólida. Primero en formalizar el proceso de minería.	Proceso lineal: no contempla iteraciones entre fases. Menos adecuado para proyectos aplicados con retroalimentación continua.	No adoptado
SEMMA	*Sample-Explore-Modify-Model-Assess* (SAS Institute). Metodología propietaria para SAS *Enterprise Miner*.	Muy centrado en el proceso técnico estadístico.	No incluye comprensión del negocio. Propietaria de SAS. No aplica para proyecto con Python/*scikit-learn*.	No adoptado
Metodología Ágil (Scrum adaptado)	*Sprints* cortos iterativos orientados a entrega continua de valor.	Flexibilidad y velocidad de entrega. Muy usado en *industria tech*.	No contempla la rigurosidad científica de validación de modelos ni documentación de decisiones estadísticas que exige un proyecto de maestría.	No adoptado

Frente a esas alternativas, se seleccionó CRISP-DM con base en [28], [29] por cuatro razones convergentes que se vinculan directamente con las características del proyecto.

Primero, la metodología deja reconocer que el desarrollo de un modelo predictivo en un entorno real no sigue una secuencia estrictamente lineal. A medida que avanzó el análisis aparecieron hallazgos que obligaron a volver sobre decisiones previas relacionadas con la comprensión de los datos, su preparación y la evaluación del modelo. En este proyecto, ese proceso se recorrió en tres momentos principales: inicialmente para construir el conjunto base de datos, luego para incorporar variables de comportamiento histórico y, al final, para ajustar el umbral de clasificación después de la validación temporal del modelo.

Segundo, CRISP-DM es el único marco que arranca de manera obligatoria con la comprensión del negocio antes de tocar los datos. Esa característica es crítica porque el sistema de crédito educativo tiene particularidades institucionales invisibles sin ese contexto previo: los créditos son de corto y largo plazo, no están regulados por la Superintendencia Financiera, el codeudor suele ser un familiar y la pandemia COVID-19 generó un *drift* estructural en el comportamiento de pago. Sin comprender el negocio primero, un analista podría tratar el *dataset* igual que uno de un banco comercial, con resultados metodológicamente incorrectos. Tercero, la literatura especializada en *credit scoring* y riesgo crediticio en educación superior [31] y los estudios de ICETEX en Colombia usa CRISP-DM o marcos equivalentes, lo que asegura la comparabilidad metodológica [32],[33]. Trabajos recientes en el contexto colombiano como los de [34],[24],[35] confirman la adopción de CRISP-DM como marco estándar para proyectos de *scoring* crediticio con aprendizaje supervisado.

Por último, CRISP-DM no prescribe herramientas técnicas específicas [26], lo que dejó seleccionar Python con *scikit-learn* como entorno de desarrollo, con reproducibilidad y acceso sin costos de licenciamiento. La Figura 4.1 muestra el diagrama del ciclo CRISP-DM adaptado al presente proyecto.

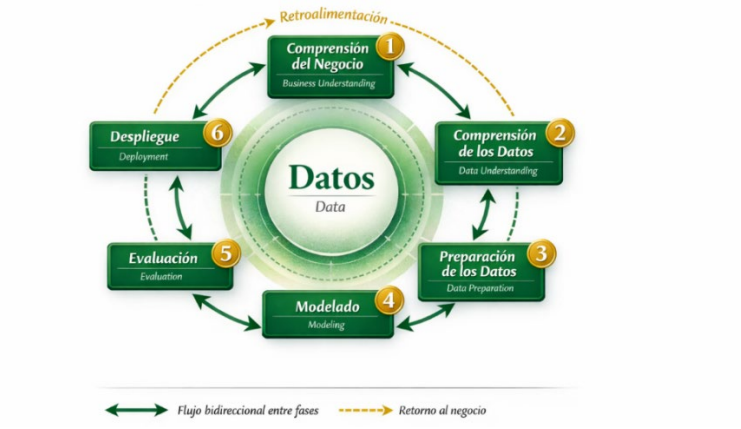


Figura 4.1 Metodología CRISP-DM. Fuente: diseño propio con base en [28].

Diseño experimental y reproducibilidad del proceso de modelado: Con el fin de fortalecer la reproducibilidad del proyecto, se documentó el diseño experimental seguido para la construcción, entrenamiento, validación y evaluación de los modelos predictivos. Este diseño permite que otro investigador pueda replicar el proceso bajo las mismas decisiones metodológicas, utilizando las fuentes institucionales autorizadas, el mismo criterio de construcción de la variable objetivo y los mismos esquemas de partición temporal.

El proceso experimental se desarrolló en las siguientes etapas:

Integración de las fuentes institucionales disponibles, correspondientes a solicitudes de crédito y cartera depurada vigente. Construcción de la variable objetivo En_Mora, a partir del estado de la obligación y de la regla institucional definida para identificar incumplimiento. Exclusión de registros no aptos para el modelado, incluyendo obligaciones refinanciadas, registros sin ventana de observación cerrada y variables con riesgo de data leakage. Construcción del vector final de variables predictoras con enfoque prospectivo, utilizando únicamente información disponible al inicio del periodo de análisis.

División temporal del conjunto de datos en entrenamiento, prueba temporal y validación out-of-time, respetando la secuencia cronológica del fenómeno financiero. Estandarización de variables numéricas y codificación de variables categóricas dentro del pipeline de modelado. Entrenamiento de modelos supervisados de clasificación: Regresión Logística, Árbol de Decisión, Random Forest y Gradient Boosting. Sintonización de hiperparámetros mediante búsqueda controlada sobre el conjunto de entrenamiento, utilizando validación cruzada estratificada.

Tratamiento del desbalanceo de clases mediante SMOTE, aplicado exclusivamente sobre el conjunto de entrenamiento, con evaluación posterior sobre datos reales no remuestreados. Esta decisión buscó mejorar la representación de la clase minoritaria sin alterar la distribución natural de los conjuntos de prueba, holdout temporal y validación out-of-time.

Evaluación de desempeño en el conjunto holdout temporal mediante AUC-ROC, Recall, Precision, F1-score, matriz de confusión, PR-AUC y Brier Score. Adicionalmente, se reconoció la relevancia de métricas propias del riesgo crediticio, como Gini y KS, las cuales se proponen como métricas complementarias para futuras iteraciones del modelo.

Selección del modelo operativo y del modelo interpretable, considerando simultáneamente desempeño predictivo, estabilidad temporal, interpretabilidad y utilidad institucional. Generación del scoring sobre la población objetivo 2025 y construcción de zonas de riesgo para la priorización de acciones preventivas. La Figura 4.2 resume el flujo experimental propuesto.

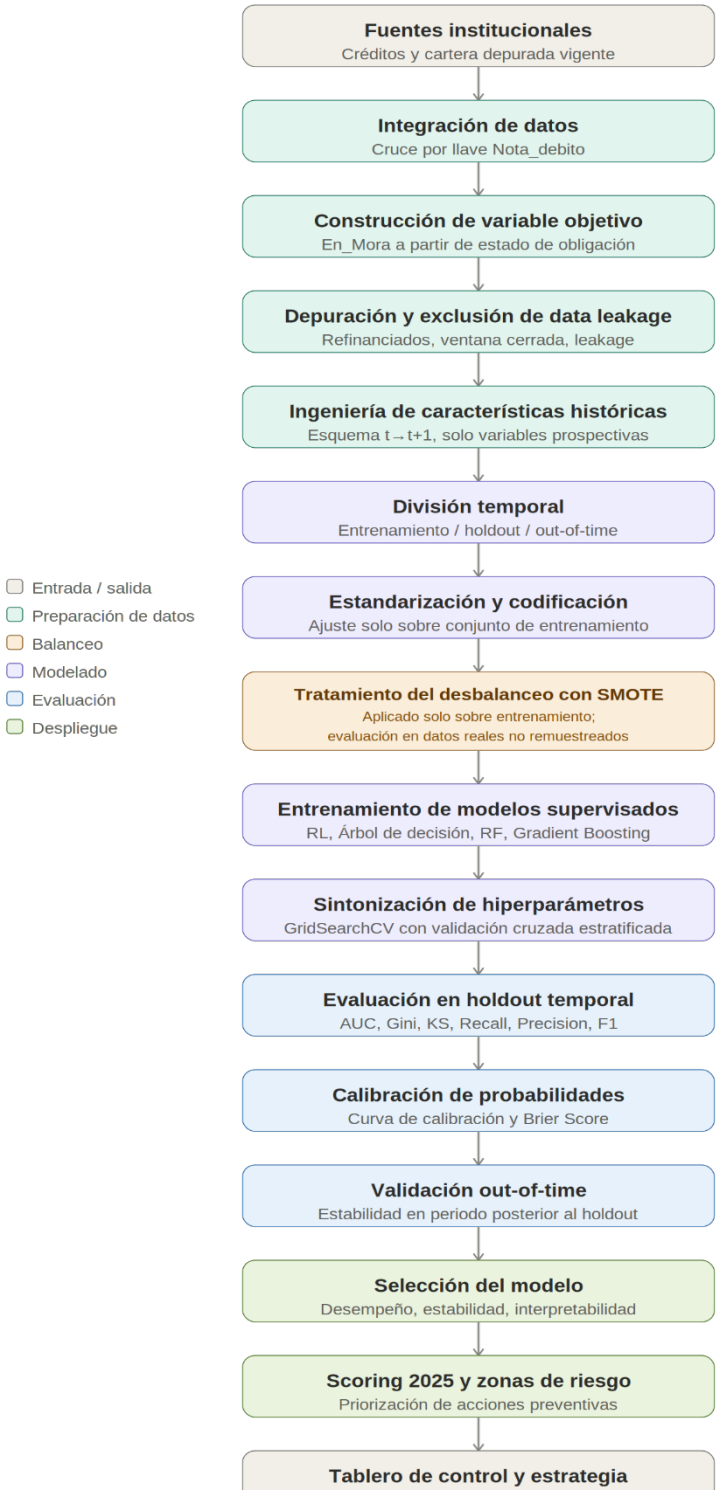


Figura 4.2. Flujo experimental y esquema de validación del modelo

Para garantizar la trazabilidad de los resultados, todos los experimentos se ejecutaron bajo una misma estructura de pipeline, evitando aplicar transformaciones sobre el conjunto total antes de la división temporal. Las transformaciones de escalamiento, codificación y remuestreo se ajustaron exclusivamente sobre el conjunto de entrenamiento y luego se aplicaron a los conjuntos de prueba y validación. Esta decisión evita que información futura o información de los conjuntos de evaluación influya en el aprendizaje del modelo.

Adicionalmente, se fijó una semilla aleatoria para los procedimientos que incluyen aleatoriedad, como la validación cruzada, la inicialización de ciertos algoritmos y la generación de observaciones sintéticas mediante SMOTE. El código fuente, el notebook de ejecución, las tablas de resultados y la versión exportada del tablero se incluyen como anexos digitales para facilitar la revisión técnica del proceso.

Esquema de partición temporal de los datos

El diseño de validación se basó en una partición temporal, debido a que el objetivo del proyecto es predecir el riesgo de incumplimiento de obligaciones futuras a partir de información histórica. Esta decisión evita una evaluación artificialmente optimista que podría generarse con particiones aleatorias, en las cuales registros de periodos futuros podrían quedar en entrenamiento y registros de periodos anteriores en prueba.

La partición se estructuró de la siguiente manera:

- Conjunto de entrenamiento: obligaciones históricas utilizadas para ajustar los modelos y aprender los patrones asociados al incumplimiento.
- Validación cruzada sobre entrenamiento: procedimiento utilizado para comparar configuraciones de hiperparámetros y evaluar la estabilidad interna del modelo sin utilizar el conjunto de prueba temporal.
- Holdout temporal: conjunto posterior al entrenamiento utilizado para evaluar el desempeño del modelo en periodos no vistos durante el ajuste.
- Validación out-of-time: periodo adicional, posterior al holdout, utilizado para validar la estabilidad temporal del modelo seleccionado y simular su comportamiento frente a una población futura.
- Este esquema permite separar tres decisiones metodológicas: el aprendizaje del modelo, la selección de la mejor configuración y la evaluación final del desempeño. De esta manera se reduce el riesgo de sobreajuste y se obtiene una estimación más realista de la capacidad predictiva del modelo en operación.
- Esta documentación permite diferenciar claramente entre los resultados obtenidos en entrenamiento, validación cruzada, prueba temporal y validación fuera de tiempo. Además, facilita que futuros equipos técnicos puedan recalibrar el modelo, incorporar nuevas variables o comparar nuevos algoritmos bajo el mismo marco experimental.

Antes de describir la aplicación fase por fase es necesario documentar la arquitectura del sistema de información sobre el que opera el proyecto. El proyecto se desarrolló a partir de dos fuentes de datos institucionales, ambas extraídas del sistema de información del Departamento de Apoyo Financiero a Estudiantes. La relación entre ambas fuentes sigue un modelo de datos relacional, ya que comparten la llave **Nota_debito / Nota_Debito**, que deja hacer el cruce directo entre los registros históricos del primer conjunto (solicitudes y seguimiento de créditos, con 191,130 registros y 25 variables) y el estado vigente de la cartera que refleja el segundo conjunto (cartera depurada vigente, con 18,452 registros y 22 variables). Esta documentación arquitectónica deja a cualquier evaluador entender el sistema como un todo antes de profundizar en cada fase del proceso.

Con el marco definido y la arquitectura documentada, el proceso CRISP-DM inicia con la comprensión del negocio en la primera fase. La fase exige que el equipo de Ciencia de Datos entienda a profundidad el contexto institucional antes de acceder a los datos, pues esa comprensión distingue un proyecto de modelado con relevancia organizacional de un ejercicio técnico desconectado de la realidad. La Universidad otorga créditos educativos institucionales de corto y largo plazo para financiar matrículas semestrales, y esos créditos tienen características que condicionan toda la metodología. A diferencia del crédito ofrecido por el ICETEX o por entidades bancarias, la Universidad solo consulta centrales de riesgo financiero para los créditos de largo plazo. Para la mayoría de los créditos, entonces, las variables disponibles para evaluar el riesgo son principalmente internas, y el comportamiento de pago histórico del estudiante con la institución cobra especial importancia. El codeudor es casi siempre un familiar cercano del estudiante, lo que implica que el riesgo del crédito está correlacionado con el riesgo económico del hogar, no únicamente con el desempeño académico individual.

Cada crédito se otorga para un semestre específico y al inicio del siguiente puede solicitarse uno nuevo cuya aprobación depende del comportamiento de pago previo. Esto establece un ciclo semestral donde el comportamiento de pago en periodos anteriores se vuelve la señal predictiva más valiosa para el siguiente. A esto se suma el impacto de la pandemia COVID-19 durante 2020-2021, cuando la Universidad flexibilizó los criterios de aprobación para no excluir financieramente a estudiantes cuyas familias habían perdido ingresos. Se generó un *drift* temporal que elevó la tasa de mora del 2.6% en 2019 al 11.0% en 2021 y que se mantuvo en aproximadamente 10.5% hasta 2024. La metodología contempla ese *drift* mediante el *split* temporal con corte en 2022.

El problema de negocio identificado es la ausencia de herramientas analíticas en el Departamento de Apoyo Financiero para gestionar proactivamente el riesgo crediticio. En concreto, la institución enfrenta tres brechas operativas:

La primera es la identificación tardía del incumplimiento, que se detecta de forma reactiva cuando el crédito ya está en mora, sin mecanismo para intervención preventiva.

La segunda es la insuficiencia de la segmentación de cartera. Aunque existen diferencias operativas según la etapa de gestión (como la remisión de casos en estado pre jurídico a un

tercero especializado), la institución no cuenta con una segmentación analítica basada en el perfil de riesgo del estudiante, la probabilidad de recuperación o las características socioeconómicas y académicas del beneficiario. Esto impide priorizar estrategias diferenciadas de cobranza antes de que el crédito alcance etapas avanzadas de deterioro.

La tercera es la falta de un análisis cuantitativo que demuestre el retorno de inversión de diferentes estrategias de intervención según el nivel de riesgo, lo que dificulta la priorización presupuestal.

La traducción de esas brechas a un problema formal de Ciencia de Datos produce la siguiente especificación técnica: Especificación técnica del problema de Ciencia de Datos

La traducción de las brechas operativas a un problema formal de Ciencia de Datos requiere definir con precisión qué se busca predecir, con qué información, bajo qué restricciones y con qué criterios se evaluará el éxito del modelo. La Tabla 4.2 presenta esta especificación técnica.

Tabla 4.2: Especificación técnica del problema de Ciencia de Datos

Componente	Especificación
Tarea	Clasificación binaria supervisada
Variable objetivo	En_Mora ∈ {0, 1}, donde 1 = crédito clasificado como CARTERA en el sistema institucional tras superar su fecha de vencimiento más 90 días de gracia; 0 = crédito cancelado completamente dentro de los plazos. Los créditos con estado REFINANCIADO se excluyen del modelado por ser un estado intermedio.
Restricción de prospectividad	El modelo solo puede usar variables conocidas al inicio del semestre, antes de que el crédito entre en mora. Variables que reflejen el resultado del pago (% Recaudo, Saldo_nota_debito, No_Creditos) están explícitamente excluidas
Criterios de éxito	AUC-ROC ≥ 0.65; *Recall* (clase 1) ≥ 0.60; *F1-score* ≥ 0.40
Alerta de *leakage*	Valores de AUC > 0.95 son señal de *data leakage* y serán rechazados
Horizonte temporal	Predicción semestre a semestre (esquema t→t+1). El modelo no predice mora a largo plazo, sino incumplimiento en el próximo período semestral

La segunda y tercera fase de CRISP-DM comprenden, respectivamente, la comprensión inicial de los datos disponibles y el análisis exploratorio. Su contenido sustantivo (los resultados del análisis sobre las fuentes de la Universidad) se desarrolla en el Capítulo 5 del proyecto, mientras que aquí se documentan las decisiones metodológicas que rigen cómo se hizo ese análisis. El análisis exploratorio se estructuró en tres niveles progresivos, cada uno con objetivos metodológicos distintos y herramientas estadísticas específicas, como resume la Tabla 4.3.

Tabla 4.3: Estructura metodológica del análisis exploratorio de datos

Nivel	Denominación	Herramientas estadísticas	Preguntas que responde y decisiones que fundamenta
Nivel 1	Análisis univariado	Frecuencias absolutas y relativas, estadísticos descriptivos (mediana, rango intercuartílico (IQR), asimetría), histogramas, *boxplots*	¿Cómo se distribuye cada variable individualmente? ¿Cuál es el perfil típico del beneficiario? ¿Existen valores extremos que deban tratarse? ¿Cuál es la prevalencia de valores faltantes? Fundamenta: decisiones de imputación, winsorización, exclusión de variables.
Nivel 2	Análisis bivariado	*Test* chi-cuadrado (χ^2) para categóricas, variables numéricas, intervalos de confianza (IC) Clopper-Pearson 95% para proporciones, correlación de Pearson	¿Cuál es la relación estadística entre cada variable y el *target* (En_Mora)? ¿Cuáles variables presentan diferencias significativas entre grupos de mora? Fundamenta: selección de predictores candidatos, detección de *data leakage*, orientación de feature engineering.
Nivel 3	Análisis multivariado	Matriz de correlación de Pearson, V de Cramér para pares categóricos, heatmaps de interacción Estrato × Sexo	¿Existen correlaciones altas entre predictores que generen multicolinealidad? ¿Hay interacciones relevantes entre variables? Fundamenta: reducción de redundancia en el feature set, diseño de términos de interacción.

El primer nivel es un análisis univariado que usa frecuencias absolutas y relativas, estadísticos descriptivos como mediana, rango intercuartílico y asimetría, junto con histogramas y *boxplots*. Responde preguntas sobre la distribución individual de cada variable, el perfil típico del beneficiario, la existencia de valores extremos y la prevalencia de datos faltantes. Sirve de base para las decisiones de imputación, winsorización y exclusión de variables. El segundo nivel es un análisis bivariado que utiliza el *test* chi-cuadrado para variables categóricas, el *test* de *Mann-Whitney U* para numéricas e intervalos de confianza exactos de Clopper-Pearson al 95% para proporciones; con esto se evalúa la relación estadística entre cada variable y el *target*, se detecta *data leakage* y se orienta la ingeniería de características. El tercer nivel es un análisis multivariado basado en la matriz de correlación de Pearson, la *V de Cramér* para pares categóricos y *heatmaps* de interacción; identifica multicolinealidad entre predictores e interacciones relevantes como la de Estrato por Sexo. La elección del *test* de *Mann-Whitney* sobre la prueba *t* de Student se justifica por la asimetría marcada de las distribuciones financieras (con asimetría superior a +20 en variables como **Saldo_nota_debito**), que viola el supuesto de normalidad. El intervalo de Clopper-Pearson se prefiere sobre el de Wald para muestras desbalanceadas donde la tasa de mora es pequeña.

El proyecto se diseñó para ser claro y fácil de entender. Se basó en criterios generales de claridad y comparación para que los resultados sean fáciles de interpretar [12]. El objetivo era que las figuras y el tablero sean comprensibles para personas que no tienen conocimientos técnicos. Para lograr esto, se utilizaron gráficos simples e indicadores ejecutivos. También se utilizó un sistema de colores tipo semáforo para ayudar a entender el riesgo crediticio. Esto permite a los usuarios leer y entender los resultados de manera sencilla.

La cuarta fase de CRISP-DM, dedicada a la preparación de datos, constituye el corazón técnico del proyecto desde el punto de vista metodológico. Es en ella donde se toman las decisiones más críticas: qué variables incluir, cómo transformarlas, cómo dividir los datos y cómo tratar el desbalanceo. Una sola decisión incorrecta (como incluir una variable con *data leakage*) puede producir un modelo aparentemente excelente que resulta completamente inútil en producción.

El *data leakage*, o fuga de información del futuro al pasado, es el error metodológico más frecuente y dañino en proyectos de modelado predictivo aplicado. Ocurre cuando una variable refleja información que en el momento real de hacer la predicción no estaría disponible, y hace que el modelo aprenda a hacer trampa con respuestas del examen y falle al aplicarse en el mundo real. Para prevenirlo se implementó un protocolo de control que excluye explícitamente las variables cuyo valor al inicio del semestre es desconocido o igual para todos los estudiantes.

La contribución de ingeniería de características más importante del proyecto es el esquema de *features* históricas *t→t+1*. Su objetivo es convertir el historial de comportamiento de pago del estudiante en señales predictivas válidas sin caer en *data leakage*. En lugar de usar el recaudo y saldo del semestre actual (que serían fuga de información), se construyen variables de comportamiento histórico a partir de semestres anteriores: si el estudiante tuvo mora en el semestre *t-1*, cuántas veces fue a mora en los últimos n semestres y cuál fue su patrón de pagos previo. Esas variables son prospectivamente válidas porque en el semestre *t* se conoce lo que ocurrió en *t-1*.

La división del *dataset* en conjuntos de entrenamiento y prueba es otra decisión metodológica determinante. En riesgo crediticio, un *split* aleatorio deja que registros de 2024 caigan en entrenamiento y registros de 2016 en prueba, lo que contamina la evaluación con información temporal cruzada. Por eso se adoptó un *split* temporal con corte en 2022: los datos hasta 2022 se usan para entrenar y los de 2023 en adelante para evaluar. Esto reproduce el escenario real en el que el modelo se entrena con historia pasada y se aplica a créditos futuros. El diseño es especialmente importante dado el *drift* temporal documentado por la pandemia.

El *dataset* de modelado presenta una distribución marcadamente desbalanceada: el 94.82% de los registros corresponde a créditos sin mora y solo el 5.18% a créditos en mora, con una proporción de 18 a 1. Este desbalanceo refleja la realidad del portafolio: el crédito educativo fue diseñado para tener tasas de mora bajas en condiciones normales, pero tiene consecuencias metodológicas que deben gestionarse. SMOTE genera muestras sintéticas de la clase minoritaria mediante interpolación entre casos reales existentes. Para cada muestra de la clase en mora identifica sus cinco vecinos más cercanos y genera una nueva muestra en un punto aleatorio del segmento que une la muestra original con uno de esos vecinos. A diferencia de la duplicación simple de casos, SMOTE produce variedad en la clase minoritaria y reduce el riesgo de sobreajuste. Es esencial que SMOTE se aplique exclusivamente sobre el conjunto de entrenamiento y nunca sobre el de prueba, pues de lo contrario los resultados de evaluación serían artificialmente optimistas.

Con los datos preparados, la quinta fase aborda la selección y configuración de los algoritmos de aprendizaje supervisado. Se decidió entrenar cuatro algoritmos simultáneamente en lugar de uno solo, lo que deja comparar de manera objetiva el desempeño y seleccionar el modelo final con fundamento. La estrategia de comparación multialgoritmo es consistente con la adoptada por Betancourt [36], quien compara regresión logística, *Random Forest* y *Gradient Boosting* para riesgo crediticio, y por Villanueva [35], quienes evalúan redes neuronales, árboles de decisión y

máquinas de vectores de soporte en el contexto colombiano. Cada algoritmo fue elegido porque cumple una función específica en el conjunto, y juntos cubren el espectro completo desde la interpretabilidad máxima hasta el desempeño predictivo máximo, como detalla la Tabla 4.4.

Tabla 4.4: Justificación metodológica de los cuatro algoritmos de clasificación

Algoritmo	Rol en el conjunto	Hiperparámetros clave	Fortalezas para este problema	Limitaciones y consideraciones
Regresión Logística	*Baseline* interpretable	**C=0.1 (post-tuning Grid) penalty='l2', solver='lbfgs' max_iter=2000 class_weight='balanced' random_state=42**	Modelo de referencia en *credit scoring* desde hace décadas Mester [37]. Produce probabilidades calibradas. Coeficientes directamente interpretables como *log-odds*. Robusto a datos de entrenamiento limitados.	Asume linealidad en el espacio *log-odds*. No captura interacciones no lineales. La relación no lineal estrato-mora (sección 5.8.3) puede no ser capturada eficientemente.
Árbol de Decisión	Modelo de reglas explícitas	**max_depth=6 min_samples_leaf=80**	Máxima interpretabilidad: produce reglas IF-THEN legibles por humanos. Captura relaciones no lineales y umbrales. Ideal para presentar a *stakeholders* no técnicos de la Dirección Financiera.	Alta varianza: sensible al sobreajuste si max_depth no se controla. Menor desempeño predictivo que los métodos de ensamble. Se usa como complemento, no como modelo final.
Random Forest	Ensamble robusto	**n_estimators=200 max_depth=8 min_samples_leaf=50**	Promedia 200 árboles de decisión con submuestreo aleatorio de variables (bagging). Reduce la varianza del árbol simple. Muy robusto al sobreajuste. Proporciona importancia de variables como media de ganancias de información.	Menor interpretabilidad que el árbol simple. Más costoso computacionalmente. El parámetro n_jobs=1 se usa para garantizar reproducibilidad exacta con SEMILLA=42.
Gradient Boosting	Modelo de máximo desempeño	**n_estimators=200 max_depth=4 learning_rate=0.05**	Construye árboles secuencialmente, donde cada árbol corrige los errores del anterior (boosting). Típicamente el algoritmo con mejor AUC en tabular data. La tasa de aprendizaje baja (0.05) con muchos árboles reduce el riesgo de sobreajuste.	Mayor riesgo de sobreajuste si no se controlan los hiperparámetros. Sensible a outliers. Los parámetros adoptados son conservadores para un *dataset* de tamaño medio.

La Regresión Logística actúa como *baseline* interpretable y es el modelo de referencia en *credit scoring* desde hace décadas. Produce probabilidades calibradas con coeficientes interpretables como *log-odds*, aunque asume linealidad y no captura interacciones no lineales. El Árbol de Decisión funciona como modelo de reglas explícitas con máxima interpretabilidad; genera reglas IF-THEN legibles por humanos, ideales para presentar a los *stakeholders* no técnicos de la Dirección Financiera. Se configura con profundidad máxima de seis niveles y al menos ochenta muestras por hoja para controlar el sobreajuste. *Random Forest* es el ensamble robusto que promedia doscientos árboles de decisión con submuestreo aleatorio de variables; reduce la varianza del árbol simple y entrega importancia de variables como media de ganancias de información. *Gradient Boosting* representa el modelo de máximo desempeño y construye árboles secuencialmente donde cada uno corrige los errores del anterior. Se configuró con una tasa de aprendizaje conservadora de 0.05 y doscientos estimadores para reducir el riesgo de sobreajuste en un *dataset* de tamaño medio.

La reproducibilidad se garantiza con una semilla global (SEMILLA = 42) que se pasa a todos los componentes del proceso: la generación de números aleatorios con NumPy, la división *train-test*, SMOTE y cada uno de los cuatro algoritmos. El valor 42 es una convención adoptada en la comunidad de Ciencia de Datos y aprendizaje automático, popularizada por su uso en la documentación oficial de *scikit-learn*, en los ejemplos de Kaggle y en textos de referencia como Hands-On *Machine Learning*. Su elección no obedece a un criterio estadístico (ninguna semilla produce resultados sistemáticamente superiores a otra en un generador pseudoaleatorio bien implementado), sino a un principio de transparencia y trazabilidad: al usar un valor reconocible por la comunidad, cualquier investigador que revise el código identifica de inmediato que se trata de una decisión deliberada de reproducibilidad y no de un valor seleccionado tras múltiples pruebas para favorecer un resultado particular (práctica conocida como *seed hacking*). Lo que importa metodológicamente no es el valor específico de la semilla, sino que sea idéntico en todos los componentes del *pipeline* y que se declare de forma explícita, para que cualquier investigador que ejecute el mismo código con los mismos datos obtenga exactamente los mismos resultados.

La evaluación de los modelos de la sexta fase va más allá del cálculo de métricas sobre el conjunto de prueba. Se implementa un marco de validación multidimensional que examina el desempeño desde cuatro perspectivas complementarias, cada una responde a una pregunta diferente sobre la calidad y utilidad del modelo. En un *dataset* con desbalanceo de 18 a 1, la elección de métricas no es trivial. Un clasificador trivial que predijera siempre "sin mora" obtendría una exactitud del 94.82% sin haber aprendido nada, así que la exactitud no se usa como criterio de selección. En su lugar se priorizan el *Recall* para la clase positiva (qué fracción de los morosos reales identificó el modelo), el área bajo la curva ROC (AUC-ROC) como métrica de discriminación global y el *F1-score* como balance entre precisión y *recall* [24],[38]. En cobranza, fallar en detectar un moroso tiene un costo operativo mucho mayor que una alerta falsa. El umbral óptimo de clasificación se determina mediante el índice de Youden J, y se incluye una validación *out-of-time* para verificar la estabilidad del modelo en períodos no vistos durante el entrenamiento.

CRISP-DM no termina con la evaluación. La séptima fase de despliegue exige pensar cómo los resultados se traducen en herramientas y procesos que la organización pueda usar. En un proyecto aplicado, el despliegue es conceptual, pero debe ser concreto al punto de que el equipo de TI y el Departamento de Apoyo Financiero puedan implementarlo sin rediseñar el sistema. El diseño comprende cuatro componentes articulados.

El primero es un sistema de segmentación por zonas de riesgo tipo semáforo, donde el score de probabilidad del modelo se traduce en cuatro zonas operativas con acciones diferenciadas de intervención. La Tabla 4.5 presenta el diseño de esa segmentación.

Tabla 4.5: Sistema de segmentación por zonas de riesgo: umbrales, acciones operativas y costos estimados de intervención

Zona	Rango de score	Acción operativa	Canal	Costo estimado por estudiante

●	**Verde**	0 – 30%	**Monitoreo sin intervención activa**	Ninguno	$0
●	**Amarillo**	30 – 60%	**Recordatorio preventivo automático**	SMS / correo electrónico	$500
●	**Naranja**	60 – 80%	**Contacto telefónico con propuesta de plan de pagos**	Llamada del asesor financiero	$5,500
●	**Rojo**	80 – 100%	**Gestión presencial con el estudiante y el codeudor**	Visita o citación presencial	$30,000

Nota. Los costos unitarios son estimaciones institucionales del Departamento de Apoyo Financiero que incluyen tiempo del asesor, infraestructura tecnológica y gastos operativos asociados a cada canal. Estos valores se utilizan en el análisis costo-beneficio del Capítulo 8.

La lógica del semáforo es progresiva: a medida que el score de riesgo aumenta, la intensidad y el costo de la intervención también lo hacen, y se reservan los recursos más costosos (gestión presencial) para los estudiantes con mayor probabilidad de incumplimiento. El diseño deja al Departamento de Apoyo Financiero asignar su presupuesto de cobranza con eficiencia, concentrando los esfuerzos donde el modelo predice mayor riesgo.

El segundo componente es una priorización por deciles mediante curva de ganancia que ordena los créditos activos de mayor a menor score de riesgo a nivel obligación, los divide en diez grupos iguales y responde la pregunta clave de operaciones: si la Universidad interviene al X% de los créditos de mayor riesgo, ¿qué porcentaje de los morosos reales captura? Esa priorización luego se traduce a nivel estudiante para el análisis costo-beneficio. El tercer componente es un tablero interactivo que integra la distribución por zonas de riesgo, el score promedio por *cluster*, la curva de ganancia acumulada y el ranking de las cinco variables más importantes; está diseñado para ser operado por analistas financieros sin conocimientos de programación [39]. El cuarto componente es un análisis costo-beneficio de la cobranza preventiva que cuantifica el retorno de inversión de intervenir en cada zona, comparando el costo unitario de intervención con la mora potencialmente mitigada esperada bajo un supuesto conservador de efectividad del 30 al 50%.

El uso de modelos predictivos para evaluar el riesgo crediticio de estudiantes universitarios plantea consideraciones éticas que este proyecto reconoce de manera explícita. En materia de confidencialidad, los datos del Departamento de Apoyo Financiero son de carácter interno y se emplean exclusivamente con fines académicos; no contienen nombres ni documentos de identidad, pues el campo Cliente es un código numérico interno sin identificación directa. Sobre la equidad del modelo, el uso del estrato socioeconómico como predictor podría generar discriminación si el modelo fuera más estricto con estudiantes de estratos bajos. Sin embargo, el análisis bivariado del Capítulo 5 muestra que la relación estrato-mora no es un gradiente lineal simple. Los estratos 3-4 presentan mayor mora que los estratos 1-2, lo que sugiere que el modelo no penaliza necesariamente más a los estudiantes de menor ingreso, aunque cualquier implementación real debe monitorear de manera regular las métricas de equidad entre grupos demográficos. El modelo produce un *score* de riesgo que debe ser una entrada para la decisión del asesor financiero, no un reemplazo de esa decisión: un *score* alto no debe significar

automáticamente la denegación del crédito, sino la asignación de un plan de acompañamiento financiero, un acuerdo de pago escalonado o una conversación con el estudiante sobre sus circunstancias. La institución tiene responsabilidad ética de usar el modelo para apoyar y proteger al estudiante, no para excluirlo financieramente. Esta interpretación es coherente con el análisis cuantitativo de *fairness* presentado a continuación: la tasa de falsos positivos (FPR) decreciente con el nivel socioeconómico (58,69% en estrato 1 vs 14,39% en estrato 6) y la tasa de mora más baja en estratos altos (1,62% en estrato 6 vs 6,47% en estrato 3) reflejan diferencias estadísticas presentes en los datos históricos, no un castigo arbitrario del algoritmo.

Para complementar la discusión cualitativa de equidad, se computó un análisis cuantitativo de *fairness* sobre el conjunto de prueba (n=26.977), con la FPR y los Falsos Negativos (FNR) por grupo demográfico. En sexo, la FPR fue de 42,85% para mujeres y 52,47% para hombres, mientras que la FNR fue similar (14,89% vs 15,26%). Eso sugiere que el modelo tiende a generar más alertas falsas en hombres, pero no presenta diferencia material en la detección efectiva. Por estrato, la FPR decrece con el nivel socioeconómico (58,69% en estrato 1 vs 14,39% en estrato 6), patrón que se explica por la mayor prevalencia de morosidad en estratos bajos del histórico de entrenamiento. El modelo no aplica un castigo arbitrario por estrato; refleja la diferencia estadística observada en los datos. Estos resultados se presentan en el output del *notebook* (celda *fairness*_analysis.png) y deben monitorizarse periódicamente para detectar *drift* en el comportamiento del modelo respecto a grupos sensibles.

En síntesis, este capítulo ha presentado el recorrido metodológico completo del proyecto siguiendo las siete fases de CRISP-DM como un proceso integrado y coherente. Desde la selección justificada del marco de trabajo y la documentación de la arquitectura de datos, pasando por la comprensión del negocio del crédito educativo, el diseño del análisis exploratorio en tres niveles, la preparación de datos con protocolo *anti-leakage* y esquema *t→t+1*, la selección y configuración de cuatro algoritmos de clasificación, la evaluación multidimensional y el despliegue conceptual en herramientas operativas, cada decisión se ha fundamentado en las características específicas del contexto institucional y en los principios de reproducibilidad, equidad y rigor estadístico que exige un proyecto aplicado de maestría. La Tabla 4.6 presenta el mapa metodológico completo que integra las fases CRISP-DM con los cuatro objetivos específicos del proyecto, las decisiones técnicas clave.

Tabla 4.6: Mapa metodológico completo: CRISP-DM × Objetivos Específicos × Decisiones técnicas

Fase CRISP-DM	Denominación	OE vinculado	Decisiones metodológicas clave
Fase 0	Configuración del entorno	Transversal	SEMILLA=42 global.

Fase 1	Comprensión del negocio	OG + todos los OE	Definición del problema como clasificación binaria. Criterios de éxito (AUC≥0.65, *Recall*≥0.60, F1≥0.40). Restricción de prospectividad.
Fase 2	Comprensión de los datos	OE1	Análisis de dos fuentes (df1: 191,130×25; df2: 18,452×22). Validación del *target* por cruce df1×df2. Análisis de nulos (máx. 40.2% en Departamento).
Fase 3	EDA (3 niveles)	OE1	EDA univariado, bivariado (χ^2, MW-U, Clopper-Pearson), multivariado (Pearson, Cramér). Principios Munzner. Detección de *drift* temporal 2015–2025.
Fase 4	Preparación de datos	OE2	Protocolo anti-*leakage* (7 variables excluidas). Feature engineering t→t+1 (3 *features* históricas). Split temporal corte 2022. SMOTE SOLO en *train*. StandardScaler fit SOLO en *train*.
Fase 5	Modelado predictivo	OE3	Entrenamiento de 4 modelos (RL, DT, RF, GB). Hiperparámetros documentados. Todos con SEMILLA=42.
Fase 6	Evaluación multidimensional	OE3	5 dimensiones de validación. Umbral óptimo Youden J. Validación OOT. SHAP. Curvas ROC, ganancia, lift.
Fase 7	Despliegue conceptual	OE4	4 componentes: segmentación semáforo, deciles de cobranza, tablero Plotly HTML, análisis costo-beneficio.

Nota. *Objetivo General. OE1–OE4: Objetivos Específicos 1 a 4. RL: Regresión Logística. DT: Árbol de Decisión. RF: Random Forest. GB: Gradient Boosting. El vínculo entre las fases CRISP-DM y los objetivos específicos garantiza que cada decisión metodológica tiene un propósito académico explícito y contribuye a la respuesta a las preguntas de sistematización del proyecto.*

5. CARACTERIZACIÓN DE LOS BENEFICIARIOS DE CRÉDITO EDUCATIVO Y FACTORES ASOCIADOS AL RIESGO DE INCUMPLIMIENTO

Objetivo específico 1: Analizar las variables académicas, socioeconómicas y de comportamiento de los estudiantes beneficiarios de créditos educativos, identificando aquellas asociadas al riesgo de incumplimiento.

Este capítulo presenta la caracterización de los beneficiarios del crédito educativo y los factores asociados al riesgo de incumplimiento. Se describen las dos fuentes de información utilizadas, la construcción y validación de la variable objetivo, y los resultados del análisis exploratorio a nivel univariado, bivariado y multivariado. El análisis se realizó sobre los datos proporcionados por el Departamento de Apoyo Financiero, procesados con Python en el *notebook* que acompaña al proyecto.

5.1 Fuentes de datos

El proyecto se desarrolló sobre dos conjuntos de datos complementarios, cada uno aporta una perspectiva distinta del ciclo de vida del crédito educativo. A continuación, se describe cada uno en detalle.

El primer conjunto, identificado como *Dataset* 1 (df1), contiene el registro histórico de las notas débito o transacciones crediticias individuales generadas por el sistema de crédito educativo entre 2015-1S y 2025-2S. Cada fila del *dataset* representa una cuota o cargo individual asociado a un crédito otorgado a un estudiante en un período académico específico.

El *dataset* original (df1) contiene 191,130 registros y 25 variables, correspondientes a 14,087 estudiantes únicos. En promedio, cada estudiante tiene asociadas aproximadamente 13.6 transacciones a lo largo de su historial crediticio con la institución.

En cuanto a la cobertura temporal, el *dataset* abarca 22 períodos académicos desde 2015-1S hasta 2025-2S (11 años). No todos los registros son aptos para entrenar un modelo predictivo. Los créditos de los períodos más recientes presentan un problema metodológico conocido como censura derecha (*right-censoring*): un crédito otorgado en 2025-2S que aparece como CARTERA en el sistema no necesariamente es un crédito moroso, sino que puede ser un crédito activo cuyas cuotas aún no han vencido. Clasificarlo como "en mora" sería incorrecto y contaminaría el análisis. Para resolver el problema se aplicaron dos filtros sucesivos.

El primero es un filtro de maduración con ventana cerrada y 90 días de gracia: solo se incluyen créditos cuya fecha de vencimiento sea anterior a la fecha de corte (fecha del análisis menos 90 días). Con eso se garantiza que cada crédito haya tenido oportunidad real de ser pagado o de entrar en mora. El filtro no excluye períodos completos, sino créditos individuales cuya última cuota aún no ha vencido. Por ejemplo, de los créditos de 2024-2S, los de corto plazo sí se incluyen

porque su ciclo de pago ya concluyó, mientras que las obligaciones que aún no han completado su ciclo de vencimiento se excluyen temporalmente del modelado, independientemente de la modalidad del crédito, porque su desenlace aún es desconocido. En total, el filtro excluyó 7,722 registros y redujo el *dataset* a 183,408 registros.

El segundo filtro es la exclusión de los *vintages* de 2025 (14,938 créditos de 2025-1S y 2025-2S), que no han completado su ciclo de vida y por tanto no tienen un desenlace definitivo. Esos créditos no se eliminan del proyecto; son la población objetivo del modelo. Son los candidatos a recibir el *score* predictivo para gestión preventiva de cobranza, es decir, los créditos sobre los que el Departamento de Apoyo Financiero aplicará las recomendaciones del sistema. Tras los dos filtros, el *dataset* de modelado queda conformado por 167,275 registros correspondientes al período 2015-2024, con una tasa de mora del 5.18%. Esta decisión es coherente con las buenas prácticas en *credit scoring*, donde solo se modelan créditos con desenlace conocido y maduro.

La Tabla 5.1 presenta el diccionario completo de las variables más relevantes del *Dataset* 1, con el tipo de dato esperado, su rol en el modelo y el porcentaje de valores nulos. Esta clasificación recoge los resultados del análisis de calidad de datos y de la revisión hecha para excluir variables con riesgo de fuga de información.

Tabla 5.1: Diccionario de variables del *Dataset* 1- Solicitudes de Crédito *Dataset*1 (df1)

	Variable	Tipo esperado	Uso en el modelo	Nulos (%)
0	Cliente	**ID**	Agrupación	**0.0**
1	ID_Estudiante	**ID**	Agrupación	**0.0**
2	Sexo	**Categórica**	Feature	**0.0**
3	Estrato_Socioeconomico	**Ordinal**	Feature	**14.8**
4	Estado_Estudiante	**Categórica**	Feature	**8.8**
5	Primiparo	**Binaria**	Feature	**8.8**
6	Graduado	**Binaria**	Excluida (*leakage*)	**8.8**
7	No_Creditos	**ID secuencial**	Excluida (es ID)	**0.0**
8	Descripcion_linea_credito	**Categórica**	Feature derivada	**0.0**
9	Nota_debito	**ID**	Excluida (ID)	**0.0**
10	Valor_nota_debito	**Numérica continua**	Feature	**0.0**
11	Cuotas	**Numérica discreta**	Feature	**0.0**
12	Periodo	**Temporal**	Feature derivada	**0.0**
13	Nombre_centro_costo	**Categórica**	Feature derivada	**0.0**
14	Estado_describe	**Categórica (*target*)**	*Target*	**0.0**
15	Valor_financiacion	**Numérica continua**	Feature	**0.0**
16	Fecha_vencimiento_ndb	**Fecha**	Filtro de maduración	**0.0**

El segundo conjunto de datos *Dataset* 2, denominado internamente **df2**, corresponde a una fotografía de la cartera crediticia activa de la Universidad al momento de la extracción de la información. A diferencia de **df1**, que tiene perspectiva histórica completa, **df2** refleja únicamente el estado de los créditos que aún tienen saldo vigente en el sistema, sea porque están al día o

porque están en mora.

Este conjunto contiene 18,452 registros y 22 variables, correspondientes a 2,662 estudiantes distintos. El saldo total de la cartera registrada asciende a $20.291.181.030, de los cuales $5.339.145.302 (26.3%) corresponden a cartera vencida con mora activa. Al igual que **df1**, este *dataset* cubre los 22 períodos académicos desde 2015-1S hasta 2025-2S. La distribución es marcadamente no uniforme: los períodos más recientes concentran la mayor cantidad de registros (2025-2S con 6,820 registros y 2025-1S con 3,034), lo que refleja que la cartera activa vigente se concentra en créditos recientes que aún no han vencido o que acaban de entrar en mora. La Tabla 5.2 presenta el diccionario de las 8 variables del *Dataset* 2. Este conjunto no se usa directamente para entrenar el modelo, sino para validar la construcción de la variable objetivo y para enriquecer el análisis descriptivo con información sobre la severidad y antigüedad de la mora.

Tabla 5.2: Diccionario de variables del Conjunto de Datos 2 — Cartera Depurada Vigente *Dataset*2 (df2)

	Variable	Tipo esperado	Descripción	Uso en el proyecto	Nulos (%)
0	Nota_Debito	**ID**	Identificador del crédito (clave de cruce con *Dataset* 1)	**Clave de cruce con df1**	**0.0**
1	Cliente	**ID**	Identificador del estudiante	**Agrupación**	**0.0**
2	Saldo	**Numérica continua**	Saldo pendiente del crédito a la fecha del corte	**Cuantificación del riesgo financiero**	**0.0**
3	Dias	**Numérica discreta**	Días de mora (negativo = días por vencer, positivo = días vencido)	**Validación del *target* En_Mora**	**N/A**
4	Edad_Cartera	**Categórica ordinal**	Clasificación de la cartera por antigüedad de la mora	**Construcción de bandas de severidad**	**0.0**
5	Genera Mora	**Binaria**	Indicador de si el crédito genera intereses de mora (Sí/No)	**Validación cruzada de mora**	**0.0**
6	Cuotas_Pendientes	**Numérica discreta**	Número de cuotas aún pendientes de pago	**Caracterización de la deuda**	**N/A**
7	Valor_Cuota	**Numérica continua**	Valor de la cuota periódica del crédito	**Caracterización de la deuda**	**N/A**
8	Fecha_Vencimiento	**Fecha**	Fecha de vencimiento de la siguiente cuota	**Referencia temporal**	**N/A**

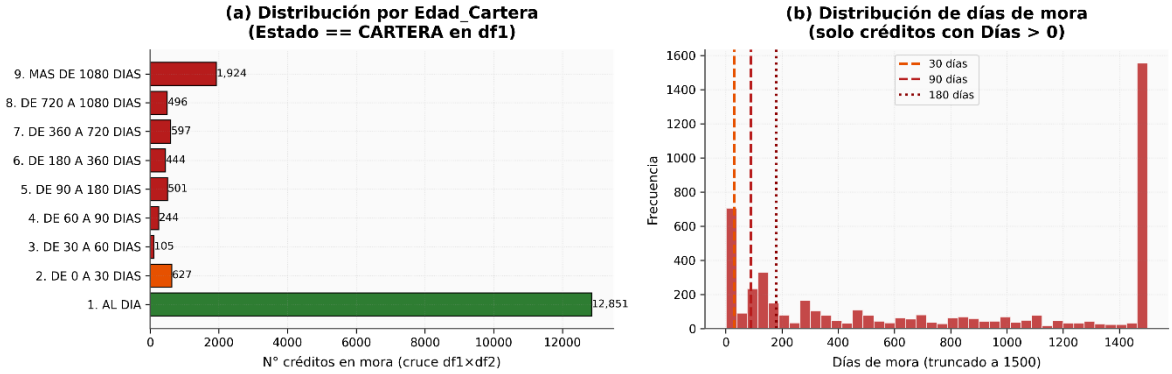
5.2 Estrategia de Integración de las Fuentes

La integración de las dos fuentes de datos cumple un doble propósito. Sirve para validar que la clasificación administrativa CARTERA, usada para construir la variable objetivo, corresponde efectivamente a créditos con mora activa. Y enriquece el análisis descriptivo con información de severidad y antigüedad de la deuda. Este cruce se hizo con el campo **Nota Débito**, que actúa como llave común única entre ambas tablas.

De los 23,572 registros clasificados como CARTERA en el primer *dataset*, se encontraron 17,789

en el segundo (75.5%). El 24.5% restante no aparece porque esos créditos ya fueron recuperados, castigados contablemente o reclasificados antes de la extracción del segundo *dataset*. El comportamiento es esperado: el segundo *dataset* representa la cartera activa en un momento dado, mientras que el primero conserva el historial completo, incluidos créditos ya cerrados.

Esta validación respalda la definición operativa utilizada para clasificar los créditos como incumplimiento. El cruce del 75,5% entre CARTERA (df1) y el *snapshot* de cartera activa (df2), una mediana de 697 días y un promedio de 1.042 días de mora entre los efectivamente vencidos, junto con un 19,5% de créditos por encima de los 180 días, son consistentes con que los casos clasificados como **En_Mora**=1 corresponden a obligaciones con riesgo efectivo de incumplimiento. El 24,2% con Días ≥ 30 en el *snapshot* no debe interpretarse como falsos morosos: el resto incluye créditos *re-aged* (volvieron al día tras una mora previa), créditos con intereses de mora ya cobrados y casos de mora incipiente (1–29 días).



La Figura 5.1 muestra los resultados de esta validación cruzada, confirmando que la clasificación administrativa CARTERA corresponde efectivamente a créditos con mora activa en el sistema de cartera vigente.

El cruce entre ambas fuentes deja analizar la distribución de la cartera por bandas de severidad de la mora. La Tabla 5.3 presenta esa distribución, donde se observa que la mediana de días de mora para los créditos efectivamente vencidos es de 697 días, lo que indica que una proporción significativa de la cartera morosa corresponde a obligaciones con deterioro prolongado.

Tabla 5.3: Distribución de la cartera activa por bandas de severidad de mora

Banda_Mora	N_creditos	Dias_Promedio	Saldo_Total	Pct_Cartera
0_AL_DIA	12.851	-444	$ 14.946.503.577	72.2%
1_LEVE_1_29	627	22	$ 602.516.212	3.5%
2_TEMPRANA_30_59	105	42	$ 54.068.461	0.6%
3_INTERMEDIA_60_89	244	81	$ 353.692.192	1.4%
4_CRITICA_90_179	501	135	$ 811.652.150	2.8%
5_SEVERA_180_MAS	3.461	1456	$ 3.475.831.841	19.5%

5.3 Construcción y Validación de la Variable Objetivo

La variable objetivo es el corazón del modelo supervisado: determina qué comportamiento se busca predecir. Su definición debe ser operacionalmente precisa, reproducible a partir de los datos disponibles y metodológicamente justificada. Esta sección documenta las decisiones tomadas en cada paso de su construcción.

En el sistema financiero regulado por la Superintendencia Financiera de Colombia, la gestión del riesgo crediticio se apoya en criterios formales de clasificación de cartera y mora [3], [40]. Los créditos educativos no operan bajo ese marco regulatorio: en su mayoría son créditos internos de corto plazo, otorgados sin consulta a centrales de riesgo externas, y su ciclo de vida se resuelve mediante la clasificación administrativa del sistema contable institucional.

La variable objetivo binaria **En_Mora** se construyó a partir del campo **Estado_describe** del *Dataset* 1, con la regla de codificación que se presenta en la Tabla 5.4.

Tabla 5.4: Regla de codificación de la variable objetivo En_Mora desde el campo Estado_describe

Estado	Significado	*Target*
CANCELADO	Crédito pagado completamente dentro de los plazos	**En_Mora = 0**
CARTERA	Crédito con saldo pendiente transferido a gestión de cobro	**En_Mora = 1**
REFINANCIADO	Crédito reestructurado (estado intermedio)	**Excluido**

Una fuente crítica de sesgo metodológico en el modelado de riesgo crediticio es la inclusión de créditos cuyo plazo de vencimiento aún no ha llegado al momento del análisis. Si un crédito vence el 30 de enero de 2026 y el análisis se realiza en noviembre de 2025, ese crédito aparecerá como CANCELADO en el sistema (no ha generado mora todavía), pero su resultado final es desconocido. Incluirlo como "sin mora" introduce sesgo de censura derecha, lo que inflará artificialmente el desempeño del modelo.

Para eliminar este sesgo, se implementó un filtro de maduración con ventana cerrada y un margen de gracia de 90 días. Bajo este criterio, solo se incluyeron las obligaciones cuya `Fecha_vencimiento_ndb` fuera anterior o igual a la fecha de corte, definida como la fecha del análisis menos 90 días de gracia. Este filtro no excluye una modalidad de crédito específica ni elimina los créditos de largo plazo como categoría de análisis; únicamente retira del conjunto de modelado aquellas obligaciones cuyo ciclo de vencimiento aún no ha finalizado y, por tanto, no cuentan con un desenlace observable. En consecuencia, los créditos de largo plazo permanecen dentro del análisis cuando sus obligaciones ya han madurado, mientras que las obligaciones no vencidas, sean de corto o largo plazo, se reservan hasta contar con información suficiente para clasificar su resultado. En total, el filtro excluyó 7.722 registros de los 191.130 originales, dejando 183.408 créditos maduros. Adicionalmente, se excluyeron 1.195 registros con estado `REFINANCIADO`, por tratarse de un estado intermedio que no corresponde ni a pago definitivo ni a mora definitiva. El efecto detallado de estos filtros se resume en la Tabla 5.5.

Tabla 5.5: Efecto del filtro de ventana cerrada sobre el *dataset* de modelado

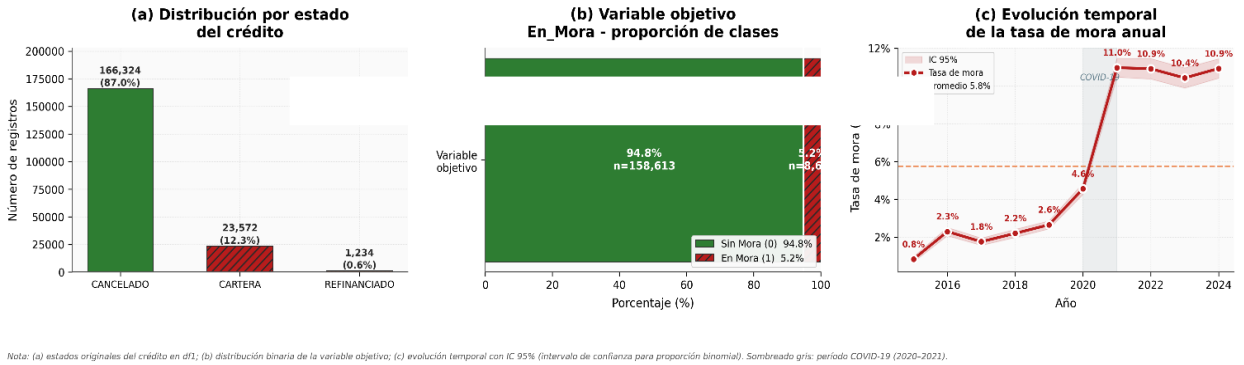
Año	Créditos	Tasa de mora	IC 95% inferior	IC 95% superior	Fase
2015	16.530	**0,8%**	0,7%	1,0%	*Baja mora*
2016	18.121	**2,3%**	2,1%	2,5%	*Baja mora*
2017	18.656	**1,8%**	1,6%	2,0%	*Baja mora*
2018	18.130	**2,2%**	2,0%	2,4%	*Baja mora*
2019	22.078	**2,6%**	2,4%	2,9%	*Baja mora*
2020	18.881	**4,6%**	4,3%	4,9%	*Ruptura COVID*
2021	15.124	**11,0%**	10,5%	11,5%	*Ruptura COVID*
2022	12.778	**10,9%**	10,4%	11,5%	*Ruptura COVID*
2023	12.771	**10,4%**	9,9%	11,0%	*Nuevo régimen*
2024	14.206	**10,9%**	10,4%	11,4%	*Nuevo régimen*

Nota. IC 95% calculado por método Wilson. Las fases se delimitan por el análisis de *drift* temporal.

Tras aplicar el filtro de ventana cerrada y excluir los registros REFINANCIADO, la distribución de la variable objetivo **En_Mora** sobre el *dataset* de modelado final (167,275 registros, tras excluir los créditos de 2025 por inmadurez de cartera) es:

Distribución de la variable objetivo **En_Mora**

Sin mora (**En_Mora** = 0): 158.613 registros (94,82%) | En mora (**En_Mora** = 1): 8.662 registros (5,18%) | Ratio de desbalanceo: 18.3:1



Nota: (a) estados originales del crédito en df1; (b) distribución binaria de la variable objetivo; (c) evolución temporal con IC 95% (intervalo de confianza para proporción binomial). Sombreado gris: período COVID-19 (2020–2021).

Figura 5.2 Distribución de la variable objetivo En_Mora, mostrando tanto la proporción de clases como su evolución temporal a lo largo del período analizado.

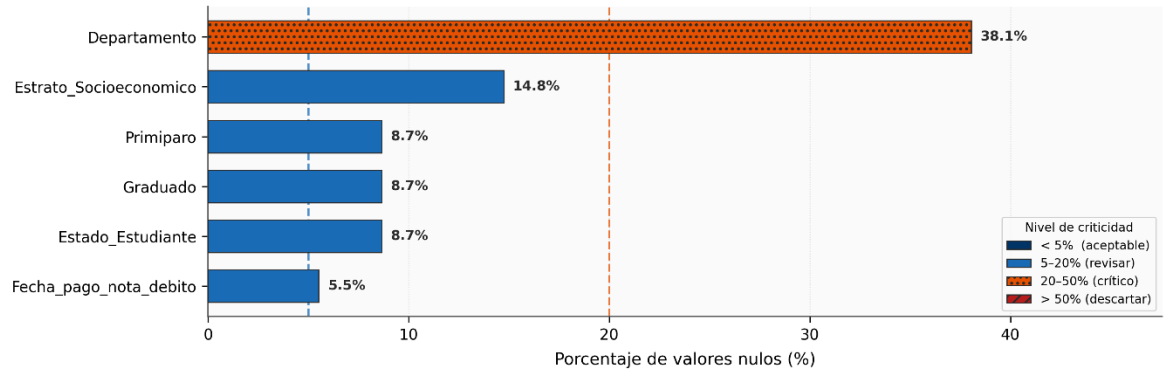
Este nivel de desbalanceo (18 créditos al día por cada uno en mora) es una característica estructural del portafolio crediticio de la Universidad, no un error en los datos. Tiene implicaciones directas en la selección de métricas y en el tratamiento de los datos para el modelado; ambos aspectos se abordan en los capítulos 6 y 7.

5.4 Análisis de Calidad de Datos

Antes del análisis exploratorio se realizó una auditoría de calidad de los datos en tres dimensiones. Esta etapa es necesaria porque los problemas de calidad no detectados y no tratados producen modelos que aprenden patrones espurios derivados de los errores en los datos, no del fenómeno real que se modela.

El análisis de duplicados confirmó que el campo **Nota_debito** es un identificador único en df1: no existe ningún par de registros con el mismo valor de **Nota_debito.** El análisis de duplicados exactos (igualdad en todas las 25 variables simultáneamente) también arrojó cero duplicados. Esa unicidad valida la integridad estructural del *dataset* y confirma que cada fila representa una transacción crediticia distinta.

Las variables financieras del portafolio de créditos presentan distribuciones marcadamente asimétricas hacia la derecha, con valores extremos que se alejan del rango típico. Es un comportamiento esperado en datos económicos: la mayoría de los créditos tienen valores dentro de un rango razonable, pero una minoría de casos (como créditos de posgrado de largo plazo con años de intereses acumulados) generan valores extremos legítimos, no errores de captura.



Nota: porcentaje de registros nulos por variable en el dataset "Solicitudes". Umbral 5%: imputable; 20%: requiere justificación metodológica; 50%: candidato a eliminar.

Figura 5.3 Análisis de valores faltantes en df1

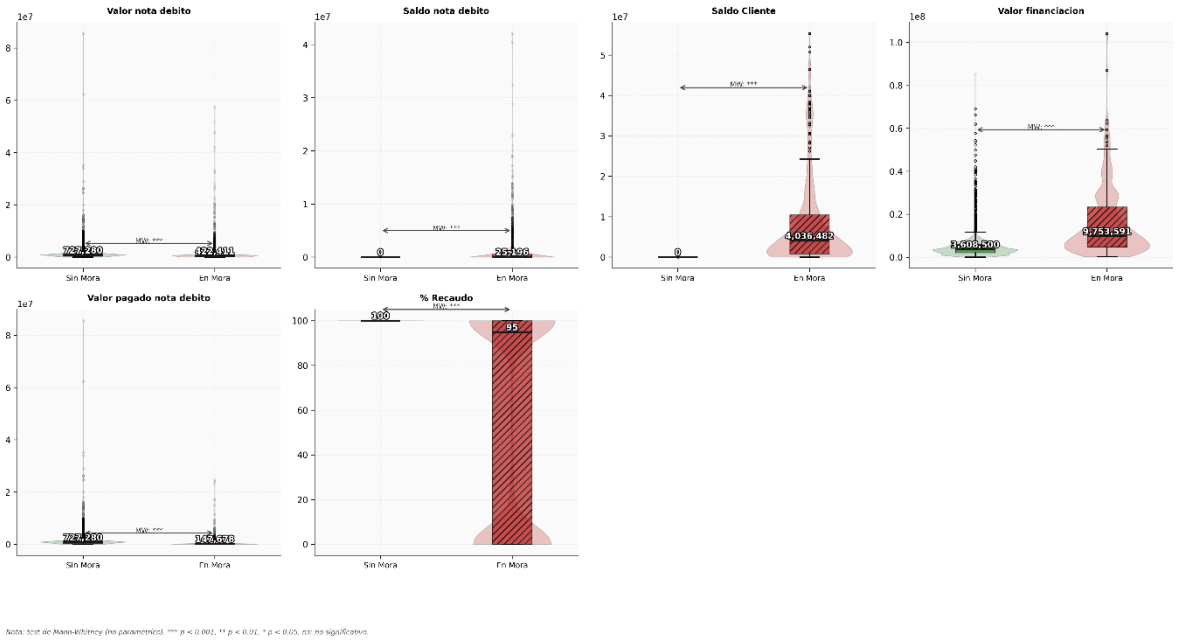


Figura 5.4 Distribución de variables financieras por estado de mora (boxplots y violines)

5.5 Evolución temporal del riesgo de mora por cohorte (2015–2024)

El análisis de cohortes agrupa los créditos por año de origen y calcula la tasa de mora para cada grupo. Este tipo de análisis longitudinal es importante en el riesgo crediticio porque deja distinguir si el nivel de incumplimiento es estable a lo largo del tiempo o si presenta tendencias que deben incorporarse en el diseño del modelo y en la interpretación de sus resultados.

La Tabla 5.6 presenta la tasa de mora por cohorte anual para el período 2015-2024. Este análisis longitudinal deja identificar cambios estructurales en el comportamiento de pago que deben considerarse en el diseño del modelo predictivo.

Tabla 5.6: Tasas de mora por cohorte anual — *Dataset* de modelado (ventana cerrada, 2015–2024)

Año	Créditos	Tasa de mora	IC 95% inferior	IC 95% superior	Fase
2015	16.530	**0,80%**	0,70%	1,00%	*Baja mora*
2016	18.121	**2,30%**	2,10%	2,50%	*Baja mora*
2017	18.656	**1,80%**	1,60%	2,00%	*Baja mora*
2018	18.130	**2,20%**	2,00%	2,40%	*Baja mora*
2019	22.078	**2,60%**	2,40%	2,90%	*Baja mora*
2020	18.881	**4,60%**	4,30%	4,90%	*Ruptura COVID*
2021	15.124	**11,00%**	10,50%	11,50%	*Ruptura COVID*

2022	12.778	**10,90%**	10,40%	11,50%	*Ruptura COVID*
2023	12.771	**10,40%**	9,90%	11,00%	*Nuevo régimen*
2024	14.206	**10,90%**	10,40%	11,40%	*Nuevo régimen*

Nota. Los créditos del año 2025 (14,938 registros) fueron excluidos del *dataset* de modelado por inmadurez de cartera: al no haber completado su ciclo de vida, su clasificación como "en mora" o "sin mora" no es definitiva. Estos créditos se reservan para *scoring* futuro una vez que maduren. La cobertura temporal del *dataset* de modelado es 2015–2024 (10 años, 20 períodos académicos).

Los datos de la Tabla 5.6 revelan tres etapas claramente diferenciadas en la historia del riesgo crediticio:

La primera etapa corresponde al período de baja mora (2015–2019). La tasa de mora se mantuvo entre 0,8% y 2,6%, con promedio del 1,9%. Este período representa la línea base del comportamiento crediticio en condiciones normales del portafolio, con crecimiento sostenido en volumen sin deterioro significativo de la calidad.

La segunda etapa corresponde a la ruptura estructural por COVID-19 (2020–2022). La tasa más que se cuadruplica en dos años: de 2,6% (2019) a 11,0% (2021). Ese salto abrupto coincide con el impacto socioeconómico de la pandemia sobre las familias de los estratos 2–4, que concentran la mayoría de los beneficiarios del crédito. El confinamiento, la pérdida de empleos informales y la contracción del ingreso familiar redujeron drásticamente la capacidad de pago.

La tercera etapa corresponde al incremento de la mora (2023–2024). La tasa no regresó a los niveles pre-pandemia y se estabilizó alrededor del 10.5%. El deterioro de la calidad crediticia del portafolio tiene un componente permanente: los efectos del COVID-19 sobre el perfil de riesgo de los beneficiarios son estructurales, no temporales. Este fenómeno de *drift* temporal justifica el diseño del split de entrenamiento/prueba por corte temporal (hasta 2022 para entrenar, 2023–2024 para evaluar).

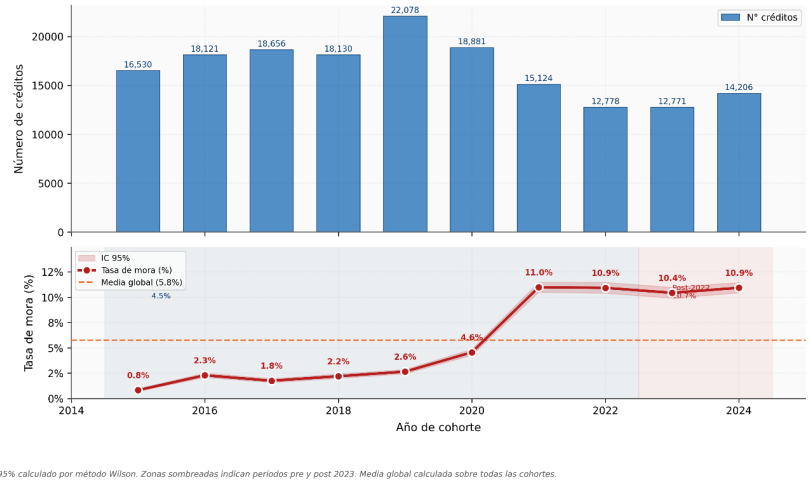


Figura 5.5 Evolución de la tasa de mora por cohorte anual con análisis de *drift* temporal (2015-2024)

5.6 Perfil univariado de los beneficiarios

El análisis univariado examina la distribución individual de cada variable para construir el perfil descriptivo de los beneficiarios del crédito educativo. Ese perfil es el punto de partida para entender a quién sirve el sistema de crédito y cuáles son las características dominantes de la población atendida.

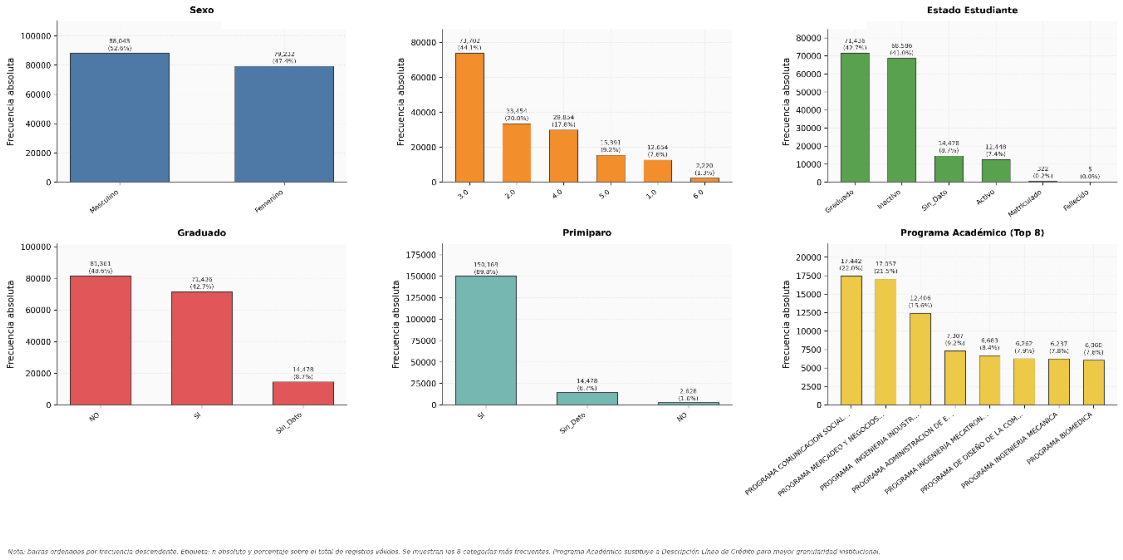


Figura 5.6 Distribución univariada de variables categóricas de los beneficiarios del crédito

5.6.1 Distribución por Sexo

El 52.6% de los beneficiarios son de sexo masculino (n=100,467) y el 47.4% femenino (n=90,663). La distribución es cercana a la equidad de género y refleja la composición general de la matrícula universitaria. A efectos del análisis de riesgo, el sexo es una variable de control demográfico cuyo efecto real sobre la mora debe evaluarse en el contexto de otras variables socioeconómicas.

5.6.2 Distribución por Estrato Socioeconómico

La distribución por estrato revela que el crédito educativo tiene un perfil marcadamente orientado a los estratos medios y medios-bajos de la pirámide socioeconómica colombiana:

Tabla 5.7: Distribución de Beneficiarios por Estrato Socioeconómico

Estrato	N	%	Acumulado	Barra
Estrato 3	56.236	**29,4%**	29,4%	
Estrato 2	38.388	**20,1%**	49,5%	
Estrato 4	34.426	**18,0%**	67,5%	
Estrato 5	16.732	**8,8%**	76,3%	
Estrato 1	14.632	**7,7%**	84,0%	

Estrato 6	2.356	**1,2%**	85,2%	
Sin información	28.360	**14,8%**	100,0%	
Total	**191.130**	**100,0%**		

Los estratos 2, 3 y 4 concentran el 67.5% de los beneficiarios. Esa concentración tiene una implicación económica importante: los estudiantes de estratos medios-bajos tienen acceso limitado a fuentes alternativas de financiamiento (no calificarían para créditos bancarios por ausencia de historial crediticio o garantías), lo que hace del crédito institucional su única opción para financiar su educación. Eso implica mayor vulnerabilidad ante choques de ingreso: una pérdida de empleo del padre o madre, una enfermedad familiar o un evento económico adverso puede interrumpir de inmediato la cadena de pagos del crédito. La escasa representación de estratos 1 (7.7%) no indica que los estudiantes de menores recursos no requieran financiamiento, sino que son atendidos por líneas de becas y subsidios antes que por crédito.

5.6.3 Estado Académico del Estudiante al Momento del Crédito

La variable **Estado_Estudiante** presenta una distribución inusual que merece análisis cuidadoso:

Tabla 5.8: Distribución por Estado Académico del Estudiante al momento del crédito

Estado	N	%	Barra
Inactivo	75.464	39,5%	
Graduado	75.360	39,4%	
Activo	23.238	12,2%	
Sin información	16.741	8,8%	
Matriculado	322	0,2%	
Fallecido	5	0,0%	
Total	**191.130**	**100,0%**	

Nota. El estado académico registrado corresponde al momento de la extracción de los datos (2025), no al momento del otorgamiento del crédito. Esto explica la alta proporción de Inactivos y Graduados.

La alta representación de estudiantes "Inactivos" (39.5%) y "Graduados" (39.4%) en un *dataset* de créditos activos se explica porque el estado académico registrado en **df1** es el estado actual del estudiante al momento de la extracción de los datos (2025), no el estado al momento en que se otorgó el crédito. Muchos créditos fueron otorgados cuando el estudiante estaba activo, pero para cuando los datos fueron exportados, ese mismo estudiante ya se había graduado o retirado. Tiene una implicación metodológica clave para el modelado: esta variable captura el estado académico en el momento del análisis, que es información posterior al otorgamiento del crédito. En el esquema predictivo *t→t+1* del modelo solo se usará el estado académico conocido en el semestre t para predecir la mora en *t+1.*

5.6.4 Líneas de crédito del sistema

El sistema de crédito educativo de la universidad opera a través de 42 líneas de crédito diferenciadas por beneficiario, nivel educativo, plazo y condiciones financieras. Las más representativas son:

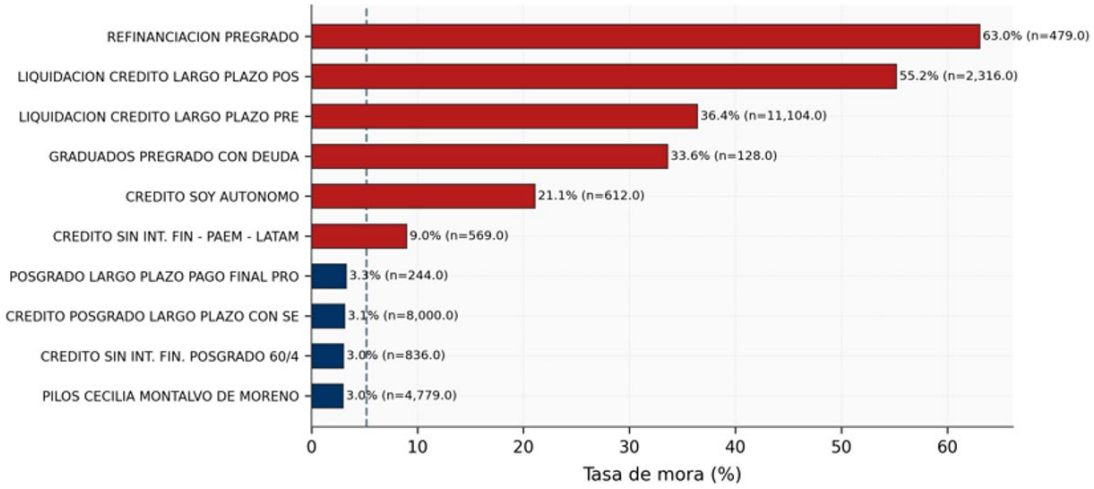


Figura 5.7 Principales líneas de crédito por volumen y tasa de mora observada

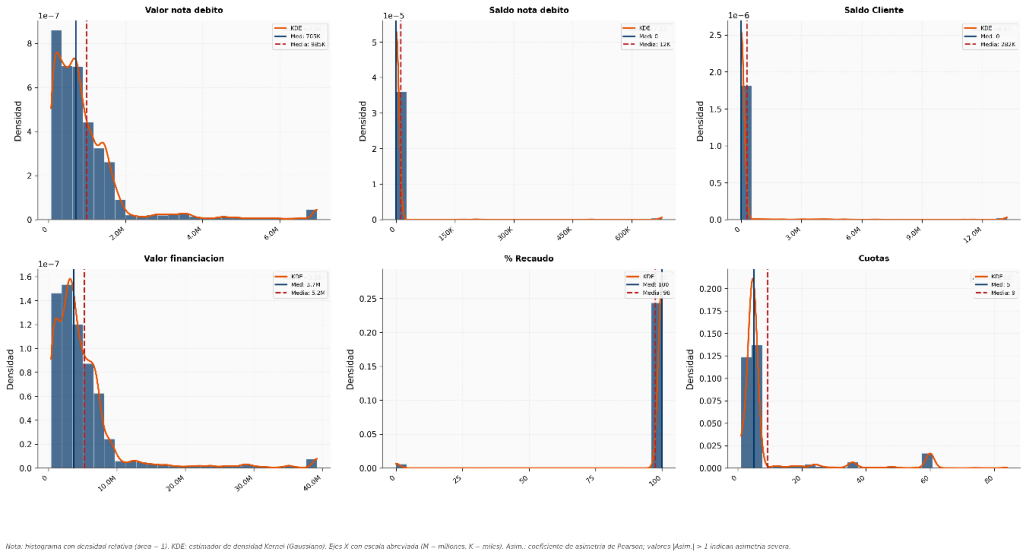


Figura 5.8 distribución de variable numéricas

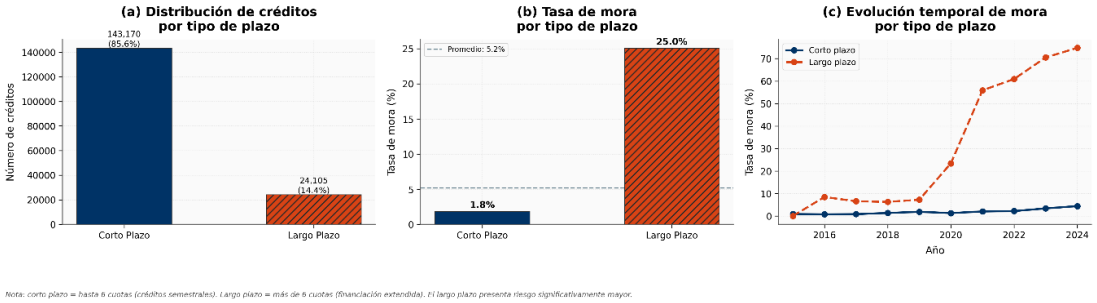


Figura 5.9 Análisis de mora por tipo de plazo: corto plazo vs largo plazo

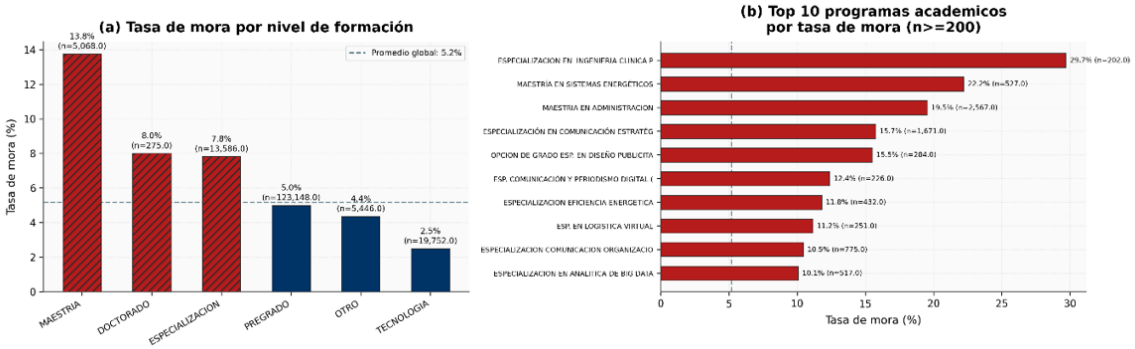


Figura 5.10. Tasa de mora por nivel de formación y programa académico

5.7 Análisis bivariado de variables asociadas al riesgo de mora

El análisis bivariado examina la relación estadística entre cada variable predictora y la variable objetivo **En_Mora.** Una variable es útil para el modelo si existe una diferencia estadísticamente significativa en la tasa de mora entre sus categorías (o una correlación significativa con **En_Mora** para las variables numéricas). Se usaron dos pruebas estadísticas complementarias: el *test* chi-cuadrado de Pearson (χ^2) para variables categóricas y el *test* de *Mann-Whitney U* (MW) para variables numéricas, junto con los intervalos de confianza exactos de Clopper-Pearson al 95% para las proporciones de mora por categoría.

¿Por qué usar *test* de *Mann-Whitney* en lugar de la prueba t?

La prueba t de Student compara medias y asume que los datos siguen una distribución normal. Las variables financieras del portafolio presentan distribuciones fuertemente asimétricas (asimetría de hasta +30) que violan ese supuesto. El *test* de *Mann-Whitney U* es una alternativa no paramétrica que compara medianas y no requiere supuesto de normalidad.

5.7.1 Estado del Estudiante

El estado académico del estudiante es la variable con mayor poder discriminante en el análisis bivariado, confirmado por la prueba chi-cuadrado ($\chi^2 = 858{,}42$; $p < 0{,}001$; $gl = 5$). Las tasas de mora observadas por categoría son:

5.7.2 Condición de Graduado

La prueba chi-cuadrado confirma una asociación estadísticamente significativa entre la condición de graduado y el riesgo de mora ($\chi^2 = 695{,}20$; $p < 0{,}001$; $gl = 2$). Los estudiantes graduados presentan una tasa de mora del 4,31% (n=71.436), inferior al 5,15% de los no graduados (n=81.361). Esa diferencia de 0,84 puntos porcentuales sugiere que la condición de graduado puede estar asociada con una menor probabilidad de incumplimiento, posiblemente porque la titulación facilita mejores condiciones de inserción laboral y, en consecuencia, una mayor capacidad para atender las obligaciones financieras adquiridas con la institución.

5.7.3 Estrato Socioeconómico

La prueba chi-cuadrado confirma una asociación estadísticamente significativa entre estrato y mora ($\chi^2 = 600{,}93$; $p < 0{,}001$; $gl = 5$). El patrón observado en los datos difiere de la hipótesis intuitiva de gradiente decreciente simple, según la cual se esperaría que los estratos más bajos presentaran siempre mayores niveles de mora. Los resultados muestran un comportamiento más heterogéneo: los estratos 3 y 4 concentran una parte importante de los beneficiarios y presentan mayor exposición al riesgo de incumplimiento, mientras que los estratos 1 y 2 podrían estar más asociados a esquemas de apoyo financiero, becas, descuentos u otras alternativas de financiación. Este hallazgo sugiere que el riesgo de mora no depende únicamente del nivel socioeconómico, sino también de la combinación entre capacidad de pago, acceso a beneficios institucionales y tipo de financiación utilizada.

5.7.4 Variables Financieras

El análisis bivariado de las variables numéricas, realizado con la prueba de *Mann-Whitney U,* mostró diferencias estadísticamente significativas ($p < 0.001$) entre los grupos en mora y sin mora en todas las variables financieras evaluadas. En la interpretación de estos resultados se dio mayor peso a las medianas, por ser más representativas que las medias en distribuciones asimétricas. Se incluyó una consideración metodológica relacionada con el riesgo de *data leakage*.

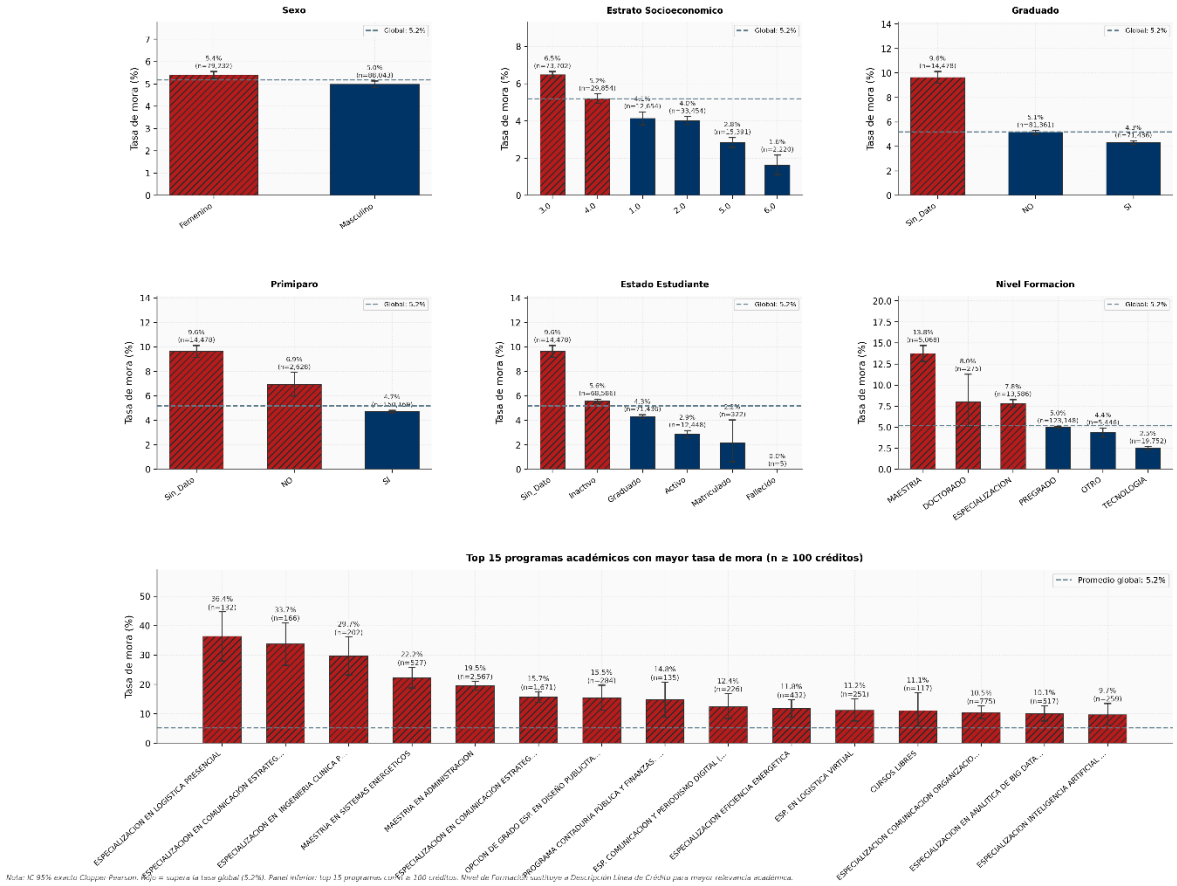
5.7.5 Identificación de variables con *data leakage*

La fuga de datos ocurre cuando se incluyen como predictoras variables que, en el momento real de la predicción, no estarían disponibles porque reflejan información futura o simultánea al evento que se quiere predecir. Incluir variables con *leakage* produce modelos con desempeño extraordinariamente alto en evaluación, pero inútiles en producción, porque en el mundo real esas variables no existen en el momento de tomar la decisión.

En el sistema de crédito, el momento de la decisión es el inicio del semestre, cuando se aprueba un nuevo crédito para un estudiante. En ese momento, las siguientes variables tienen valor cero o no están disponibles para todos los estudiantes:

La variable % Recaudo no resulta adecuada para la predicción prospectiva, ya que su valor se consolida al final del ciclo y refleja el comportamiento de pago acumulado en el período. Aunque presenta una alta correlación con En_Mora (r = -0.73), esa relación no debe interpretarse como capacidad predictiva, sino como una consecuencia simultánea del evento de mora.

Las variables **Saldo_nota_debito** y **Saldo_Cliente** requieren un tratamiento metodológico especial, dado que pueden reflejar información simultánea al evento de mora y, por tanto, introducir riesgo de *leakage* si se incorporan sin control temporal. **Saldo_nota_debito** corresponde al valor pendiente de la cuota y, cuando mantiene valores altos al final del período, puede estar capturando directamente el no pago. **Saldo_Cliente** refleja la deuda acumulada del estudiante, la cual se incrementa en presencia de mora activa. En ambos casos, su uso exige precaución, ya que no representan necesariamente señales prospectivas sino consecuencias del mismo fenómeno que se busca predecir.



Nota: IC 95% exacto Clopper-Pearson. Rojo = supera la tasa global (5.2%). Panel inferior: top 15 programas con n ≥ 100 créditos. Nivel de Formación sustituye a Descripción Línea de Crédito para mayor relevancia académica.

Figura 5.11 Tasa de mora por variables categóricas con intervalos de confianza al 95% (Clopper-Pearson)

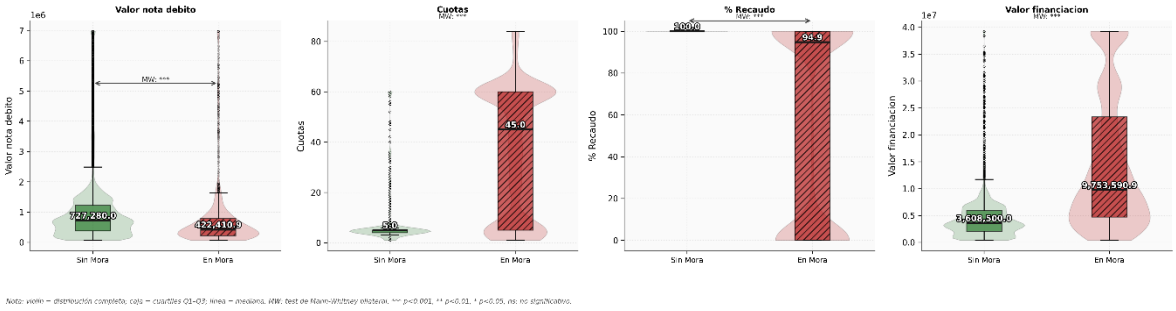


Figura 5.12 Distribución de variables numéricas según estado de mora (*test* de *Mann-Whitney*)

5.8 Análisis Multivariado

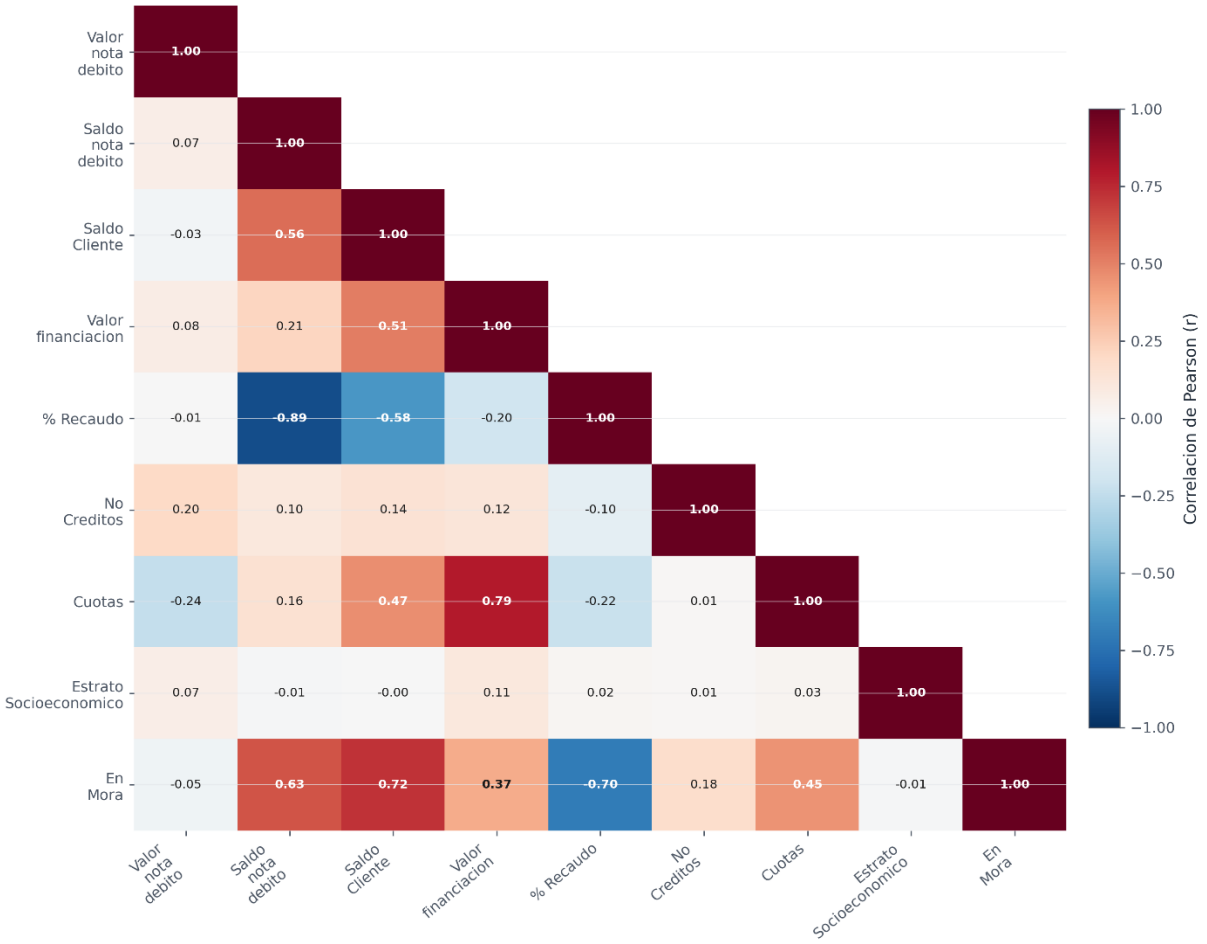
El análisis multivariado va más allá de las relaciones par a par (bivariado) y examina las interacciones entre múltiples variables simultáneamente. Es importante para detectar dos problemas que podrían afectar la calidad del modelo: la multicolinealidad entre predictores (variables que aportan información redundante) y los efectos de interacción (situaciones donde el efecto de una variable sobre la mora cambia según el nivel de otra variable).

5.8.1 Correlación de Pearson entre Variables Numéricas

La matriz de correlación de Pearson dejó identificar pares de variables numéricas con alta colinealidad. Su cálculo se hizo sobre las variables numéricas candidatas a predictores, excluyendo aquellas con riesgo de *data leakage*. En general, se observó una alta correlación entre las variables de valor monetario, especialmente entre **Valor_nota_debito**, **Valor_financiacion** y sus variables derivadas, lo cual es esperable porque representan distintas medidas de un mismo monto de crédito. Por eso, en el proceso de ingeniería de características desarrollado en el Capítulo 6 se seleccionó la variable más informativa de este grupo para evitar redundancias.

En relación con la variable objetivo, las asociaciones lineales más altas, aunque de magnitud moderada, se encontraron en **Valor_financiacion** (r = +0.29) y Cuotas (r = +0.34). Ese resultado sugiere que los créditos de mayor monto y con mayor número de cuotas presentan una mayor probabilidad de mora, en concordancia con lo observado previamente en el análisis de las líneas de crédito.

Estrato mostró una correlación baja con **En_Mora** (r = -0.013; p < 0.001). Aunque esa asociación es estadísticamente significativa, su magnitud es pequeña, lo que indica que la relación entre ambas variables no sigue un patrón lineal simple. Es posible que el modelo capture relaciones no lineales que la correlación de Pearson no alcanza a reflejar.



Nota: solo triangulo inferior (matriz simetrica). Magnitudes segun Cohen (1988): |r| < 0.10 trivial, 0.10-0.30 debil, 0.30-0.50 moderada, >= 0.50 fuerte. Negrita resalta |r| >= 0.30. Con n = 167,275 la significancia estadistica no es criterio util; se reporta tamano del efecto.

Figura 5.13 Matriz de correlación de Pearson entre variables numéricas del *dataset* de modelado. Solo se muestra el triángulo inferior (la matriz es simétrica). Escala de color divergente: azul = correlación positiva, rojo = correlación negativa, blanco = ausencia de correlación. Los valores en negrita destacan correlaciones con $|r| > 0.30$

5.8.2 V de Cramér entre Variables Categóricas

La relación entre las variables categóricas se examinó con la V de Cramér. Los resultados muestran una asociación alta entre Graduado y **Estado_Estudiante** (V = 0.89), lo que sugiere que ambas variables capturan información cercana y que no es necesario incluirlas simultáneamente en el modelo. Se priorizó **Estado_Estudiante** por su mayor capacidad discriminante. Sexo, Primíparo y Estrato presentaron asociaciones bajas con el resto de variables, lo que indica que aportan información más independiente al análisis.

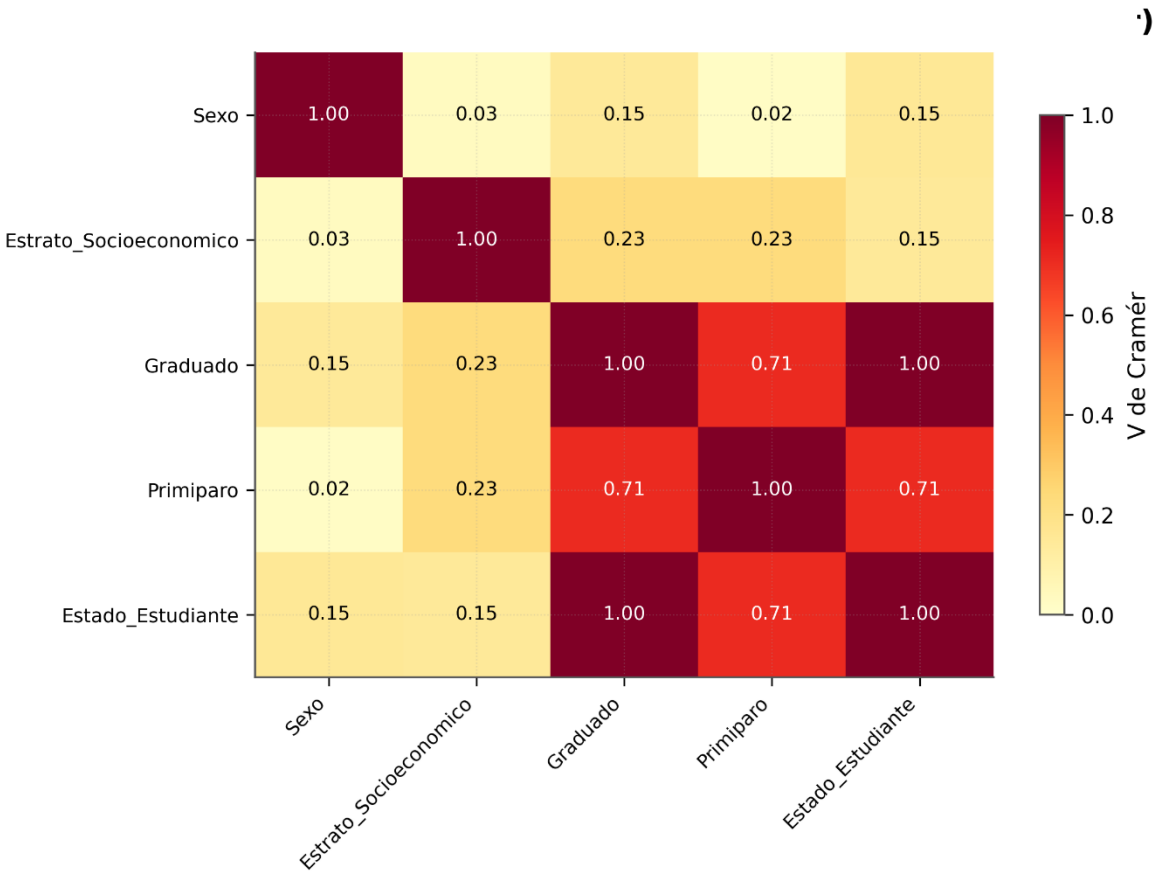


Figura 5.14 Asociación entre variables categóricas (V de Cramér)

5.8.3 Interacción entre estrato y sexo

El *heatmap* de la Figura 5.15 cruza simultáneamente estrato socioeconómico y sexo para examinar si el efecto del estrato sobre la mora es uniforme entre ambos géneros o si existe interacción. Este tipo de análisis es relevante para el diseño de políticas focalizadas de apoyo financiero.

Los datos observados muestran que la relación estrato-mora tiene patrones diferenciados por sexo: en estratos 1-2, las mujeres presentan tasas de mora marginalmente superiores a los hombres, mientras que en estratos 3-4 la situación se invierte. En estratos 5-6, ambos sexos convergen en tasas bajas. Esta interacción, aunque moderada en magnitud, justifica incluir el término de interacción Estrato × Sexo como *feature* candidata en el proceso de ingeniería de características del Capítulo 6, para que el modelo pueda capturarla si tiene suficiente poder estadístico.

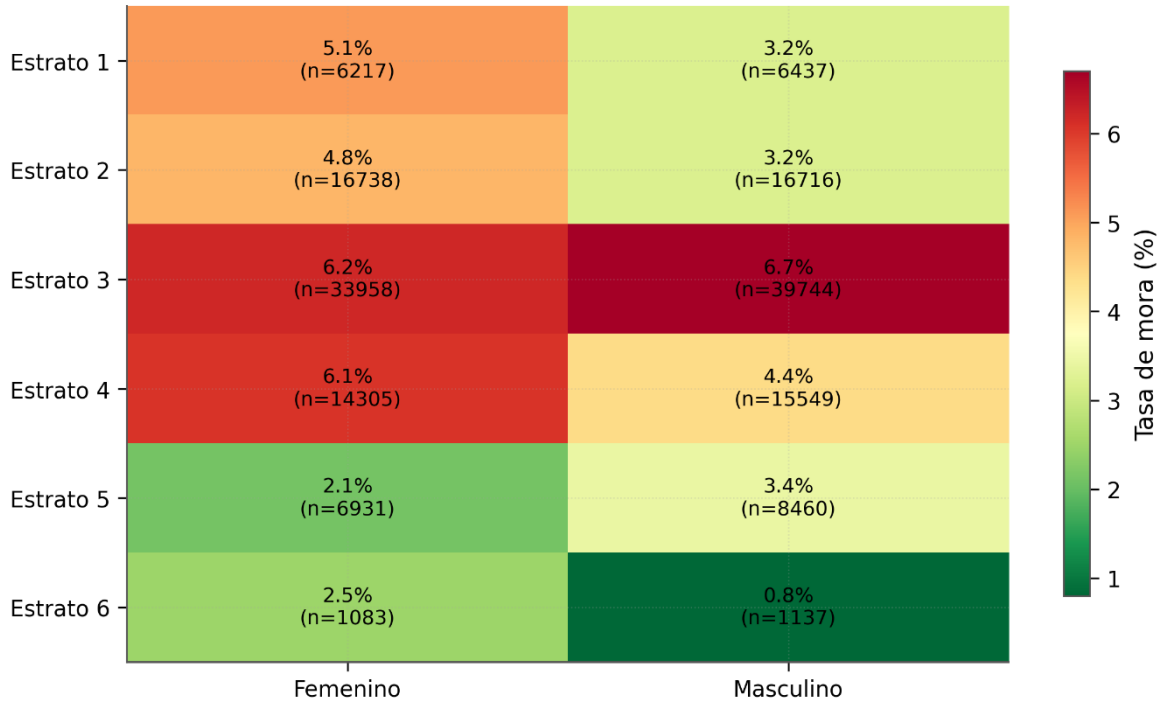


Figura 5.15 Mapa de calor de tasa de mora por Estrato Socioeconómico y Sexo (interacción bivariada). Escala de color RdYlGn inversa: verde = tasa baja, rojo = tasa alta. Cada celda muestra la tasa de mora en porcentaje y el número de observaciones (n). Las celdas más intensas en rojo identifican los segmentos de mayor riesgo conjunto

6. CONSTRUCCIÓN DEL CONJUNTO DE DATOS PARA EL MODELADO PREDICTIVO DEL RIESGO CREDITICIO EDUCATIVO

Objetivo específico 2: Preparar, transformar y depurar los datos históricos del sistema de financiación institucional para su uso en modelos de aprendizaje supervisado.

Este capítulo desarrolla la Fase 4 de CRISP-DM (Preparación de Datos) y conecta el análisis exploratorio del Capítulo 5 con el modelado predictivo del Capítulo 7. Si en el capítulo anterior se identificaron los patrones y factores asociados al riesgo, aquí se documentan las decisiones que permitieron convertir esos hallazgos en un conjunto de datos válido para alimentar los algoritmos de clasificación.

En la práctica, la preparación de datos es donde se toman las decisiones más importantes de un proyecto de Ciencia de Datos: qué variables incluir, cuáles excluir, cómo codificarlas, cómo dividir los datos y cómo tratar el desbalanceo. Una variable mal incluida, que contenga información del futuro (*data leakage*), puede producir un modelo con métricas excelentes en evaluación, pero inútil al aplicarse a créditos reales. El riesgo es alto en crédito educativo porque muchas columnas del sistema contable reflejan el resultado del pago, no las condiciones previas al otorgamiento.

Las decisiones de este capítulo se apoyaron en la teoría de aprendizaje automático, en el conocimiento del proceso de crédito de la universidad adquirido en las fases anteriores y en las recomendaciones de la literatura especializada [30],[31],[41]. El capítulo se organiza en cinco secciones: diagnóstico de *leakage*, ingeniería de características, conformación del vector final, división temporal y tratamiento del desbalanceo.

6.1 Diagnóstico y prevención de *data leakage*

El *data leakage* ocurre cuando el modelo tiene acceso a información que no estaría disponible al momento real de hacer la predicción. Baesens [41] lo señala como la causa principal de fracaso en modelos de *credit scoring* puestos en producción. Un modelo contaminado con *leakage* puede alcanzar un AUC de 0.95 en evaluación, pero al aplicarse a créditos nuevos su desempeño cae drásticamente porque las variables que le daban poder predictivo no existen cuando se toma la decisión.

El momento de la decisión es el inicio del semestre, cuando se aprueba un nuevo crédito. En ese momento el asesor financiero conoce el estrato del estudiante, su programa, el tipo de crédito y su historial de pagos previos, pero no sabe cuánto pagará al final del semestre ni cuál será su saldo acumulado. Cualquier variable cuyo valor al inicio del semestre sea desconocido o refleje el resultado del pago constituye *leakage*.

6.1.1 Variables excluidas y justificación individual

Tras analizar cada una de las 25 columnas originales del *dataset*, se identificaron y excluyeron

siete variables con riesgo de *leakage*. Cada exclusión se justifica a continuación:

% Recaudo tiene la mayor correlación con el *target* (r = −0.73), pero es una correlación espuria. Al inicio del semestre ningún estudiante ha pagado: el % Recaudo es cero para todos. Su valor final refleja cuánto pagó al cierre del ciclo. Un modelo que use esta variable aprende que los que no pagaron tienen bajo recaudo, lo cual no es una predicción sino una descripción del mismo evento.

No_Creditos fue la exclusión más relevante del proyecto. En versiones anteriores del modelo aparecía como la variable más importante con 25.1% de importancia. Al examinar su distribución se descubrió que no es un conteo de créditos del estudiante sino un identificador secuencial del sistema contable: los valores crecen monótonamente con el tiempo. El modelo aprendía una señal temporal artificial que coincidía con el *drift* post-pandemia, pero no reflejaba un patrón real de riesgo individual. Al excluirla, el ranking de importancia cambió por completo y las verdaderas variables predictivas quedaron expuestas: Cuotas (26.8%), Año (19.9%) y **Nivel_Formacion** (15.2%).

Saldo_nota_debito y **Saldo_Cliente** crecen únicamente cuando hay mora activa. Para un estudiante al día estos valores son cero; para uno en mora reflejan la deuda acumulada. Son variables simultáneas al evento de mora, no previas.

Valor_pagado_nota_debito refleja pagos realizados durante el ciclo del crédito. Su valor al inicio del semestre es cero para todos.

Graduado captura el estado académico al momento de la extracción (2025), no al otorgar el crédito. Un crédito de 2018 aparece con Graduado = SI porque el estudiante ya se graduó años después. Usarla como predictor implicaría que el modelo "sabe" si el estudiante se va a graduar, lo cual es imposible en la práctica.

Estado_describe es la variable a partir de la cual se construyó el *target* (**En_Mora**). Incluirla como predictor le daría al modelo acceso directo a la respuesta.

La Tabla 6.1 resume las siete variables excluidas, su correlación con el *target* y la justificación de la exclusión.

Tabla 6.1: Variables excluidas por *data leakage*: correlación con **En_Mora**, valor al inicio del semestre y justificación

Variable	Razón de exclusión
No_Creditos	Identificador secuencial del crédito (correlación 0.998 con **Nota_debito**). Reemplazado por **N_Creditos_Previos**
Saldo nota débito	Resultado del pago, no disponible al otorgar
Saldo cliente	Refleja pagos realizados
% Recaudo	Mide lo que se busca predecir

Valor pagado	Información posterior al evento
Graduado	*Leakage* severo: graduarse implica haber pagado
Estado describe	Fuente directa del *target*

6.2 Ingeniería de características con enfoque histórico

Una vez excluidas las variables con *leakage*, se necesitaba compensar la pérdida de información predictiva con variables que fueran válidas prospectivamente. Para eso se diseñaron tres variables de comportamiento histórico bajo el esquema t→t+1: al momento de otorgar un crédito en el semestre t, se conoce con certeza lo que ocurrió en los semestres anteriores.

Mora_Semestre_Anterior es una variable binaria que indica si el estudiante tuvo mora en el semestre t−1. Para estudiantes de primer ingreso se asigna el valor 0. Su poder predictivo es notable: la tasa de mora entre quienes tuvieron mora el semestre anterior es del 55.4%, frente al 4.9% de quienes no tienen historial de incumplimiento, una diferencia de 11 veces [6].

N_Moras_Hist es el conteo acumulado de semestres con mora a lo largo de toda la historia del estudiante. Los estudiantes con al menos un episodio previo presentan una tasa de mora del 49.1%, frente al 4.8% de quienes no tienen historial.

Pct_Mora_Hist normaliza el conteo anterior por la antigüedad del estudiante: un estudiante con 2 moras en 10 semestres (20%) tiene un perfil diferente al de uno con 2 moras en 3 semestres (67%). Sin esa normalización, el modelo podría confundir antigüedad con propensión al riesgo.

Las tres variables son prospectivamente válidas porque al momento de aprobar un crédito ya se conoce todo lo que ocurrió en semestres anteriores. Su aporte colectivo al modelo es del 11.8% de importancia.

6.2.1 Variables derivadas adicionales

Se derivaron seis variables adicionales a partir de las columnas originales.

Es_Femenino codifica el sexo como variable binaria.

Es_Primiparo indica si es el primer crédito del estudiante.

Es_Activo captura si el estado académico era Activo al momento del registro.

Es_Largo_Plazo clasifica los créditos con más de 6 cuotas como de largo plazo, una distinción que el Capítulo 5 confirmó como el factor de riesgo individual más importante: mora del 25.0% en largo plazo versus 1.8% en corto plazo. El umbral de 6 cuotas corresponde a la clasificación operativa del Departamento de Apoyo Financiero.

Nivel_Formacion_Ord codifica ordinalmente el nivel de formación: 1 = Tecnología (n=19,752), 2 = Pregrado (n=128,594), 3 = Especialización (n=13,586), 4 = Maestría (n=5,068), 5 = Doctorado (n=275).

Estrato_num convierte el estrato socioeconómico a numérico preservando su ordinalidad natural.

El *dataset* de modelado quedó conformado por 167,275 registros y 14 variables prospectivas, con una tasa de mora del 5.18%. Se clasifican en dos grupos:

Variables estructurales (9): **Estrato_num, Es_Femenino, Es_Primiparo, Es_Activo, Cuotas, Valor_nota_debito, Valor_financiacion, Es_Largo_Plazo y Nivel_Formacion_Ord**. Son características del crédito o del estudiante conocidas al momento de la solicitud.

Variables coyunturales e históricas (5): **Año, Semestre, Mora_Semestre_Anterior, N_Moras_Hist y Pct_Mora_Hist**. Capturan condiciones temporales y comportamiento previo del estudiante. La reducción de 25 columnas a 14 *features* no implica pérdida de información útil, ya que solo se eliminaron variables con nulos excesivos (Departamento, 40.2%), redundancias entre variables monetarias correlacionadas y las 7 variables de *leakage*.

La Tabla 6.2 presenta el vector completo con estadísticas descriptivas.

Tabla 6.2: Vector final de 14 *features* prospectivas

No	***Features* prospectivas**
1	Estrato_num
2	Es_Femenino
3	Es_Primiparo
4	Es_Activo
5	Cuotas
6	Valor_nota_debito
7	Valor_financiacion
8	Mora_Semestre_Anterior
9	N_Moras_Hist
10	Pct_Mora_Hist
11	Es_Largo_Plazo
12	Año
13	Semestre
14	Nivel_Formacion_Ord

6.3 División temporal del *dataset*

En riesgo crediticio, la forma de dividir los datos tiene implicaciones directas sobre la validez de los resultados. El *split* aleatorio, que es estándar en clasificación genérica, produce estimaciones optimistas en este contexto porque mezcla registros de diferentes épocas: datos de 2024 (régimen de mora elevada) pueden caer en entrenamiento mientras datos de 2016 (mora baja) caen en prueba. Thomas et al. [31] demuestran que eso contamina la evaluación con información temporal cruzada que no existiría en un escenario real.

Se adoptó un *split* temporal con corte en 2022. El conjunto de entrenamiento comprende 140,298 registros (2015 a 2022) con tasa de mora del 4.1%, incluyendo tanto el régimen pre-pandemia como la transición. El conjunto de prueba comprende 26,977 registros (2023 y 2024) con tasa de mora del 10.7%, perteneciente al nuevo régimen post-pandemia.

La diferencia de mora entre conjuntos (4.1% vs 10.7%) es deliberada: es una prueba exigente donde el modelo debe generalizar a un entorno con incidencia 2.6 veces mayor. Si mantiene un desempeño aceptable bajo esas condiciones, su capacidad predictiva es genuina. El corte en 2022 se eligió por ser el último año completo antes de la estabilización del nuevo régimen.

6.4 Estandarización y tratamiento del desbalanceo con SMOTE

Las variables numéricas presentan escalas diferentes: **Valor_financiacion** se mide en millones de pesos mientras que **Es_Femenino** es binaria. Se aplicó estandarización con StandardScaler de *scikit-learn* [42], ajustado exclusivamente sobre el conjunto de entrenamiento y luego aplicado a ambos conjuntos. Si el scaler se ajustara sobre el *dataset* completo, los parámetros incorporarían información de la prueba al proceso de normalización y eso sería una forma sutil de *leakage*.

Como se mencionó en el capítulo 5, el *dataset* de modelado presenta una distribución marcadamente desbalanceada: 94.82% sin mora y 5.18% en mora, con un ratio de 18.3 a 1. Este desbalanceo no es un error en los datos sino una característica inherente del portafolio crediticio. El crédito educativo institucional fue diseñado para tener tasas de mora bajas en condiciones normales.

El desbalanceo tiene consecuencias técnicas que deben gestionarse [7]. Si un modelo se entrena directamente sobre estos datos, la función de pérdida que guía el aprendizaje se optimiza prediciendo siempre la clase mayoritaria (sin mora), porque eso minimiza el error global. Un clasificador trivial que predijera «sin mora» para todos los créditos obtendría una exactitud del 94.82% sin haber aprendido absolutamente nada sobre los morosos. Para un sistema de alerta temprana cuyo propósito es identificar a los estudiantes en riesgo antes de que incumplan, ese comportamiento es inaceptable.

Para contrarrestar el desbalanceo se aplicó la técnica SMOTE, que es el estándar de facto en la

literatura de *credit scoring* para el tratamiento de clases desbalanceadas. A diferencia de la duplicación simple de registros (*Oversampling* naïve), que copia registros existentes de la clase minoritaria y genera riesgo de sobreajuste, SMOTE genera muestras sintéticas nuevas que no existían en el *dataset* original.

Antes de SMOTE, el conjunto de entrenamiento contenía 134,521 registros sin mora y 5,777 en mora. Después de SMOTE ambas clases quedaron equilibradas en 134,521 registros cada una. Es esencial destacar que SMOTE se aplicó exclusivamente sobre el conjunto de entrenamiento y nunca sobre el de prueba. Si se aplicara también sobre el conjunto de prueba, las muestras sintéticas compartirían estructura con las reales del entrenamiento y el modelo las reconocería trivialmente, lo que produciría métricas artificialmente optimistas que no representarían el rendimiento real en producción. Esta precaución, que parece obvia pero es frecuentemente violada en la práctica [43], es una de las garantías metodológicas más importantes del proyecto.

La Figura 6.1 muestra el efecto del balanceo SMOTE sobre la distribución de clases en el conjunto de entrenamiento, con la proporción antes y después del tratamiento.

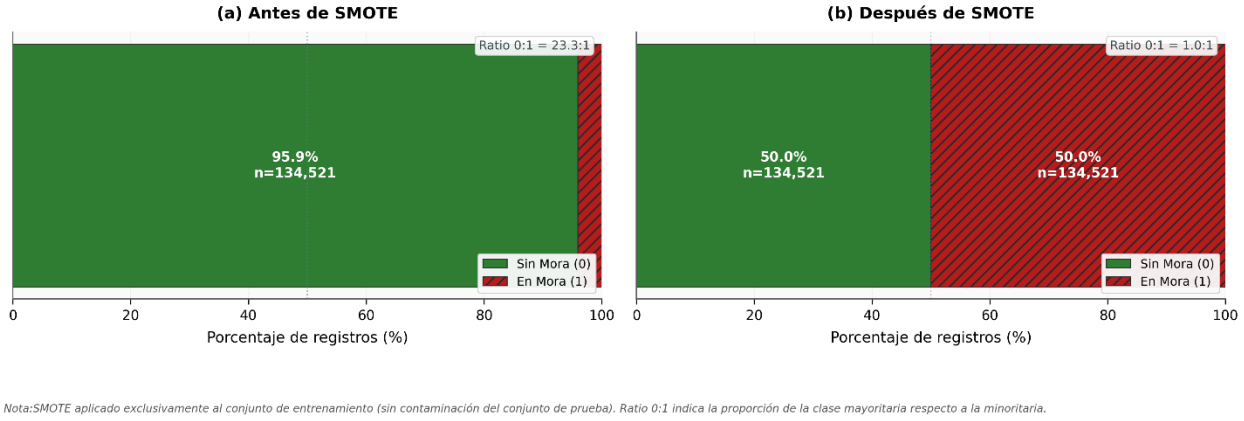


Figura 6.1. Efecto del balanceo SMOTE sobre la distribución de clases en el conjunto de entrenamiento. Panel izquierdo: distribución del conjunto de entrenamiento (ratio 23,3:1, n=140.298). Panel derecho: distribución tras SMOTE (ratio 1:1). Nota: el ratio global del *dataset* modelable es 18,3:1; el ratio 23,3:1 corresponde específicamente al subconjunto de entrenamiento (≤2022).

Este capítulo documentó la transformación de un *dataset* operativo de 25 columnas en un conjunto curado de 14 *features* prospectivas, dividido temporalmente y balanceado. Las decisiones más importantes fueron: la exclusión de siete variables con *leakage* (en particular **No_Creditos**, cuya importancia aparente del 25.1% resultó ser un artefacto del sistema contable), la construcción de *features* históricas t→t+1 (validadas empíricamente: 55.4% de mora con historial previo frente a 4.9% sin historial), y la elección del split temporal con corte en 2022 para evaluar la generalización del modelo ante el *drift* post-pandemia.

6.5 Justificación del uso de SMOTE y alcance metodológico del remuestreo

El uso de SMOTE se incorporó como una estrategia para atender el desbalanceo de clases observado en el problema de predicción de mora. En este contexto, la clase mayoritaria corresponde a las obligaciones que no presentan incumplimiento, mientras que la clase minoritaria corresponde a las obligaciones clasificadas en mora o cartera. Esta distribución desbalanceada puede afectar el aprendizaje de los modelos, dado que algunos algoritmos tienden a favorecer la clasificación de la clase mayoritaria y reducen su capacidad para identificar los casos de mayor interés institucional.

Para mitigar este efecto, se aplicó SMOTE únicamente sobre el conjunto de entrenamiento. Esta decisión metodológica evita contaminar los conjuntos de evaluación y permite que el desempeño del modelo sea medido sobre datos reales no remuestreados. En consecuencia, ni el conjunto holdout temporal ni la validación out-of-time fueron modificados mediante técnicas de balanceo. Ambos conservaron la distribución original de clases, con el fin de representar de manera más realista el comportamiento esperado en operación institucional.

Es importante precisar que el uso de SMOTE en este proyecto no se interpreta como una mejora demostrada frente a modelos sin remuestreo, dado que no se ejecutó un experimento comparativo formal entre ambos escenarios. Por tanto, el estudio no concluye que SMOTE mejore necesariamente el desempeño del modelo, sino que lo utiliza como una alternativa metodológica razonable frente al desbalance de clases, manteniendo controles para evitar fuga de información y sobreestimación del desempeño.

Asimismo, se reconoce que el remuestreo puede afectar la interpretación de las probabilidades predichas, debido a que modifica la prevalencia de la clase positiva durante el entrenamiento. Por esta razón, las probabilidades generadas por el modelo deben analizarse con cautela y evaluarse sobre datos reales no remuestreados mediante herramientas como curvas de calibración y Brier Score. En este sentido, los scores producidos por el modelo se entienden principalmente como una herramienta de ordenamiento y priorización del riesgo, más que como probabilidades absolutas definitivas.

Como línea de mejora futura, se recomienda ejecutar una comparación sistemática entre modelos entrenados con y sin SMOTE, utilizando la misma partición temporal, los mismos algoritmos y las mismas métricas de evaluación. Esta comparación permitiría determinar empíricamente si el remuestreo aporta mejoras en Recall, F1-score, KS, Gini y calibración, o si su efecto debe ser sustituido por otras estrategias como ajuste de pesos de clase, calibración posterior o selección de umbrales operativos.

7. MODELOS PREDICTIVOS DE CLASIFICACIÓN PARA LA IDENTIFICACIÓN TEMPRANA DEL RIESGO DE INCUMPLIMIENTO

Objetivo específico 3: Implementar y evaluar modelos de clasificación para anticipar el riesgo de incumplimiento, seleccionando el más adecuado según métricas de desempeño.

Después de caracterizar los factores asociados al riesgo de mora y de construir un conjunto de datos válido para el modelado, este capítulo desarrolla las etapas de modelado y evaluación de CRISP-DM. Su propósito es determinar en qué medida los algoritmos de clasificación dejan anticipar el incumplimiento en créditos educativos y cuáles ofrecen mejores condiciones de desempeño, estabilidad temporal e interpretabilidad para su uso institucional.

7.1 Selección y justificación de algoritmos

Se decidió entrenar cuatro algoritmos de clasificación supervisada simultáneamente en lugar de uno solo. Esta estrategia de comparación multialgoritmo no es caprichosa: deja una selección fundamentada del modelo final con base en evidencia empírica, en lugar de asumir a priori que un algoritmo particular será el mejor. La estrategia es consistente con la adoptada por Betancourt [36], quien compara regresión logística, *Random Forest* y *Gradient Boosting* para riesgo crediticio en Colombia, por Villanueva et al. [35], quienes evalúan múltiples técnicas en el contexto colombiano, y por Lessmann et al. [43], quienes demuestran que en *credit scoring* ningún algoritmo domina sistemáticamente a los demás en todos los *dataset*s.

Los cuatro algoritmos seleccionados dejan comparar alternativas con distintos niveles de interpretabilidad y desempeño predictivo. Eso es particularmente importante en instituciones educativas, donde la utilidad del modelo depende tanto de su precisión como de la posibilidad de explicar sus resultados.

7.1.1 Regresión Logística

La regresión logística se incluyó como modelo base por ser una de las técnicas más utilizadas en problemas de *credit scoring* y por su amplia trayectoria en la evaluación del riesgo crediticio [37], [6]. Su principal fortaleza en este proyecto radica en la interpretabilidad, ya que deja estimar probabilidades de mora y analizar con mayor claridad el efecto de cada variable sobre la predicción.

En términos prácticos, este modelo deja identificar cómo cambian las probabilidades de incumplimiento según características del estudiante y del crédito. Eso resulta especialmente útil en un contexto institucional donde no solo importa predecir bien, también importa entender y explicar los resultados obtenidos.

Para reducir el riesgo de sobreajuste se aplicó regularización L2. El parámetro C se ajustó mediante validación cruzada, evaluando valores entre 0.01 y 10.0. Dado que el desempeño del

modelo se mantuvo estable en ese rango, se seleccionó C = 0.1 como un valor conservador para el entrenamiento final.

7.1.2 Árbol de Decisión

El árbol de decisión se incluyó por su facilidad de interpretación, ya que deja representar las decisiones del modelo con reglas simples y comprensibles. Esta característica es especialmente útil en un contexto institucional, donde es importante que los resultados puedan explicarse de manera clara a usuarios no técnicos.

Para su entrenamiento se definió una profundidad máxima de seis niveles y un mínimo de 80 muestras por hoja. Estos parámetros se establecieron con el fin de evitar un crecimiento excesivo del árbol y reducir el riesgo de sobreajuste. Con esa configuración, el modelo conserva capacidad para identificar patrones relevantes en los datos sin generar reglas demasiado específicas que dificulten su generalización.

7.1.3 *Random Forest*

Random Forest [44] se incluyó como una alternativa más robusta que el árbol de decisión individual, ya que combina múltiples árboles entrenados sobre distintas muestras de los datos y luego integra sus resultados en una sola predicción. Esa lógica reduce la inestabilidad propia de un único árbol y mejora la capacidad de generalización del modelo. En el proyecto se configuró con 200 árboles, profundidad máxima de 8 y un mínimo de 50 muestras por hoja. Además, se utilizó **n_jobs**=1 para garantizar la reproducibilidad de los resultados bajo la semilla definida.

7.1.4 *Gradient Boosting*

Gradient Boosting [45] se incorporó como un modelo orientado a maximizar el desempeño predictivo. Aprovecha una estrategia secuencial en la que cada árbol busca corregir los errores del anterior [46]. Esa característica lo hace especialmente útil en problemas de clasificación sobre datos tabulares, aunque con una menor interpretabilidad que otros modelos. Para este proyecto se adoptó una configuración conservadora, con 200 estimadores, profundidad máxima de 4 y una tasa de aprendizaje de 0.05, con el fin de reducir el riesgo de sobreajuste y favorecer un comportamiento más estable en los datos.

7.2 Configuración experimental y sintonización de hiperparámetros

Con el fin de garantizar una comparación ordenada y reproducible entre los modelos evaluados, se definió una configuración experimental homogénea para los cuatro algoritmos de clasificación: Regresión Logística, Árbol de Decisión, Random Forest y Gradient Boosting. Todos los modelos fueron entrenados sobre el mismo conjunto de variables prospectivas, con la misma partición temporal y bajo el mismo esquema de validación.

La sintonización de hiperparámetros se realizó sobre el conjunto de entrenamiento, utilizando

validación cruzada estratificada de cinco particiones. Este esquema permitió evaluar distintas configuraciones internas de cada algoritmo sin utilizar el conjunto holdout temporal, el cual se reservó para la evaluación posterior del desempeño. La métrica principal de selección fue AUC-ROC, debido a que permite medir la capacidad discriminante del modelo sin depender de un único umbral de clasificación. No obstante, la decisión final también consideró Recall, F1-score y estabilidad temporal, dado que el objetivo operativo del proyecto es identificar tempranamente obligaciones con riesgo de mora.

7.3 Métricas de evaluación

Debido al desbalanceo del conjunto de datos, la exactitud no se consideró una métrica suficiente para evaluar el desempeño de los modelos, ya que podía ofrecer una impresión optimista sin reflejar realmente su capacidad para identificar casos de mora [7]. Por esta razón, se priorizaron métricas más informativas para este tipo de problema.

Entre ellas se incluyeron el *recall*, por su relevancia para medir la detección de estudiantes en mora; la precisión, para valorar la calidad de las alertas emitidas; el *F1-score*, como síntesis del balance entre ambas; y el AUC-ROC, para evaluar la capacidad general de discriminación del modelo. De manera complementaria, se revisaron también área bajo la curva precisión-*recall* (PR-AUC) y *Brier Score*, con el propósito de analizar el desempeño del modelo en un escenario de clases desbalanceadas y la calibración de las probabilidades estimadas.

Aunque el marco teórico reconoce la relevancia de métricas como Gini y KS en modelos de riesgo crediticio, el análisis empírico se concentró principalmente en AUC-ROC, Recall, Precision, F1-score, PR-AUC y Brier Score. Para futuras iteraciones se recomienda incorporar el cálculo formal de KS y reportar Gini de manera sistemática a partir del AUC-ROC.

7.4 Alcance metodológico del uso de SMOTE

El uso de SMOTE en este proyecto se incorporó como una estrategia para mitigar el desbalanceo de clases observado en la variable objetivo. Esta técnica fue aplicada únicamente sobre el conjunto de entrenamiento, conservando sin remuestreo los conjuntos de evaluación, holdout temporal y validación out-of-time. De esta manera, el desempeño del modelo fue evaluado sobre datos reales, manteniendo la distribución original de la cartera institucional.

Es importante precisar que el proyecto no desarrolló un experimento comparativo formal entre modelos entrenados con y sin SMOTE. Por tanto, los resultados no deben interpretarse como evidencia concluyente de que SMOTE mejora necesariamente el desempeño del modelo. Su uso se justifica como una decisión metodológica frente al desbalance de clases, pero se reconoce que su efecto debe ser evaluado en futuras iteraciones mediante una comparación controlada.

Dado que SMOTE puede modificar la prevalencia de la clase positiva durante el entrenamiento,

las probabilidades predichas deben interpretarse con cautela. En este sentido, los scores generados por el modelo se entienden principalmente como una herramienta de ordenamiento relativo del riesgo y priorización operativa, más que como probabilidades absolutas definitivas. Por esta razón, se recomienda complementar el uso del modelo con análisis de calibración, monitoreo periódico y recalibración conforme se incorporen nuevos datos.

7.5 Resultados del entrenamiento en *holdout* temporal

Los cuatro modelos se entrenaron sobre el conjunto de entrenamiento balanceado con SMOTE (269,042 registros, ratio 1:1) y se evaluaron sobre el conjunto de prueba temporal (2023-2024, 26,977 registros con distribución natural de 10.7% de mora). La Tabla 7.1 presenta los resultados comparativos.

Tabla 7.1 Métricas comparativas de los cuatro modelos en *holdout* temporal (2023-2024)

Modelo	Accuracy	*Precision*	*Recall*	*F1-score*	AUC-ROC	Gini ★
Regresión Logística	0.5544	0.1746	0.8496	0.2897	0.8462	0.6924
Árbol de Decisión	0.9374	0.7631	0.6007	0.6722	0.7830	0.5660
Random Forest operativo	0.9372	0.7319	0.6510	0.6890	0.8564	0.7128
Gradient Boosting	0.9358	0.7161	0.6617	0.6878	0.8559	0.7118

	Accuracy	*Precision*	*Recall*	*F1-score*	AUC-ROC
Regresión Logística	0.5544	0.1746	0.8496	0.2897	0.8462
Árbol de Decisión	0.9374	0.7631	0.6007	0.6722	0.7830
Random Forest	0.9372	0.7319	0.6510	0.6890	0.8564
Gradient Boosting	0.9358	0.7161	0.6617	0.6878	0.8559

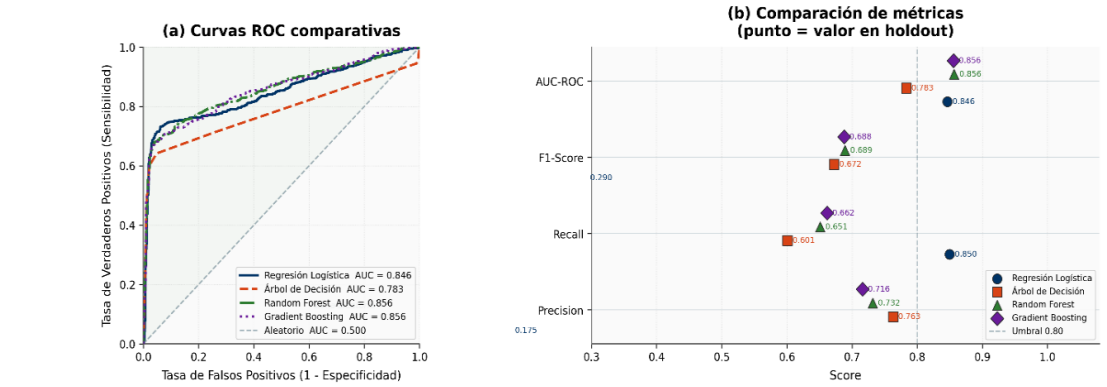
El coeficiente de Gini, derivado del AUC-ROC mediante la expresión Gini = 2 × AUC-ROC − 1, confirma la capacidad discriminante de los modelos evaluados. Random Forest obtiene el valor más alto (0.7128), seguido de Gradient Boosting (0.7118) y Regresión Logística (0.6924), mientras que el Árbol de Decisión presenta la menor separación entre clases (0.5660). Estos resultados son consistentes con el ordenamiento obtenido mediante AUC-ROC y refuerzan la selección de Random Forest como modelo operativo. El cálculo formal del estadístico KS se propone como complemento en futuras iteraciones del modelo, con el fin de fortalecer la comparación con las prácticas habituales de credit scoring institucional.

La Regresión Logística obtiene el *Recall* más alto (0.8496), detectando el 85% de los morosos reales, pero con una Precisión baja (0.1746): de cada 100 alertas que genera, solo 17 son morosos reales. En el contexto operativo eso no es necesariamente negativo, pues el costo de enviar un SMS preventivo a un estudiante que no iba a entrar en mora ($500 COP) es insignificante comparado con el costo de no detectar un moroso (pérdida potencial de $1.6 millones, que es el valor medio del crédito).

Random Forest logra el mejor equilibrio global con un AUC-ROC de 0.8564 (el más alto), *F1-score* de 0.6890 (el más alto) y Precisión de 0.7319. Eso significa que cuando *Random Forest* alerta sobre un moroso, tiene razón el 73% de las veces, lo que lo hace adecuado para *scoring* masivo donde se necesita priorizar los recursos de cobranza con eficiencia.

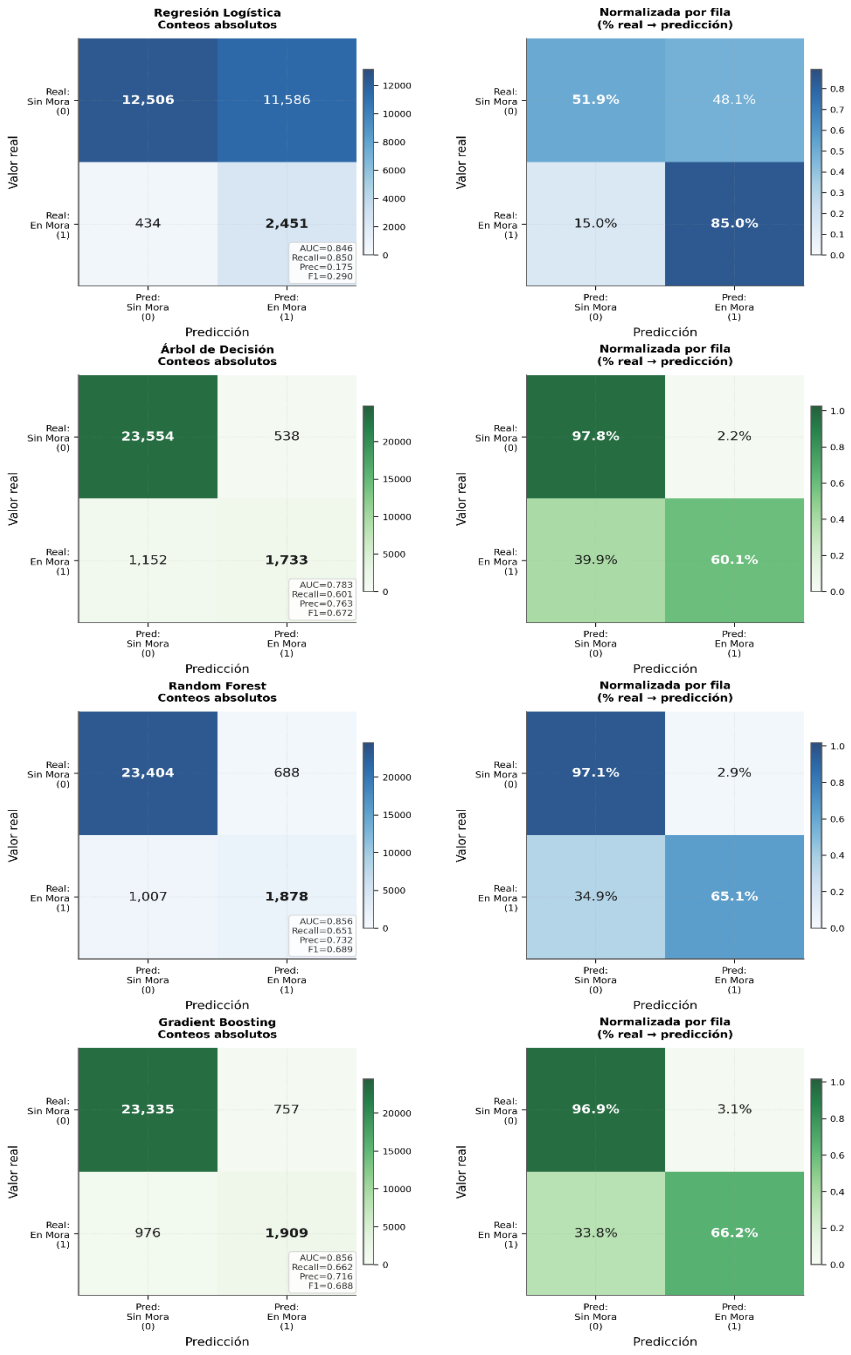
El Árbol de Decisión presenta el menor AUC-ROC (0.7830) pero la mejor Precisión (0.7631), lo que refleja su tendencia a ser conservador en sus predicciones. *Gradient Boosting* obtiene resultados virtualmente idénticos a *Random Forest* (AUC 0.8559 vs 0.8564), lo que confirma que ambos ensambles capturan patrones similares en estos datos.

Sobre las métricas complementarias, Random Forest obtiene el mejor PR-AUC (0.639) y el menor Brier Score (0.1020), lo que sugiere un mejor comportamiento relativo de sus probabilidades frente a los demás modelos evaluados. No obstante, la calibración de probabilidades debe profundizarse en futuras iteraciones mediante curvas de calibración y análisis por deciles de riesgo sobre datos reales no remuestreados.



Nota: (a) curvas ROC en holdout temporal; (b) dot plot de métricas en holdout. AUC: Área Bajo la Curva ROC. Umbral de clasificación = 0.50 (default).

Figura 7.1 Curvas ROC comparativas de los cuatro modelos. Cada curva muestra el trade-off entre tasa de verdaderos positivos (*Recall*) y tasa de falsos positivos para diferentes valores de umbral



Nota: izquierda = conteos absolutos; derecha = tasa recall por fila (diagonal = clasificaciones correctas). FP: créditos en mora no detectados; FN: créditos sanos mal clasificados.

Figura 7.2 Matrices de confusión (absoluta y normalizada) para cada modelo en el *holdout* temporal

7.6 Validación cruzada y diagnóstico de estabilidad temporal

Las métricas obtenidas en el *holdout* temporal dejan evaluar el desempeño de los modelos en un escenario específico correspondiente a los años 2023 y 2024. Para valorar su robustez y consistencia también se hizo una validación cruzada estratificada con cinco particiones. En ese proceso, el balanceo con SMOTE se aplicó únicamente sobre los datos de entrenamiento de cada partición, lo que evita que la información de validación influya en el ajuste del modelo.

Los resultados del AUC-ROC medio fueron de 0.8870 para Regresión Logística, 0.9077 para Árbol de Decisión, 0.9301 para *Random Forest* y 0.9338 para *Gradient Boosting*. A primera vista, este último parecería ofrecer el mejor desempeño. La comparación debe interpretarse con cautela, ya que la validación cruzada se basa en particiones aleatorias que no reflejan por completo el cambio temporal observado en los datos. Por eso fue necesario complementar el análisis con un diagnóstico de estabilidad temporal frente al *holdout* postpandemia.

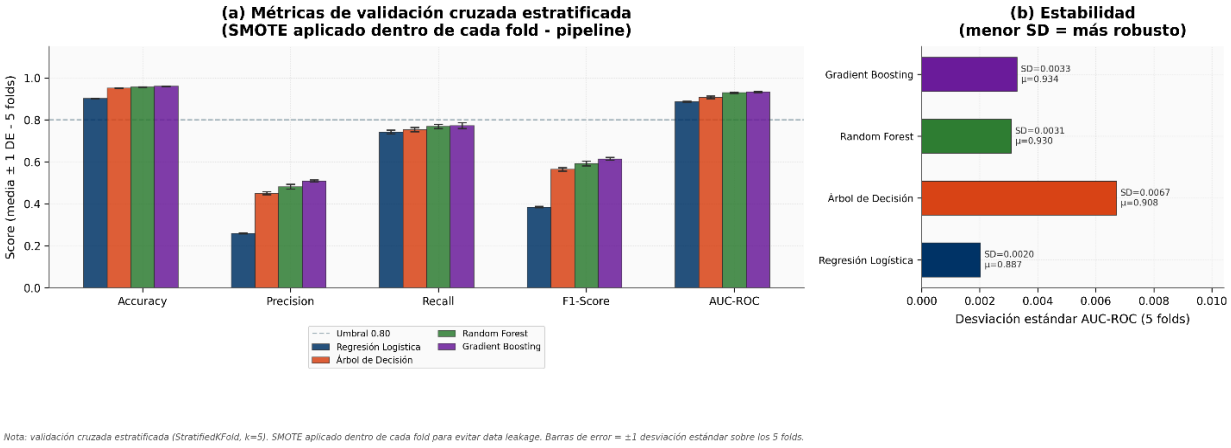


Figura 7.3 Resultados de validación cruzada estratificada (5 folds) por modelo y métrica

7.6.1 Diagnóstico de estabilidad temporal

El diagnóstico de estabilidad temporal permitió comparar el desempeño de cada modelo en la validación cruzada y en el *holdout* temporal, con el fin de identificar qué tanto cambiaban sus resultados al aplicarse sobre datos de un período posterior. Esta comparación resulta importante porque una diferencia alta entre ambos escenarios puede indicar que el modelo se ajusta bien a los datos de entrenamiento pero pierde capacidad de generalización frente a nuevas condiciones.

Los resultados muestran que la Regresión Logística presentó la menor variación, con una diferencia de 4,6% entre la validación cruzada y el *holdout* temporal. El Árbol de Decisión registró una caída más pronunciada, del 13,7%, lo que sugiere un menor nivel de estabilidad. *Random Forest* y *Gradient Boosting* mostraron diferencias de 7,9% y 8,3% respectivamente, manteniéndose dentro de un rango aceptable para el contexto del proyecto.

Tabla 7.2: Diagnóstico de estabilidad validación cruzada (CV) vs *Holdout* temporal: degradación por modelo

Estado	Modelo	AUC-ROC CV	± Desv.	AUC-ROC *Holdout*	Degradación	Interpretación
✓ OK	**Regresión Logística**	0,8870	±0,0020	**0,8462**	**+4,6%**	*Modelo más estable. Seleccionado como modelo interpretable.*
⚠	**Árbol de Decisión**	0,9077	±0,0067	**0,7830**	**+13,7%**	*Degradación alta: sobreajuste a patrones temporales del train.*
✓ OK	***Random Forest***	0,9301	±0,0031	**0,8564**	**+7,9%**	*Mejor AUC holdout. Seleccionado como modelo operativo.*
✓ OK	***Gradient Boosting***	0,9338	±0,0033	**0,8559**	**+8,3%**	*Similar a RF. Degradación aceptable.*

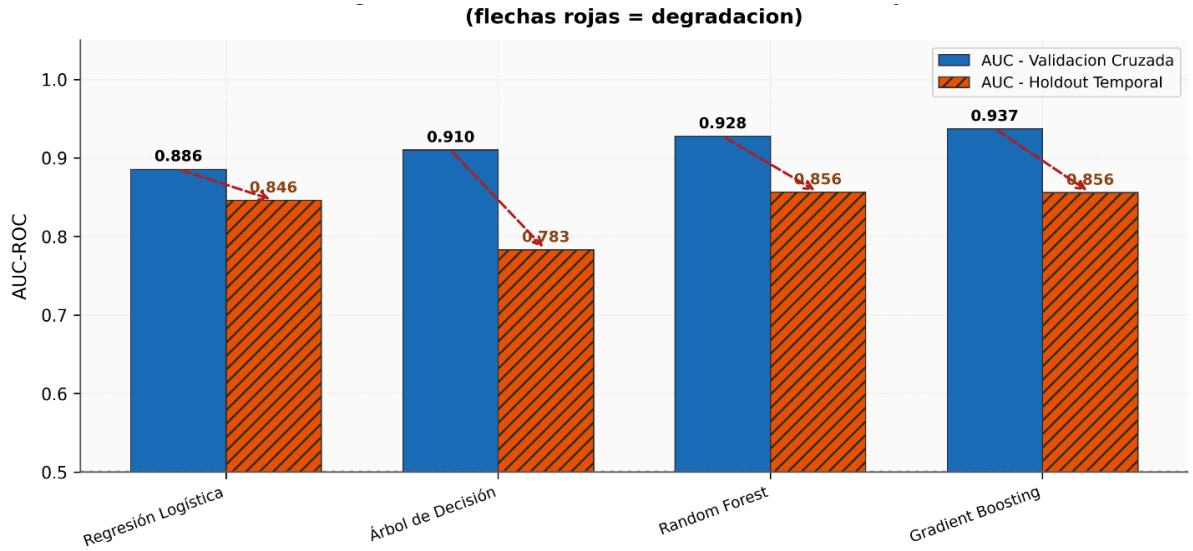


Figura 7.4 Diagnóstico de estabilidad: AUC-ROC en validación cruzada vs *holdout* temporal, con porcentaje de degradación

7.6.2 Selección dual de modelos

A partir del diagnóstico de estabilidad temporal, se optó por una selección dual de modelos que atiende dos necesidades distintas de la institución. La Regresión Logística se conservó como modelo interpretable, debido a que presentó la menor variación entre la validación cruzada y el *holdout* temporal, así como el mayor *recall* entre los modelos evaluados. Esas características la hacen útil para respaldar el análisis de factores de riesgo y facilitar la interpretación de los resultados por parte de la Dirección Financiera.

Random Forest se seleccionó como modelo operativo para el proceso de *scoring* y priorización de cobranza, ya que presentó el mejor desempeño global en el *holdout* temporal, con el mayor AUC-ROC y el mejor equilibrio entre precisión y *recall*. Ofreció condiciones más favorables para su

aplicación sobre la cartera y para la clasificación de estudiantes según nivel de riesgo.

Ambos modelos cumplieron con los criterios de desempeño definidos para el proyecto y dejaron responder tanto a la necesidad de interpretación institucional como a la de aplicación operativa.

7.7 Importancia de variables

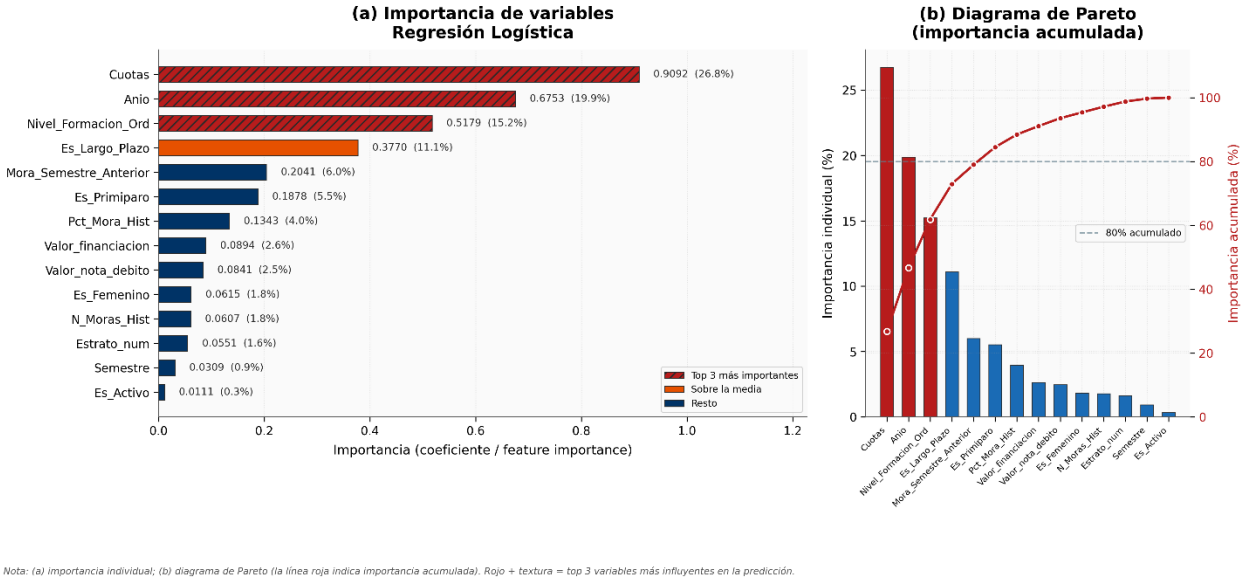
El análisis de importancia de variables en la Regresión Logística permitió identificar cuáles características tuvieron mayor peso en la predicción del riesgo de mora. Este resultado ayuda a entender mejor el comportamiento del modelo y aporta información útil para la toma de decisiones institucionales.

Las variables con mayor aporte fueron Cuotas (26.8%), Año (19.9%) y **Nivel_Formacion_Ord** (15.2%). En conjunto concentraron el 61.9% de la importancia total del modelo. En el caso de Cuotas, su relevancia es coherente con el comportamiento observado en el análisis exploratorio, donde los créditos de largo plazo presentaron mayores niveles de mora que los de corto plazo. Año recogió el efecto del cambio temporal en el comportamiento de pago, especialmente a partir de la pandemia, mientras que **Nivel_Formacion_Ord** mostró diferencias de riesgo entre los distintos niveles educativos.

Después de esas variables se ubicaron **Es_Largo_Plazo** (11.1%), **Mora_Semestre_Anterior** (6.0%), **Es_Primiparo** (5.5%) y **Pct_Mora_Hist** (4.0%). En conjunto, las variables históricas de comportamiento previo mostraron un aporte importante, lo que confirma que los antecedentes de mora y el historial de pago ofrecen información valiosa para anticipar el incumplimiento. El ranking completo de importancia individual y acumulada se presenta en la Tabla 7.3.

Tabla 7.3: Ranking de importancia de variables — coeficientes absolutos estandarizados de Regresión Logística

#	Variable	Coeficiente |β| estand.	Importancia (%)	Acumulado (%)	Barra
1	**Cuotas**	**0,9092**	**26,80%**	26,80%	
2	**Año**	**0,6753**	**19,90%**	46,60%	
3	**Nivel_Formacion_Ord**	**0,5179**	**15,20%**	61,90%	
4	**Es_Largo_Plazo**	0,377	11,10%	73,00%	
5	**Mora_Semestre_Anterior**	0,2041	6,00%	79,00%	
6	**Es_Primiparo**	0,1878	5,50%	84,50%	
7	**Pct_Mora_Hist**	0,1343	4,00%	88,40%	
8	**Valor_financiacion**	0,0894	2,60%	91,10%	
9	**Valor_nota_debito**	0,0841	2,50%	93,50%	
10	**Es_Femenino**	0,0615	1,80%	95,40%	
11	**N_Moras_Hist**	0,0607	1,80%	97,10%	
12	**Estrato_num**	0,0551	1,60%	98,80%	
13	**Semestre**	0,0309	0,90%	99,70%	
14	**Es_Activo**	0,0111	0,30%	100,00%	



Nota: (a) importancia individual; (b) diagrama de Pareto (la línea roja indica importancia acumulada). Rojo + textura = top 3 variables más influyentes en la predicción.

Figura 7.5 Ranking de importancia de variables con diagrama de Pareto acumulado.

7.8 Interpretabilidad con SHAP

El análisis con SHAP [10] complementó la interpretación global del modelo al dejar identificar cómo contribuye cada variable en las predicciones individuales. Mientras el ranking de importancia muestra qué variables son más relevantes en términos generales, SHAP permitió entender por qué un estudiante específico fue clasificado con mayor o menor riesgo de mora. Esta información resulta especialmente útil en un contexto aplicado, ya que facilita explicar los resultados del modelo y construir perfiles de riesgo más claros para la gestión institucional.

A partir de ese análisis se construyó un perfil comparativo entre estudiantes de alto riesgo y bajo riesgo. Para ello se tomaron como referencia los estudiantes ubicados por encima del percentil 75 del score y aquellos por debajo del percentil 25. Los resultados muestran diferencias marcadas en variables asociadas al comportamiento previo y a las condiciones del crédito. Entre los factores más diferenciadores están el número de moras históricas, la condición de largo plazo del crédito, la mora en el semestre anterior y el número de cuotas. En general, los estudiantes clasificados en alto riesgo presentan más antecedentes de incumplimiento, una mayor participación en créditos de largo plazo y un mayor número de cuotas que los estudiantes de bajo riesgo.

La validez de esta diferenciación se confirma al observar la tasa de mora real en ambos grupos: en el grupo de alto riesgo fue del 32.6%, mientras que en el de bajo riesgo fue del 3.1%. Estos resultados muestran que el análisis con SHAP no solo ayuda a interpretar mejor el modelo; también ayuda a caracterizar perfiles de riesgo útiles para orientar acciones de seguimiento y cobranza preventiva. La Tabla 7.4 presenta el perfil comparativo completo y la Figura 7.6 lo ilustra

gráficamente.

Tabla 7.4: Perfil comparativo: estudiantes de alto riesgo (score ≥ P75) vs bajo riesgo (score ≤ P25)

Variable	Alto Riesgo (Score ≥ P75)	Bajo Riesgo (Score ≤ P25)	Diferencia	Señal
Estrato_num	2,608	3,394	-23,1%	↓
Es_Femenino	0,417	0,634	-34,2%	↓
Es_Primiparo	0,715	0,926	-22,7%	↓
Es_Activo	0,262	0,144	**+82,4%**	↑
Cuotas	18,035	4,552	**+296,2%**	↑↑
Valor_nota_debito	1.362.464	1.224.530	+11,3%	↑
Valor_financiacion	10.195.032	4.864.312	**+109,6%**	↑↑
Mora_Semestre_Anterior	0,039	0,000	**+3941,9%**	🔴
N_Moras_Hist	0,215	0,000	**+21473,0%**	🔴
Pct_Mora_Hist	0,038	0,000	**+3765,9%**	🔴
Es_Largo_Plazo	0,388	0,004	**+9594,6%**	🔴
Año	2.023,667	2.023,193	—	—
Semestre	1,600	1,410	+13,5%	↑
Nivel_Formacion_Ord	2,110	2,414	-12,6%	↓

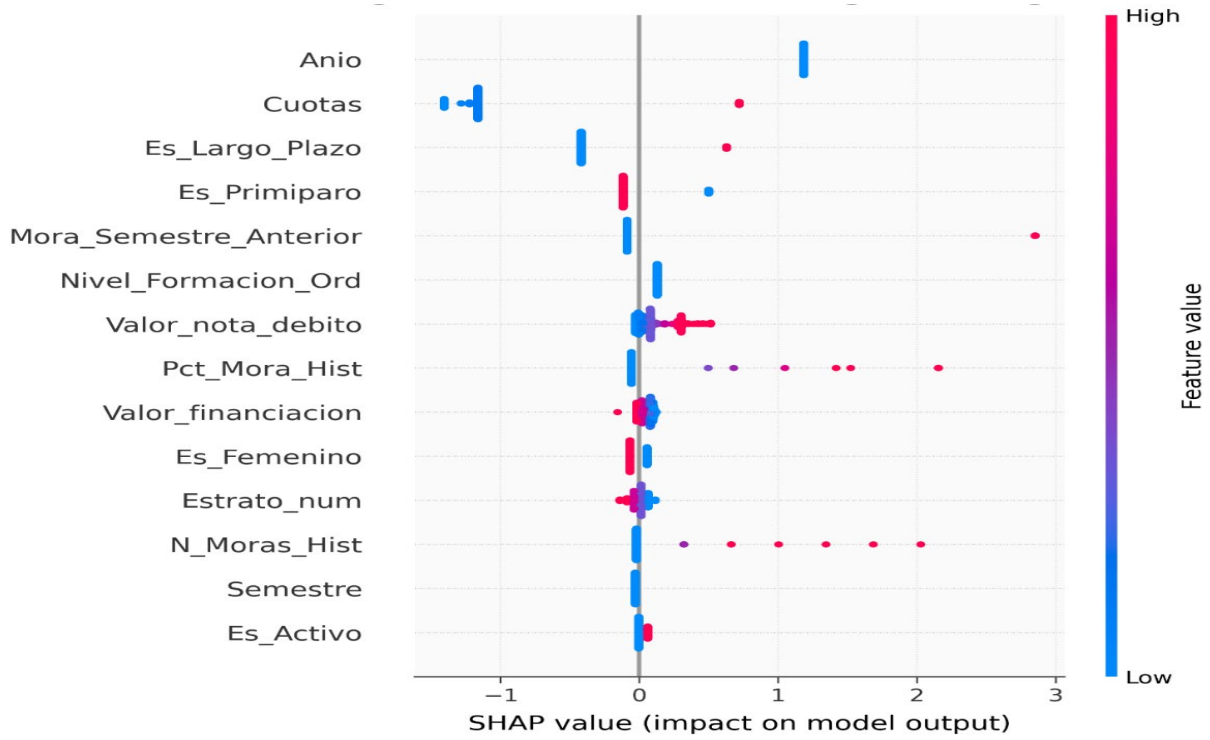


Figura 7.6 Análisis SHAP (*bee swarm plot*): contribución de cada variable a las predicciones individuales. Cada punto es un crédito; el color indica el valor de la variable (rojo = alto, azul = bajo)

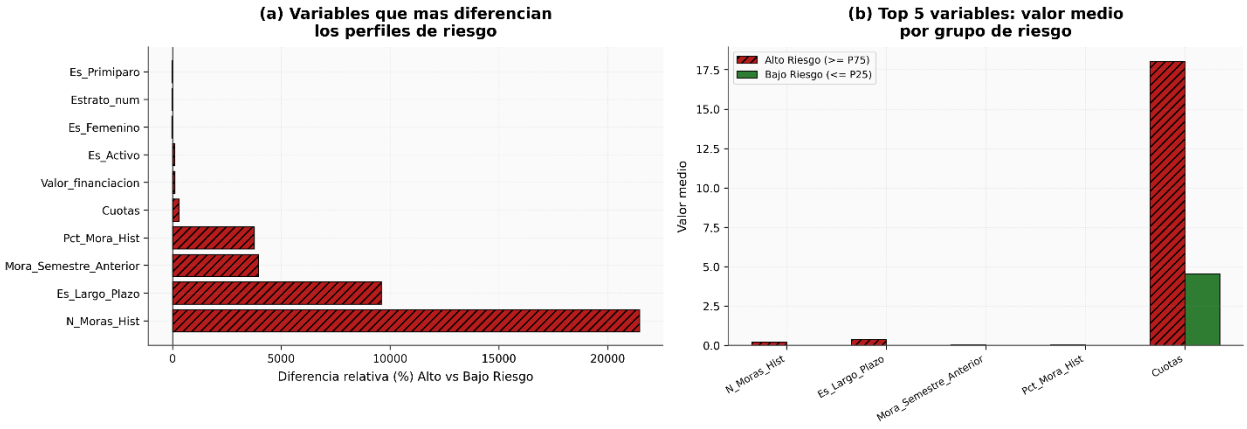


Figura 7.7 Perfil comparativo alto riesgo vs bajo riesgo: valores medios de cada variable por grupo con diferencias porcentuales

7.9 Validación *Out-of-time*

Como prueba adicional de estabilidad temporal se hizo una validación *out-of-time* con un esquema más exigente que el *holdout* principal. El modelo se entrenó con información hasta 2023 y se evaluó únicamente con datos de 2024, lo que permitió aproximarse a un escenario más cercano a su uso real.

Los resultados obtenidos fueron consistentes con los observados antes. El modelo alcanzó un AUC-ROC de 0.8573, un *recall* de 0.8596 y un *F1-score* de 0.3356. En conjunto, esos resultados indican que el desempeño se mantiene estable al aplicarse sobre un período no visto durante el entrenamiento y que el modelo conserva una buena capacidad para identificar estudiantes en riesgo de mora.

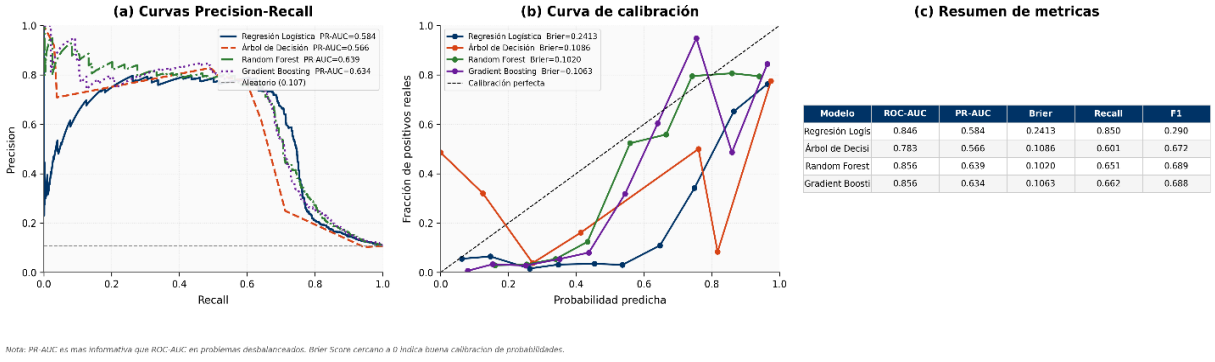


Figura 7.8 Curvas *Precision-Recall* AUC y análisis de calibración de probabilidades del modelo seleccionado

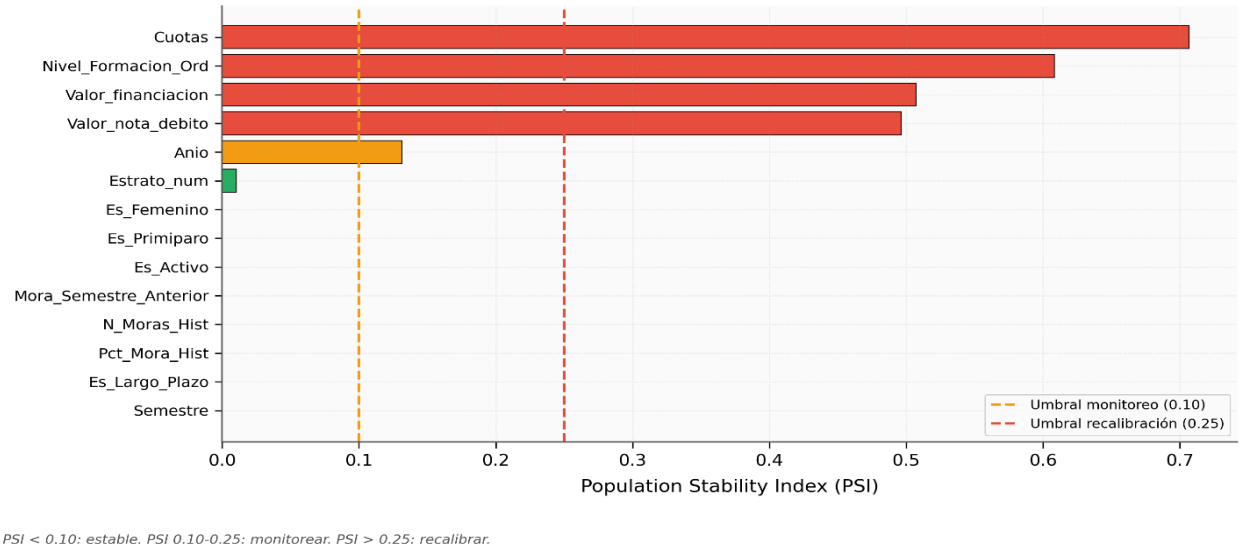


Figura 7.9 Population Stability Index (PSI) por variable: cuantificación del *drift* entre las distribuciones de *train* y *test*

7.10 Umbral óptimo de clasificación

La definición del umbral de clasificación fue clave para convertir las probabilidades estimadas por el modelo en decisiones operativas. Aunque el valor de 0.50 suele usarse por defecto, en este proyecto se consideró necesario evaluar otras alternativas, dado que el costo de no identificar oportunamente un caso de mora resulta más crítico para la institución que el de generar una alerta preventiva adicional.

Con ese criterio se compararon diferentes opciones de umbral, cada una con implicaciones distintas en términos de *recall*, precisión y equilibrio general del modelo. Entre ellas se consideró el índice de Youden J como referencia para apoyar la selección. Los resultados comparativos se presentan en la Tabla 7.5.

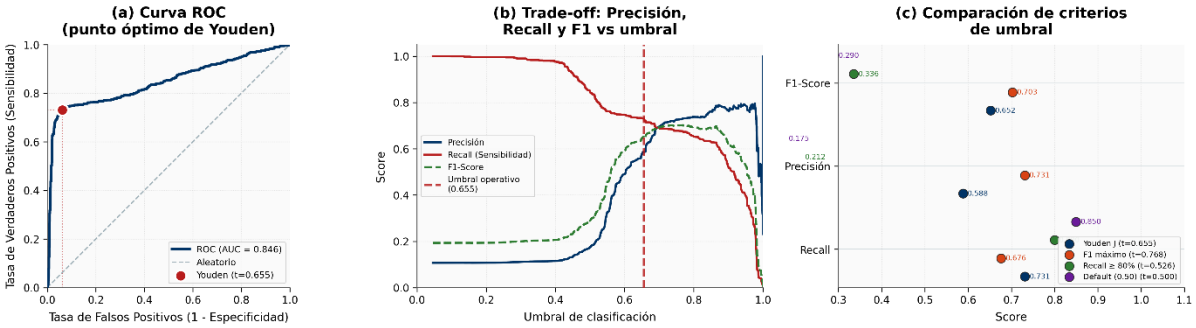
Tabla 7.5: Comparación de criterios para la selección del umbral de clasificación

Criterio	Umbral	*Recall*	Precisión	F1	Seleccionado
Youden J	**0,655**	**0,731**	**0,588**	**0,652**	**✓ Seleccionado**
F1 máximo	0,768	0,676	0,731	0,703	
Recall ≥ 80%	0,526	0,800	0,212	0,336	
Default (0.50)	0,500	0,850	0,175	0,290	

El criterio de Youden J permitió seleccionar un umbral de 0.655, con un *recall* de 0.731, una precisión de 0.588 y un *F1-score* de 0.652. El criterio de F1 máximo alcanzó un valor más alto de esa métrica, pero lo hizo a costa de una menor capacidad para detectar casos de mora. Los criterios orientados a maximizar el *recall* redujeron de forma importante la precisión y generaron

un volumen mayor de alertas falsas.

Por esa razón se eligió el umbral de Youden J como referencia operativa, ya que ofreció el mejor equilibrio entre identificar estudiantes en riesgo y mantener un nivel manejable de falsas alertas. El resultado se consideró adecuado para el propósito del modelo, sobre todo si se tiene en cuenta que las acciones preventivas asociadas a los niveles intermedios de riesgo implican costos relativamente bajos para la institución.



Nota: (a) punto de Youden maximiza J = TPR − FPR (sensibilidad + especificidad); (b) curvas de trade-off versus umbral; (c) dot plot comparativo entre criterios. Umbral operativo seleccionado: 0.655.

Figura 7.10 Selección del umbral óptimo: curvas de trade-off entre *Recall*, Precisión y F1 para diferentes valores de umbral. La línea vertical marca el umbral Youden J seleccionado (0.655)

7.11 Discusión comparativa con la literatura

Con esta información se puede ver que sí es posible anticipar el riesgo de incumplimiento en los créditos educativos con un nivel de desempeño útil para la gestión institucional. La evaluación comparativa permitió identificar un modelo con mayor capacidad interpretativa, como la Regresión Logística, y otro con mejor desempeño operativo, como *Random Forest*. El análisis de variables confirmó que las predicciones del modelo son consistentes con la dinámica de la cartera y con factores ya observados en el análisis previo. Con base en estos hallazgos, el capítulo siguiente se orienta a traducir el resultado del modelado en herramientas de segmentación y visualización para apoyar la toma de decisiones.

Sobre el desempeño predictivo, los resultados obtenidos en este estudio se ubican dentro del rango reportado por la literatura especializada en *credit scoring*. Eso sugiere que el desempeño alcanzado es consistente con los estándares del campo, aun cuando el proyecto se desarrolló con un número reducido de variables y sin acceso a información de centrales de riesgo externas, una condición diferente a la de muchos estudios del sector financiero.

La elección de la Regresión Logística como modelo interpretable también coincide con lo señalado en la literatura, que resalta la importancia de contar con modelos comprensibles en contextos donde, además de predecir, se requiere explicar las decisiones. En este proyecto esa interpretabilidad tuvo un valor especial, ya que permitió construir un perfil de estudiante de alto

riesgo que puede contrastarse con el conocimiento operativo del Departamento de Apoyo Financiero.

8. SEGMENTACIÓN DE PERFILES DE RIESGO Y VISUALIZACIÓN ESTRATÉGICA PARA LA GESTIÓN DEL CRÉDITO EDUCATIVO

Objetivo específico 4: Aplicar técnicas de segmentación y diseñar un tablero de control que facilite la visualización de perfiles de riesgo y apoye la toma de decisiones institucionales.

Este capítulo desarrolla la fase de despliegue de la metodología CRISP-DM y se orienta a traducir los resultados del modelo en herramientas útiles para la gestión del crédito educativo. Después de analizar los datos, preparar el conjunto de modelado y evaluar los algoritmos de clasificación, en este punto el interés se centra en cómo convertir esos hallazgos en acciones concretas para el Departamento de Apoyo Financiero a Estudiantes.

Con ese propósito, el capítulo presenta un conjunto de componentes orientados a la segmentación de la cartera, la priorización de acciones de cobranza preventiva, la estimación de su impacto potencial y la visualización de resultados mediante un tablero de control. En conjunto, estas herramientas buscan facilitar la interpretación de los resultados y apoyar su posible uso en la operación institucional.

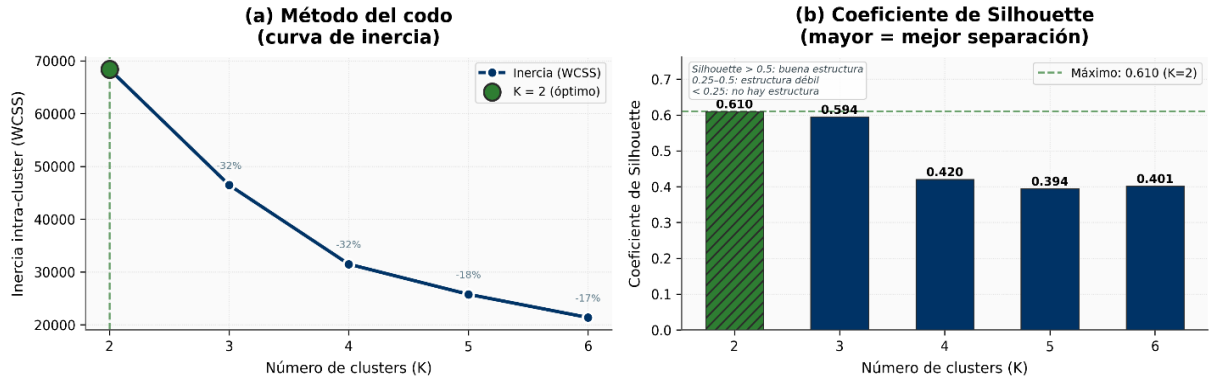
8.1 *Clustering* integrado con predicciones

Antes de definir la segmentación por zonas de riesgo a partir de umbrales del score, se hizo un análisis de *clustering* con el fin de identificar perfiles diferenciados dentro del portafolio. A diferencia de la segmentación por umbrales, que responde a una decisión operativa, el *clustering* deja explorar agrupaciones que surgen de la propia estructura de los datos y que pueden aportar una lectura complementaria del riesgo. Se utilizó *K-means* con K=2, valor seleccionado por dos criterios convergentes: (i) el coeficiente de silhouette alcanza su máximo en K=2 dentro del rango evaluado [2-6], y (ii) dos segmentos (alto riesgo / bajo riesgo) son los únicos directamente accionables por la Dirección Financiera; valores K>2 producen subdivisiones sin diferencia operativa clara para el equipo de cobranza.

Para este análisis se aplicó *K-means* usando cuatro variables: **Valor_nota_debito**, **Cuotas**, **Estrato_num** y **Prob_Mora**. El número de grupos se definió comparando diferentes alternativas y evaluando su nivel de separación interna. Los resultados mostraron que la mejor solución correspondía a dos *clusters*, por lo que se adoptó esa configuración para el análisis posterior. La Tabla 8.1 presenta el perfil de los grupos identificados.

Tabla 8.1. Perfil de *clusters* con predicciones del modelo

	N	Mora_Real_Pct	Score_Mora_Medio	Valor_Credito_Medio	No_Creditos_Medio	Segmento	Estrategia_Cobranza
Cluster							
0.0000	24800,000	0.0450	0.4940	$ 14.032.899.690	24800	RIESGO MEDIO	Contacto preventivo: SMS + correo antes del vencimiento.
10000	2177,000	0.8170	0.9340	$ 4.759.081.210	2177	ALTO RIESGO	Gestión inmediata: llamada + visita presencial + contacto codeudor



Nota:K optimo seleccionado = 2. Método del codo: K donde la reducción marginal de inercia se aplana. Silhouette Score: mide cohesión intra-cluster vs separación inter-cluster (rango [−1, 1]).

Figura 8.1 Selección del número óptimo de *clusters*: coeficiente de silueta y curva de inercia (método del codo) para K = 2 a 6

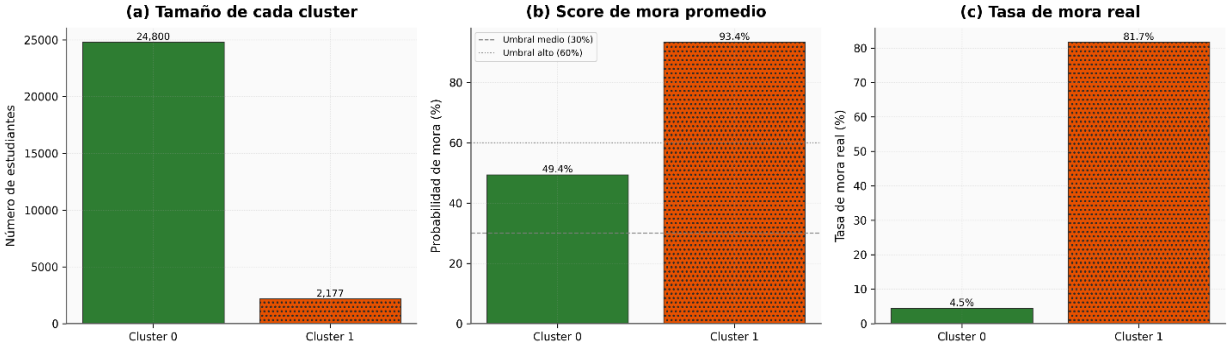


Figura 8.2 Perfil de los *clusters* identificados: distribución de variables clave y score de riesgo por grupo

La integración del *clustering* con las predicciones del modelo permitió complementar la lectura individual del riesgo con una visión más amplia de los perfiles presentes en la cartera. Eso da al Departamento de Apoyo Financiero una base adicional para diferenciar estrategias de intervención según las características del crédito y del estudiante.

8.2 Segmentación operativa por zonas de riesgo

El score de probabilidad del modelo se tradujo en cuatro zonas operativas tipo semáforo, una convención estándar en la gestión de riesgo financiero que deja a cualquier analista, sin importar su formación técnica, interpretar de inmediato el nivel de riesgo de cada crédito. Los umbrales se definieron en función de la distribución del score y de la tasa de mora observada en cada rango, buscando que cada zona tenga un perfil de riesgo claramente diferenciado y una acción operativa asociada.

La zona Verde (score menor al 30%) comprende 1,260 créditos con una tasa de mora real del 2.2%. Estos créditos presentan un riesgo muy bajo y no requieren intervención activa; el

monitoreo rutinario del sistema es suficiente. La zona Amarillo (score 30-60%) agrupa 21,328 créditos con mora real del 3.3%. Aunque el riesgo individual es bajo, el volumen es el más alto de las cuatro zonas, por lo que se recomienda un recordatorio preventivo automático por SMS o correo electrónico con un costo estimado de $500 COP por estudiante. La zona Naranja (score 60-80%) contiene 1,837 créditos con mora real del 14.9%, casi 5 veces el promedio del portafolio.

Esos créditos requieren una intervención más personalizada, como la llamada telefónica del asesor financiero con propuesta de plan de pagos, a un costo estimado de $5.500 COP. La zona Rojo (score 80% o más) agrupa 2,552 créditos con mora real del 73.7%, 33 veces mayor que la zona verde. Esos casos demandan gestión presencial con el estudiante y el codeudor, a un costo de $30.000 COP.

La tasa de mora observada en la zona roja (73,7%) es sustancialmente superior a la registrada en la zona verde (2,2%), con una diferencia de 33 veces, lo que confirma la capacidad discriminante del sistema. La Tabla 8.2 presenta la segmentación completa y la Figura 8.3 la ilustra gráficamente.

Tabla 8.2 Segmentación operativa por zona de riesgo: umbrales, volumen, mora real y acción recomendada

	Zona	Score	Estrategia	Acción	N créditos	Mora real
●	Verde	0 – 30%	Monitoreo	Sin intervención activa	1.260	2,2%
●	Amarillo	30 – 60%	Preventiva	Recordatorio automático por SMS o correo electrónico ($500/est.)	21.328	3,3%
●	Naranja	60 – 80%	Intensiva	Llamada telefónica del asesor financiero + plan de pagos ($5,500/est.)	1.837	14,9%
●	Rojo	80 – 100%	Crítica	Gestión presencial con el estudiante y el codeudor ($30,000/est.)	2.552	73,7%
	TOTAL				26.977	10,7%

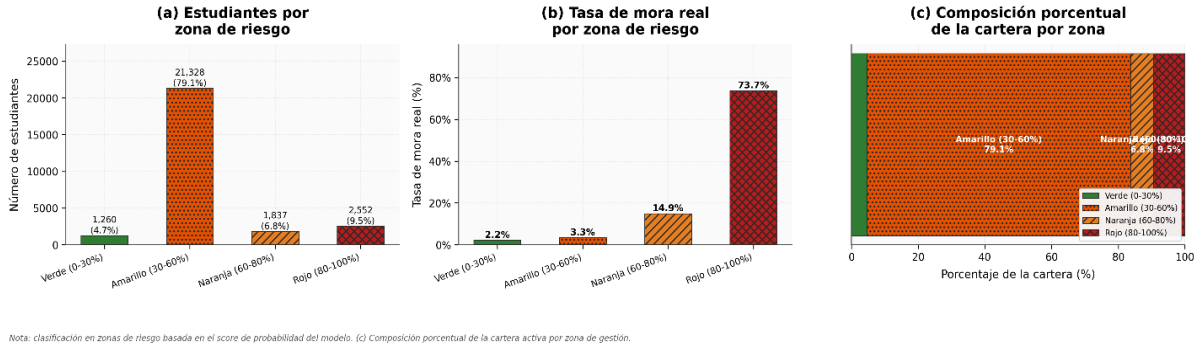


Figura 8.3 Distribución de créditos por zona de riesgo y tasa de mora real observada en cada zona

8.3 Priorización por deciles y curva de ganancia

La curva de ganancia acumulada (gain curve) responde la pregunta clave para la gestión de cobranza: si la Universidad interviene al X% de estudiantes con mayor score de riesgo, ¿qué

porcentaje de los morosos reales captura? [47]. Esta herramienta traduce el modelo predictivo en una regla de priorización directamente aplicable por el equipo de cobranza. Es importante aclarar que la curva de ganancia se construye a nivel obligación (crédito) sobre el *holdout* temporal, mientras que la priorización financiera y el análisis costo-beneficio descrito en la sección 8.5 se agrega después a nivel estudiante usando el mayor score observado entre sus obligaciones activas.

Para construirla, los 26,977 créditos del conjunto de prueba se ordenaron de mayor a menor score de riesgo y se dividieron en diez grupos iguales (deciles). El decil 10 contiene los 2,694 créditos con mayor score (score medio 0.933) y concentra 1,955 morosos reales, lo que equivale al 67.8% del total de morosos con un lift de 6.8x (la concentración de morosos es 6.8 veces mayor que si se seleccionaran créditos al azar). El decil 9 agrega 219 morosos adicionales (8.1% de mora) y eleva la captura acumulada al 75.4%. Los deciles 8 a 1 aportan progresivamente menos morosos, lo que refleja que el modelo asigna scores bajos a créditos con menor riesgo.

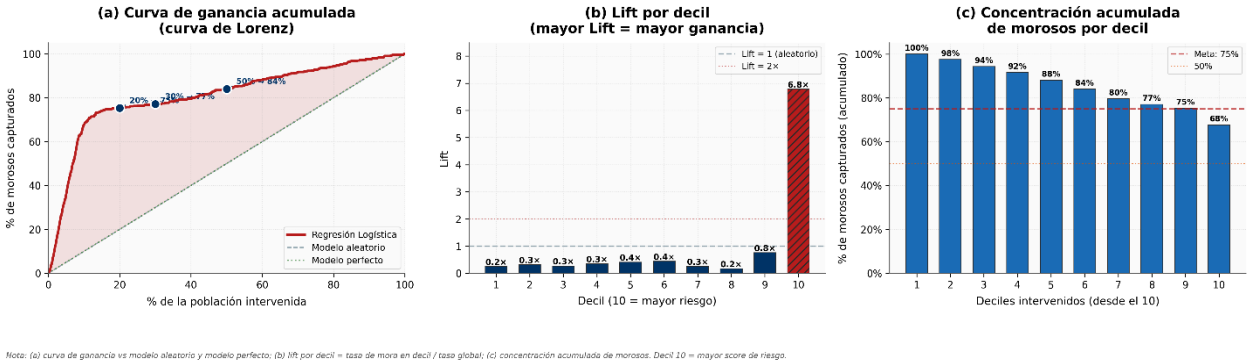
En términos operativos: interviniendo al 10% de mayor riesgo, la Universidad capturaría el 67.8% de los morosos. Al 20%, el 75.4%. Al 30%, el 77.1%. Al 50%, el 84.1%. Así la Dirección Financiera puede decidir cuántos recursos asignar a cobranza en función del porcentaje de morosos que desea capturar, con una curva de rendimientos decrecientes que justifica focalizar los esfuerzos en los segmentos de mayor riesgo. La Tabla 8.3 presenta los deciles completos y la Figura 8.4 ilustra las curvas de ganancia y *lift*.

Tabla 8.3: Tabla de deciles: priorización de cobranza con score medio, tasa de mora, captura acumulada y lift

Decil	N	Morosos	Tasa mora	Score medio	% Acum. morosos	Lift	Prioridad
10	2.694	**1.955**	**72,6%**	**0,933**	**67,8%**	**6,8x**	**Crítica**
9	2.702	219	**8,1%**	**0,627**	**75,4%**	0,8x	**Alta**
8	2.694	48	1,8%	0,552	**77,1%**	0,2x	Media
7	2.699	75	2,8%	0,534	**79,7%**	0,3x	Media
6	2.695	127	4,7%	0,515	**84,1%**	0,4x	Media
5	2.689	119	4,4%	0,492	**88,2%**	0,4x	Media
4	2.710	100	3,7%	0,467	**91,7%**	0,3x	Baja
3	2.698	78	2,9%	0,447	**94,4%**	0,3x	Baja
2	2.696	92	3,4%	0,425	**97,6%**	0,3x	Baja
1	2.700	72	2,7%	0,303	**100,0%**	0,2x	Baja

Interviniendo al 10% de mayor riesgo (decil 10) → se captura el 67.8% de los morosos con lift de 6.8x
Interviniendo al 20% → 75.4% | Al 30% → 77.1% | Al 50% → 84.1%

Nota. Lift = concentración de morosos en el decil respecto al promedio. Lift > 1 indica que el decil concentra más morosos que el azar. El % Acumulado muestra el porcentaje de todos los morosos capturados al intervenir desde el decil 10 hacia abajo. Datos calculados sobre el conjunto de prueba temporal (2023-2024, n = 26,977).



Nota: (a) curva de ganancia vs modelo aleatorio y modelo perfecto; (b) lift por decil = tasa de mora en decil / tasa global; (c) concentración acumulada de morosos. Decil 10 = mayor score de riesgo.

Figura 8.4 Curva de ganancia acumulada (panel izquierdo) y curva de lift (panel derecho). La línea diagonal representa el modelo aleatorio; la distancia entre la curva del modelo y la diagonal es la ganancia informativa

8.4 Simulación de impacto en cartera

La simulación de impacto traduce las métricas del modelo en cifras financieras que la Dirección Financiera y la Rectoría pueden evaluar directamente. Se parte de dos parámetros: la cartera total en mora sin intervención ($2,869,548,878 COP, calculada como la suma de los valores de los créditos que efectivamente entraron en mora en el conjunto de prueba) y una tasa de éxito base de la intervención preventiva del 25%, un supuesto conservador que asume que uno de cada cuatro estudiantes contactados antes del vencimiento generaría una mitigación efectiva de la mora esperada.

Aplicando una tasa de éxito base del 25% sobre la cartera capturada por priorización de score, e interviniendo al 30% de estudiantes con mayor riesgo (1.095 estudiantes únicos), se podría mitigar potencialmente $524.371.130 de cartera morosa, equivalente al 18,3% del total. Esa cifra representa una simulación de mitigación esperada bajo el supuesto base del 25%, distinta del análisis costo-beneficio de la sección 8.5, donde la efectividad se diferencia por zona (Amarillo 15%, Naranja 25%, Rojo 35%) para reflejar la distinta intensidad de cada acción de intervención. La Figura 8.5 muestra la curva completa de mora potencialmente mitigada en función del porcentaje de cartera intervenida.

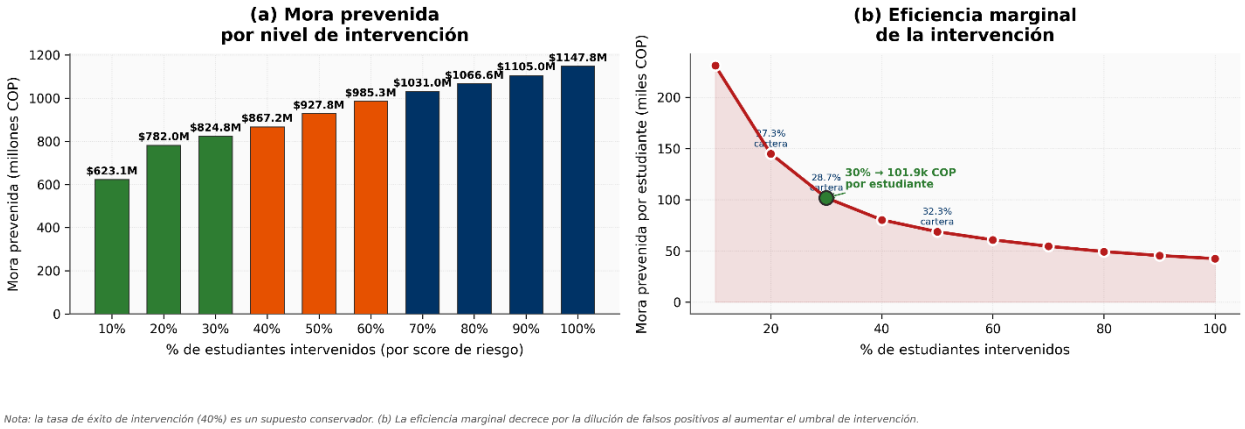


Figura 8.5. Simulación de impacto en cartera a nivel estudiante: mora potencialmente mitigada (en millones COP) por porcentaje de estudiantes intervenidos, bajo escenario base de 25% de efectividad

8.5 Análisis costo-beneficio de la estrategia de cobranza preventiva

Para garantizar coherencia operativa entre el modelo predictivo y la gestión real de cobranza, el análisis costo-beneficio se estructuró a nivel estudiante y no a nivel de obligación individual. Operativamente, el Departamento de Apoyo Financiero contacta al estudiante una sola vez según su nivel de riesgo máximo (**Prob_Mora_Max** sobre sus notas débito activas), aunque tenga varias obligaciones en el *holdout*. El beneficio se modela como mitigación esperada parcial mediante una efectividad diferenciada por zona (Amarillo 15%, Naranja 25%, Rojo 35% en el escenario base), con sensibilidad ±10pp para construir escenarios conservador y optimista. Se eligió esa estructura porque la intensidad de la intervención varía con el riesgo: un SMS preventivo tiene menor conversión que la gestión presencial con codeudor. La interpretación correcta es: simulación de mitigación esperada, no recuperación garantizada.

El análisis costo-beneficio permitió estimar si la estrategia de cobranza preventiva resulta financieramente conveniente para la institución. Para ello se compararon los costos asociados a las acciones de intervención en cada zona de riesgo con el valor estimado de la cartera que podría mitigarse potencialmente mediante una gestión oportuna.

Bajo el rediseño a nivel estudiante con efectividad diferenciada por zona, la zona Rojo (175 estudiantes) implica un costo estimado de $5.250.000 COP y un beneficio de $551.231.982 COP (el retorno sobre la inversión (ROI) +10.400%, efectividad 35%). En la zona Naranja (428 estudiantes), el costo es de $2.354.000 COP con beneficio estimado de $92.817.774 COP (ROI +3.843%, efectividad 25%). La zona Amarillo (2.860 estudiantes) registra un costo de $1.430.000 COP y un beneficio de $131.705.751 COP (ROI +9.110%, efectividad 15%).

En conjunto, bajo el escenario base, la estrategia alcanza un ROI total de +8.487% (rango de sensibilidad: +5.361% conservador a +11.663% optimista). El ROI elevado se explica por la

asimetría entre el costo unitario de intervención (entre $500 y $30.000 por estudiante según zona) y la exposición financiera en mora ($2.870 millones totales). Hay que señalar que se trata de mitigación esperada, no recuperación garantizada.

Tabla 8.4: Análisis costo-beneficio por zona de riesgo: costo de intervención, mora potencialmente mitigada y ROI

	Zona	No. estudiantes	Costo total	Beneficio Base (mitigación esperada)	ROI Base	Cartera en mora real	Retorno por $1
●	**Rojo (≥80%)**	175	$ 5.250.000	**$ 551.231.982**	**+10.400%**	**$ 1.574.948.521**	**$105,0**
●	**Naranja (60–80%)**	428	$ 2.354.000	**$ 92.817.774**	**+3.843%**	**$ 371.271.095**	**$39,4**
●	**Amarillo (30–60%)**	2.860	$ 1.430.000	**$ 131.705.751**	**+9.110%**	**$ 878.038.343**	**$92,1**
●	**Verde (<30%)**	187	$ 0	**$ 0**	*N/A*	**$ 45.290.919**	—
	TOTAL	**3.650**	**$ 9.034.000**	**$ 775.755.508**	**+8.487%**	$ 2.869.548.878	**$85,9**

Tabla 8.5 Análisis de sensibilidad del costo-beneficio.

Escenario	Efectividad por zona	Beneficio esperado	ROI total
Conservador	Am 5% / Nar 15% / Rojo 25%	$493.329.712	+5.361%
Base	**Am 15% / Nar 25% / Rojo 35%**	**$775.755.508**	**+8.487%**
Optimista	Am 25% / Nar 35% / Rojo 45%	$1.062.710.395	+11.663%

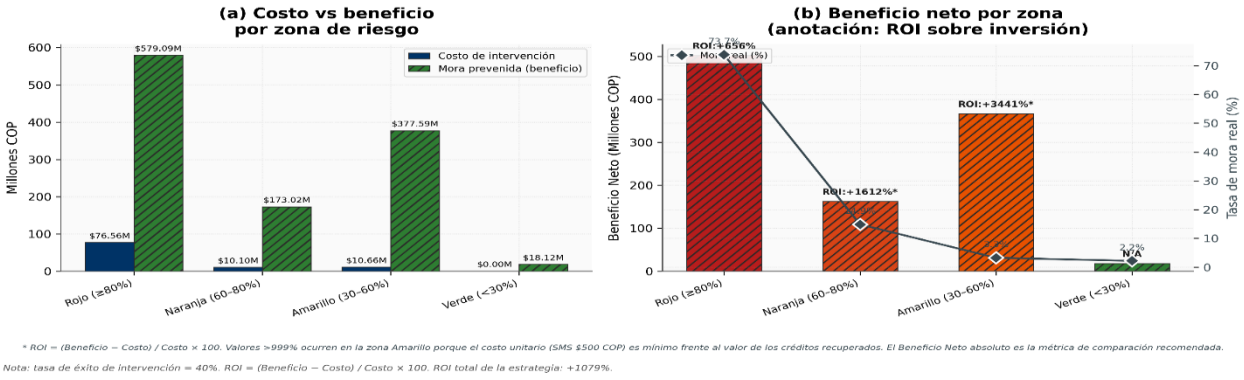


Figura 8.6 Análisis costo-beneficio por zona de riesgo: costo de intervención vs beneficio estimado (mora potencialmente mitigada)

8.6 *Scoring* de la población objetivo 2025

Como cierre del ciclo CRISP-DM, el modelo seleccionado se aplica sobre los créditos del año 2025 que fueron reservados durante la preparación de los datos (sección 6.1.1) por inmadurez de cartera. Estos créditos constituyen la población objetivo del sistema de *scoring* y representan los candidatos a recibir la priorización de cobranza preventiva durante el próximo semestre.

El proceso reaplicó la misma ingeniería de características descrita en el capítulo 6 (codificación binaria de Sexo y Estado del estudiante, escala ordinal del Nivel de Formación y el Estrato Socioeconómico, y construcción del historial de mora a partir de los registros 2015-2024). Luego se aplicó el StandardScaler entrenado sobre el conjunto de *train* y se obtuvo la probabilidad de mora con el modelo operativo (*Random Forest*, AUC-*Holdout* 0,856).

Cada crédito recibió una etiqueta operativa según tres zonas de riesgo, alineadas con la lógica de cobranza diferenciada presentada en la sección 8.2: ALTO si la probabilidad de mora es mayor o igual a 0,60; MEDIO si está entre 0,30 y 0,60; y BAJO si es inferior a 0,30. La Tabla 8.6 presenta la distribución resultante sobre los 14.938 créditos 2025 que conforman la población objetivo.

Tabla 8.6. Distribución de los créditos 2025 por zona de riesgo según el modelo de *scoring*.

Zona	Umbral	N créditos	% Créditos 2025	Acción recomendada
ALTO	Prob_Mora ≥ 0,60	487	3,30%	Gestión intensiva: llamada del asesor financiero, plan de pagos o contacto presencial con el codeudor.
MEDIO	0,30 ≤ Prob_Mora < 0,60	4.453	29,80%	Intervención preventiva: SMS y correo electrónico antes del vencimiento.
BAJO	Prob_Mora < 0,30	9.998	66,90%	Monitoreo pasivo, sin intervención activa.
Total		**14.938**	**100,00%**	—

Nota. Resultados generados con el modelo operativo Random Forest (AUC-Holdout 0,856) aplicado a los 14.938 créditos otorgados durante 2025 que fueron reservados de la muestra de modelado por inmadurez de cartera. Las cifras provienen del archivo scoring_2025_para_cobranza.csv exportado por el notebook en la sección 7.7. Los umbrales de zona se definieron en función de los criterios operativos del Departamento de Apoyo Financiero discutidos en la sección 8.2. Cabe aclarar que aunque el score se calcula originalmente por obligación, la priorización operativa de cobranza debe agregarse a nivel estudiante usando el mayor score observado entre sus obligaciones activas, coherente con la lógica del análisis costo-beneficio descrito en la sección 8.5.

La distribución obtenida es operativamente accionable. El 3,3 % de los créditos marcados como riesgo ALTO (487 obligaciones) concentran la prioridad inmediata del equipo de cobranza, mientras que el 29,8 % en zona MEDIO deja activar campañas masivas de bajo costo (SMS y correo). El 66,9 % restante en zona BAJO no requiere intervención activa, lo que libera recursos para concentrarlos en los segmentos donde el modelo predice mayor pérdida potencial.

El resultado se exportó al archivo *scoring*_2025_para_cobranza.csv con las columnas **Nota_debito, Cliente, Prob_Mora y Zona_Riesgo**, listo para ser consumido por el Departamento de Apoyo Financiero a Estudiantes como insumo del plan de cobranza preventiva del próximo semestre. Este entregable cierra el ciclo CRISP-DM al traducir el modelo validado y evaluado en los capítulos anteriores en una acción concreta sobre la cartera vigente.

8.7 Tablero de control

Como componente final del despliegue, se construyó un tablero de control orientado a apoyar la gestión preventiva del riesgo crediticio en el Departamento de Apoyo Financiero. El tablero consolida los principales resultados del proyecto en una visualización gerencial organizada en

cinco secciones: resumen ejecutivo, diagnóstico de la cartera 2015–2024, explicación del modelo predictivo, estrategia de intervención por zonas de riesgo y plan de acción para los créditos de 2025.

El tablero deja visualizar indicadores clave como el número de créditos analizados, la distribución por zonas de riesgo, la evolución histórica de la tasa de mora, las variables que más explican el riesgo, la estrategia de intervención recomendada y la estimación del retorno esperado de la cobranza preventiva. Así, los resultados técnicos del modelo se traducen en información comprensible y accionable para la toma de decisiones institucionales.

El diseño del tablero sigue principios de visualización orientados a la claridad, la comparación y la priorización de información, de acuerdo con los criterios de Visualization Analysis and Design de Munzner [12]. Se priorizó el uso de indicadores ejecutivos, gráficos comparativos, segmentación por colores tipo semáforo y mensajes interpretativos que facilitan la lectura de los resultados por parte de usuarios no técnicos.

8.8 Monitoreo de *drift* y plan de recalibración

El modelo requiere seguimiento periódico, ya que su desempeño puede verse afectado cuando cambian las características del portafolio con el paso del tiempo. Para revisar ese comportamiento se analizó la estabilidad de las variables entre el período de entrenamiento y el período de prueba, con el fin de identificar posibles cambios relevantes en su distribución.

Los resultados mostraron variaciones importantes en variables como Cuotas, **Nivel_Formacion_Ord, Valor_financiacion y Valor_nota_debito**, mientras que las variables demográficas se mantuvieron estables. Ese comportamiento es coherente con los cambios observados en la cartera después de la pandemia y sugiere que el modelo no debe asumirse como una herramienta fija en el tiempo.

A partir de esos hallazgos se recomienda hacer una recalibración periódica del modelo, idealmente cada dos semestres, incorporando información más reciente. Así se busca mantener su capacidad predictiva y asegurar que continúe siendo útil para la gestión del riesgo crediticio.

8.9 Consideraciones éticas

El uso de modelos predictivos para evaluar el riesgo crediticio de estudiantes universitarios plantea consideraciones éticas que este proyecto reconoce de manera explícita. En materia de confidencialidad, los datos del Departamento de Apoyo Financiero son de carácter interno y se emplearon exclusivamente con fines académicos; los identificadores de estudiantes (**ID_Estudiante**, **Cliente**) se utilizaron solo como claves de agrupación para construir las *features* históricas t→t+1 y no se incluyen como variables del modelo ni se reportan en los resultados.

Sobre la equidad del modelo, el uso del estrato socioeconómico como predictor podría generar

preocupaciones si el modelo fuera sistemáticamente más estricto con estudiantes de estratos bajos. Como se documentó en el Capítulo 5, la relación estrato-mora no es un gradiente lineal simple: los estratos 3-4 presentan mayor mora que los estratos 1-2, lo que sugiere que el modelo no penaliza necesariamente a los estudiantes de menor ingreso. Cualquier implementación real debe monitorear regularmente las métricas de equidad entre grupos demográficos (tasa de falsos positivos por estrato, por sexo, por programa académico) para detectar y corregir sesgos emergentes.

Por último, el modelo produce un score de riesgo que debe ser una entrada para la decisión del asesor financiero, no un reemplazo de esa decisión. Un score alto no debe significar automáticamente la denegación del crédito, sino la asignación de un plan de acompañamiento financiero, un acuerdo de pago escalonado o una conversación con el estudiante sobre sus circunstancias. La institución tiene responsabilidad ética de usar el modelo para apoyar y proteger al estudiante, no para excluirlo financieramente.

CONCLUSIONES

El proyecto buscó predecir el riesgo de incumplimiento en créditos educativos siguiendo la metodología CRISP-DM. El trabajo partió de la información histórica de créditos educativos de la institución para el periodo 2015–2024. A partir de esos registros se caracterizó la población beneficiaria, se preparó el conjunto de datos para modelado predictivo, se evaluaron algoritmos de clasificación y se plantearon herramientas de apoyo para la gestión preventiva del riesgo.

El primer objetivo específico buscaba caracterizar a los beneficiarios de crédito educativo y los factores asociados al riesgo de incumplimiento. El análisis exploratorio mostró que la mora no depende de un solo factor; aparece más bien por la combinación de condiciones financieras, académicas, temporales y de comportamiento. El número de cuotas, el nivel de formación, el año del crédito y el historial de pago tuvieron relación con el riesgo. El estrato socioeconómico, además, no siguió un patrón lineal. Esto indica que el riesgo crediticio en educación superior no se explica solo por la capacidad económica aparente del estudiante, sino por las condiciones de financiación, permanencia y apoyo institucional.

Con el segundo objetivo específico, dedicado a preparar y estructurar los datos para el análisis predictivo, se obtuvo una base de modelado más adecuada para anticipar la mora. En esa revisión quedaron por fuera variables que contenían información posterior o simultánea al evento de incumplimiento; también se retiraron campos que parecían útiles al comienzo, pero eran identificadores operativos del sistema. Al construir variables históricas, el conjunto de datos incorporó antecedentes de comportamiento de pago y ganó utilidad analítica. Esto coincide con buenas prácticas de modelado predictivo, en especial en problemas de credit scoring, donde la prevención del data leakage y la definición temporal correcta de las variables son necesarias para usar los modelos en escenarios reales.

El tercer objetivo específico se orientó a evaluar modelos de clasificación para predecir el riesgo de incumplimiento. Los resultados indicaron que los datos internos de la institución sí permiten anticipar perfiles con mayor probabilidad de mora. La Regresión Logística resultó útil para interpretar, desde la gestión, los factores de riesgo; Random Forest tuvo mejor desempeño cuando el propósito era el scoring masivo. Esa combinación atiende dos necesidades distintas: explicar el fenómeno y disponer de una herramienta operativa para priorizar el seguimiento. El uso de modelos de clasificación en riesgo crediticio mantiene coherencia con la literatura sobre credit scoring y con estudios recientes que aplican aprendizaje supervisado a decisiones financieras.

El cuarto objetivo específico abordó el diseño de herramientas para apoyar la gestión del riesgo crediticio. La segmentación por zonas tipo semáforo tradujo los resultados del modelo a criterios prácticos de priorización. Con esta clasificación, la institución puede concentrar la cobranza preventiva en estudiantes con mayor probabilidad de incumplimiento y usar mejor sus recursos. El tablero de control y el análisis costo-beneficio propuesto también ofrecen una base para tomar

decisiones; sus resultados deben leerse como estimaciones bajo supuestos de efectividad, no como valores de recuperación garantizada. Al visualizar resultados y segmentar perfiles, los modelos predictivos quedan más cerca de usuarios no técnicos y de los procesos de gestión institucional.

La respuesta a la pregunta de investigación está en la utilidad de la Ciencia de Datos para convertir registros históricos de crédito educativo en información predictiva y accionable. Con la integración, depuración, modelado y visualización de los datos, el proyecto identificó patrones asociados a la mora, estimó la probabilidad de incumplimiento y clasificó los créditos en zonas de riesgo. Así, el enfoque analítico apoya el paso de una gestión reactiva de la cartera a una gestión preventiva, segmentada y basada en evidencia.

Visto en conjunto, el proyecto muestra que una institución de educación superior puede avanzar hacia una gestión preventiva del riesgo crediticio con datos internos, herramientas de código abierto y una metodología replicable. Su aporte principal va más allá del desempeño técnico de los modelos: vincula la analítica de datos con decisiones concretas del Departamento de Apoyo Financiero, como la priorización de casos, el seguimiento temprano, la segmentación de estudiantes y el monitoreo de cartera.

El desarrollo del proyecto dejó ver algunas limitaciones. La base de datos entregada no incluye variables como la edad del estudiante ni información detallada del codeudor, por ejemplo, ocupación, ingresos o tipo de vinculación laboral. Incluir ese tipo de variables podría mejorar la capacidad predictiva del modelo, sobre todo en estudiantes de primer ingreso que todavía no tienen historial crediticio institucional. La literatura de credit scoring reconoce que la información socioeconómica, financiera y comportamental ayuda a estimar mejor el riesgo [31], [41].

La variable Estado_Estudiante también puede tener una limitación temporal. En la base analizada corresponde al estado disponible al momento de extraer la información, no necesariamente al estado académico exacto cuando se otorgó cada crédito. Se usó Es_Activo como aproximación; aun así, contar con el estado académico histórico por periodo daría una medición más precisa del riesgo.

El sistema de scoring y el tablero de control quedaron desarrollados como prototipos funcionales. Para llevarlos a la operación se requeriría una fase posterior de integración tecnológica, validación institucional y definición de protocolos de uso. Como continuación del trabajo, el modelo puede fortalecerse con nuevas fuentes de información y con variables socioeconómicas y académicas adicionales. También queda pendiente validar su desempeño en nuevos periodos y avanzar hacia una implementación institucional que integre la analítica predictiva a la gestión cotidiana del crédito educativo.

Desde el punto de vista metodológico, la validez del modelo no se fundamenta únicamente en el desempeño predictivo medido por AUC-ROC. El proyecto incorporó una evaluación amplia que consideró métricas de clasificación relevantes para problemas desbalanceados, como Recall,

Precision y F1-score, así como esquemas de validación temporal, validación cruzada y validación out-of-time para analizar la estabilidad del modelo en periodos no vistos durante el entrenamiento.

En relación con el tratamiento del desbalanceo, el uso de SMOTE se asumió como una decisión metodológica orientada a mejorar la representación de la clase minoritaria durante el entrenamiento, aplicándolo únicamente sobre los datos de entrenamiento y conservando los conjuntos de evaluación con su distribución real. No obstante, se reconoce como limitación que no se ejecutó un experimento comparativo formal entre modelos con y sin SMOTE; por tanto, no se concluye que esta técnica mejore necesariamente el desempeño del modelo. Su efecto sobre la discriminación, la calibración y la estabilidad deberá evaluarse en futuras iteraciones, utilizando datos reales no remuestreados.

En consecuencia, el modelo debe entenderse como una herramienta de apoyo a la gestión institucional del riesgo, útil para priorizar acciones preventivas y orientar la toma de decisiones, pero sujeta a monitoreo, evaluación de calibración, recalibración periódica y revisión por parte de los responsables del proceso financiero.

REFERENCIAS BIBLIOGRÁFICAS

Las siguientes referencias bibliográficas fueron utilizadas en el desarrollo de este documento.

[1] Ministerio de Educación Nacional, “Estadísticas de deserción,” Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES), 2025. [En línea]. Disponible en: https://www.mineducacion.gov.co/sistemasinfo/spadies/secciones/Estadisticas-de-desercion/. [Consultado: 19-may-2026].

[2] A. L. Leal Fica, M. A. Aranguiz Casanova y J. Gallegos Mardones, “Análisis de riesgo crediticio, propuesta del modelo credit scoring,” *Rev. Fac. Cienc. Econ.*, vol. 26, n.º 1, pp. 181–207, jun. 2018, doi: 10.18359/rfce.2666.

[3] Superintendencia Financiera de Colombia, “Circular Básica Contable y Financiera: Circular Externa 100 de 1995,” Cap. II, numeral 1.1, “Riesgo crediticio (RC).” [En línea]. Disponible en: https://www.superfinanciera.gov.co/publicaciones/15466/normativanormativa-generalcircular-basica-contable-y-financiera-circular-externa-de-15466/. [Consultado: 19-may-2026].

[4] C. E. Villacrés Armas y O. Pérez Barral, “Riesgo crediticio y proceso de cobranzas en instituciones privadas de educación,” *Contabilidad y Negocios*, vol. 19, n.º 38, pp. 119–140, 2024, doi: 10.18800/contabilidad.202402.005.

[5] C. C. Aggarwal, *Data Mining: The Textbook*. Cham, Switzerland: Springer International Publishing, 2015, doi: 10.1007/978-3-319-14142-8.

[6] L. C. Thomas, “A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers,” *International Journal of Forecasting*, vol. 16, n.º 2, pp. 149–172, 2000, doi: 10.1016/S0169-2070(00)00034-0.

[7] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[9] P. Berkhin, “A survey of clustering data mining techniques,” en *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas y M. Teboulle, Eds. Berlin/Heidelberg, Germany: Springer-Verlag, 2006, pp. 25–71, doi: 10.1007/3-540-28349-8_2.

[10] S. M. Lundberg y S.-I. Lee, “A unified approach to interpreting model predictions,” en *Advances in Neural Information Processing Systems 30* (NeurIPS 2017), Long Beach, CA, EE. UU., 2017, pp. 4765–4774. [En línea]. Disponible en: https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions. [Consultado: 09-may-2026].

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Comput. Surv., vol. 54, no. 6, Art. 115, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.

[12] T. Munzner, Visualization Analysis and Design. Boca Raton, FL, USA: CRC Press, A K Peters Visualization Series, 2014.

[13] D. García Pérez de Lema, El riesgo financiero de la pequeña y mediana empresa en Europa. Madrid, España: Pirámide, 2004.

[14] S. A. Ross, R. W. Westerfield, J. F. Jaffe, y F. López Herrera, Finanzas corporativas, 9.ª ed. México: McGraw-Hill, 2012.

[15] W. H. Beaver, "Financial Ratios as Predictors of Failure," J. Account. Res., vol. 4, pp. 71–111, 1966, doi: 10.2307/2490171.

[16] M. Tamari, "Financial Ratios as a Means of Forecasting Bankruptcy," Manag. Int. Rev., vol. 6, no. 4, pp. 15–21, 1966.

[17] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," J. Finance, vol. 23, no. 4, pp. 589–609, 1968, doi: 10.2307/2978933.

[18] G. A. Romani Chocce, P. Aroca González, N. Aguirre Aguirre, P. Leiton Vega y J. Muñoz Carrazana, "Modelos de clasificación y predicción de quiebra de empresas: una aplicación a empresas chilenas," *Forum Empresarial*, vol. 7, n.º 1, pp. 1–20, jun. 2002, doi: 10.33801/fe.v7i1.3371.

[19] E. Arias Ortiz, G. Elacqua, Á. López, J. Téllez Fuentes, R. Peralta Castro, M. Ojeda, Y. Blanco Morales, F. Pedró, D. Vieira do Nascimento y J. F. Roser Chinchilla, "Educación superior y COVID-19 en América Latina y el Caribe: financiamiento para los estudiantes," Banco Interamericano de Desarrollo, Nota Técnica IDB-TN-02206, 2021, doi: 10.18235/0003380. [En línea]. Disponible en: https://publications.iadb.org/es/educacion-superior-y-covid-19-en-america-latina-y-el-caribe-financiamiento-para-los-estudiantes. [Consultado: 19-may-2026].

[20] F. J. Madrigal Moreno, L. Chávez Contreras y A. Díaz Vázquez, "Evaluación de las 5 C's de crédito en condiciones de incertidumbre," en *Estudios Organizacionales en las Ciencias Administrativas ante los Retos del Siglo XXI*, Morelia, Michoacán, México, 2017. [En línea]. Disponible en: https://www.teczamora.mx/documentos/posgrado_investigacion/articulos/Evaluacion%20de%20las%205%20C%27s%20de%20cr%C3%A9dito%20en%20condiciones%20de%20incertidumbre.pdf. [Consultado: 19-may-2026].

[21] Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior (ICETEX), "Manual del Sistema de Administración de Riesgo Crediticio, v7," ICETEX, Bogotá, Colombia, 2023. [En línea]. Disponible en: https://web.icetex.gov.co/documents/20122/1096965/manual-sistema-de-administracion-de-riesgos-crediticio-v7.pdf. [Consultado: 19-may-2026].

[22] M. A. Granda Rodríguez, "Determinantes del riesgo de incumplimiento en créditos educativos: un análisis para Colombia (ICETEX)," tesis de maestría, Escuela de Economía y Finanzas, Universidad EAFIT, Medellín, Colombia, 2020. [En línea]. Disponible en: https://repository.eafit.edu.co/handle/10784/24951. [Consultado: 19-may-2026].

[23] Y. M. Zapata Rojas, "Guía metodológica para la aplicación de modelos de medición de riesgo de crédito," tesis de maestría, Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia, 2010. [En línea]. Disponible en: https://repository.unab.edu.co/handle/20.500.12749/13823. [Consultado: 19-may-2026].

[24] O. A. Urrutia Dorado y J. A. Franco Vanegas, “Análisis predictivo del score de riesgo crediticio mediante Machine Learning: una herramienta para la toma de decisiones financieras,” tesis de maestría, Universidad Nacional Abierta y a Distancia (UNAD), Colombia, dic. 2024. [En línea]. Disponible en: https://repository.unad.edu.co/handle/10596/67365. [Consultado: 19-may-2026].

[25] F. Montalván, “Credit scoring, aplicando técnicas de regresión logística y redes neuronales, para una cartera de microcrédito,” tesis de maestría, Universidad Andina Simón Bolívar, Sede Ecuador, Quito, Ecuador, 2019. [En línea]. Disponible en: https://repositorio.uasb.edu.ec/handle/10644/6872. [Consultado: 19-may-2026].

[26] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, no. 3, pp. 37–54, 1996, doi: 10.1609/aimag.v17i3.1230.

[27] A. M. Shimaoka, R. C. Ferreira, and A. Goldman, "The Evolution of CRISP-DM for Data Science: Methods, Processes and Frameworks," SBC Reviews on Computer Science, vol. 4, no. 1, pp. 28–43, Oct. 2024, doi: 10.5753/reviews.2024.3757.

[28] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth, “CRISP-DM 1.0: step-by-step data mining guide,” CRISP-DM Consortium, 2000. [En línea]. Disponible en: https://www.kde.cs.uni-kassel.de/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf. [Consultado: 19-may-2026].

[29] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining (PADD'00), Manchester, U.K., Apr. 2000, pp. 29–40.

[30] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," J. Oper. Res. Soc., vol. 54, no. 6, pp. 627–635, Jun. 2003, doi: 10.1057/palgrave.jors.2601545.

[31] L. C. Thomas, J. N. Crook, and D. B. Edelman, Credit Scoring and Its Applications, 2nd ed. Philadelphia, PA, USA: SIAM, 2017, doi: 10.1137/1.9781611974560.

[32] C. Schröer, F. Kruse, and J. Marx Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Comput. Sci., vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[33] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana y P. Flach, “CRISP-DM twenty years later: from data mining processes to data science trajectories,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, n.º 8, pp. 3048–3061, ago. 2021, doi: 10.1109/TKDE.2019.2962680.

[34] J. G. Salazar Vergara, "Diseño de un modelo predictivo para otorgar créditos," Semestre Económico, vol. 24, no. 57, pp. 320–347, Jul.–Dec. 2021, doi: 10.22395/seec.v24n57a13.

[35] D. Borrero-Tigreros and O. F. Bedoya-Leiva, "Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial," Rev. UIS Ing., vol. 19, no. 4, pp. 37–52, 2020, doi: 10.18273/revuin.v19n4-2020004.

[36] C. M. Betancur Londoño, "Modelo de aprendizaje automático para riesgo crediticio de microempresarios regionales según perfil socioeconómico," Trabajo de grado de especialización,

Esp. Estadística Aplicada, Fundación Univ. Los Libertadores, Bogotá, Colombia, 2022. [Online]. Available: http://hdl.handle.net/11371/4733

[37] L. J. Mester, "What's the Point of Credit Scoring?," Bus. Rev., Federal Reserve Bank of Philadelphia, no. Sep/Oct, pp. 3–16, 1997.

[38] L. C. Guerrero Bernal, "Optimización del riesgo crediticio: Predicción de morosidad a través de *Machine Learning*," trabajo de grado de especialización, Esp. Estadística Aplicada, Fundación Univ. Los Libertadores, Bogotá, Colombia, oct. 2024. [En línea]. Disponible en: https://repository.libertadores.edu.co/items/1d840456-4a9a-4fa1-a6cf-abab81fe92b3 [Consultado: 19-may-2026].

[39] E. Daza Alzate, "Modelo predictivo para la detección de fraudes financieros y estimación del riesgo crediticio en tiempo real en el sector bancario," trabajo de grado de especialización, Especialización en Ciencia de Datos y Analítica, Universidad Nacional Abierta y a Distancia (UNAD), Colombia, 2025. [En línea]. Disponible en: https://repository.unad.edu.co/handle/10596/79330. [Consultado: 19-may-2026].

[40] Basel Committee on Banking Supervision, "Basel II: international convergence of capital measurement and capital standards: a revised framework—comprehensive version," Bank for International Settlements, Basel, Switzerland, jun. 2006. [En línea]. Disponible en: https://www.bis.org/publ/bcbs128.htm. [Consultado: 19-may-2026].

[41] B. Baesens, D. Rösch, and H. Scheule, Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS (Wiley and SAS Business Series). Hoboken, NJ, USA: John Wiley & Sons, 2016.

[42] F. Pedregosa "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [En línea]. Disponible en: https://jmlr.org/papers/v12/pedregosa11a.html. [Consultado: 19-may-2026].

[43] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," Eur. J. Oper. Res., vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.

[44] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[45] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[46] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD'16), San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[47] D. J. Hand, "Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve," Mach. Learn., vol. 77, no. 1, pp. 103–123, Oct. 2009, doi: 10.1007/s10994-009-5119-5.

LISTA DE ANEXOS

Los anexos digitales del proyecto incluyen el notebook desarrollado en Python/Jupyter, disponible en un repositorio de GitHub, en el cual se documentan los procedimientos técnicos utilizados para la preparación de los datos, el análisis exploratorio, la construcción de variables, el entrenamiento de los modelos predictivos, el proceso de *scoring* y la generación de resultados asociados al proyecto.

Este archivo se incorpora con el fin de facilitar la trazabilidad metodológica, la validación técnica y la consulta del flujo analítico aplicado en el marco del proyecto. El notebook permite revisar de manera estructurada el desarrollo del análisis, incluyendo código, salidas, tablas, gráficos y resultados generados durante la ejecución del proyecto

También incluimos la versión exportada en PDF del tablero de control desarrollado en Power BI. Este tablero presenta de manera ejecutiva los principales resultados del modelo de scoring de crédito educativo, el diagnóstico de la cartera 2015–2024, la segmentación de los créditos por zonas de riesgo, la estrategia de intervención preventiva y el plan de acción propuesto para la gestión de los créditos 2025. El archivo se incorpora como evidencia visual de la herramienta diseñada para facilitar la interpretación de los resultados y apoyar la toma de decisiones institucionales en la gestión del riesgo y la cobranza.

Archivo digital anexo:

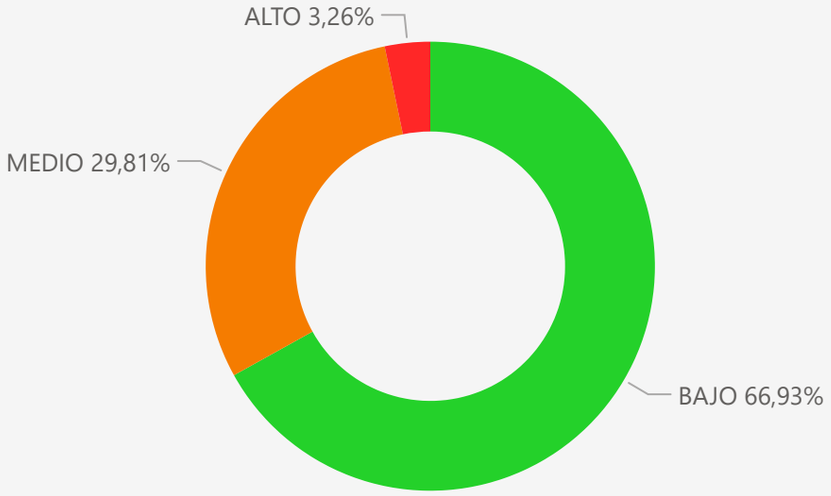
https://github.com/Nabetse1109/UAO_riesgo_credito/blob/main/UAO_Riesgo_Crediticio_CRISP_DM_v15_FINAL2.ipynb

Dashboard_UAO_Riesgo_Crediticio.pdf

GESTIÓN DEL RIESGO Y COBRANZA — SCORING DE CRÉDITO EDUCATIVO UAO

14,938	$775,755,508	$9,034,000	+8.487%
Créditos 2025	Mora evitada	Inversión cobranza	Retorno

Distribución de créditos 2025 por zona de riesgo



LOGROS DEL MODELO

✓ Detecta 85 de cada 100 morosos reales

✓ Validado con 10 años de información (2015-2024)

✓ Estable en períodos no entrenados (2024)

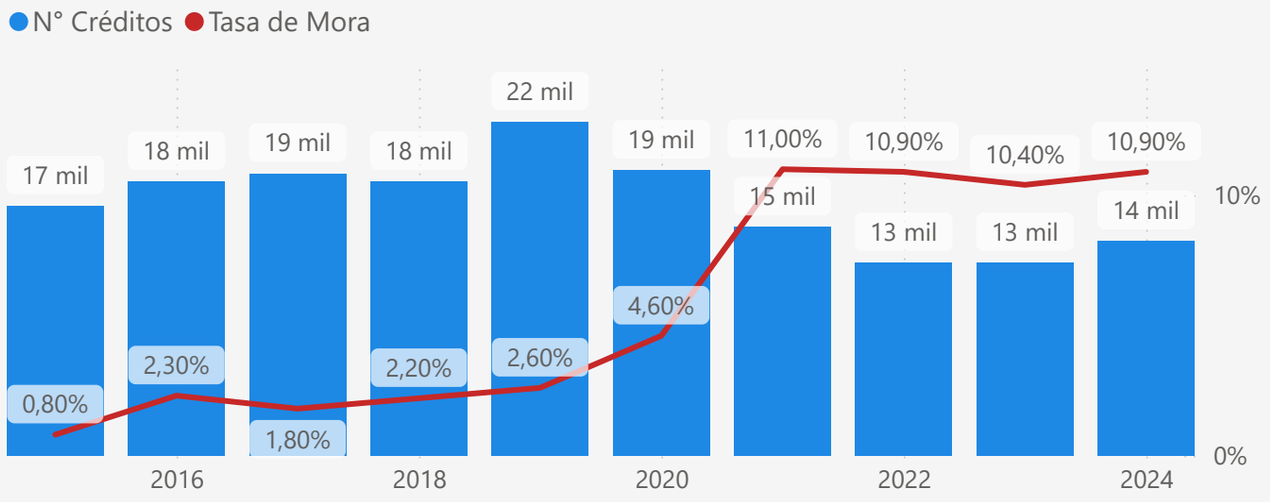
✓ Retorno esperado de $85,90 por cada $1 invertido

El modelo permite intervenir preventivamente sobre los 487 créditos de alto riesgo del 2025, con un retorno esperado de 85,9 a 1.

DIAGNÓSTICO — CARTERA UAO 2015–2024

5,18%	14,087	167,275	2021 (11,0%)
Tasa de mora global	Estudiantes Únicos	Creditos Históricos	Cohorte Crítica

Evolución de la tasa de mora por cohorte anual (2015-2024)



Año	N° Créditos	Tasa de Mora	Fase
2015	16.530	0,8%	Baja mora
2016	18.121	2,3%	Baja mora
2017	18.656	1,8%	Baja mora
2018	18.130	2,2%	Baja mora
2019	22.078	2,6%	Baja mora
2020	18.881	4,6%	Ruptura COVID
2021	15.124	11,0%	Ruptura COVID
2022	12.778	10,9%	Ruptura COVID
2023	12.771	10,4%	Nuevo régimen
2024	14.206	10,9%	Nuevo régimen

EL PLAZO ES EL PRINCIPAL FACTOR DE RIESGO

Corto plazo: **1,8%** de mora

Largo plazo: **25,0%** de mora

Los créditos a largo plazo tienen 13,9 veces más probabilidad de caer en mora que los de corto

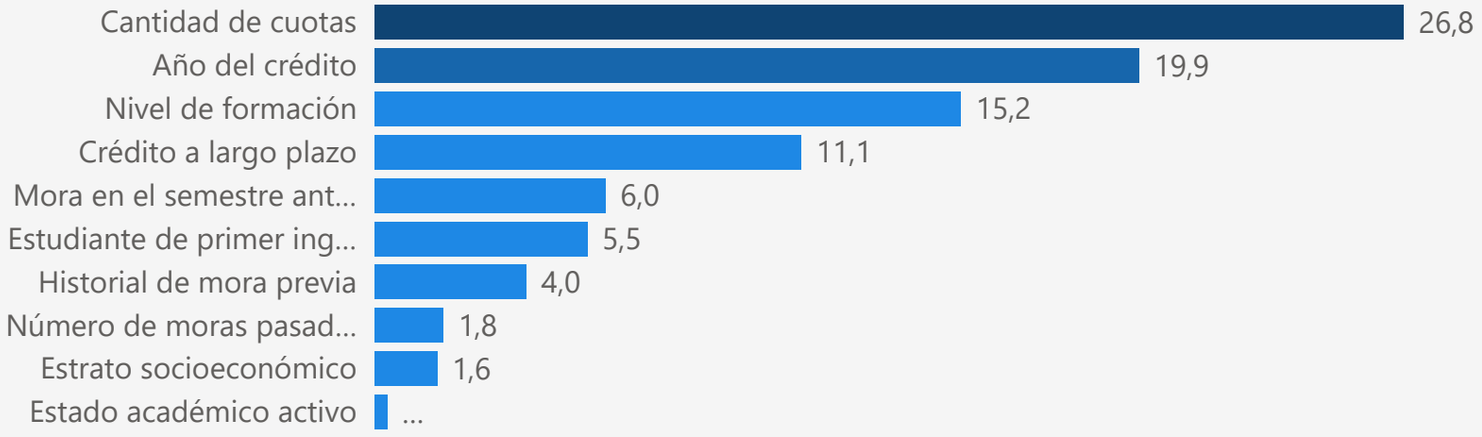
¿CÓMO IDENTIFICAMOS EL RIESGO?

85 de cada 100
Morosos detectados

Alta
Confiabilidad del modelo

10 años
Validación temporal

Variables que más explican el riesgo de mora (%)



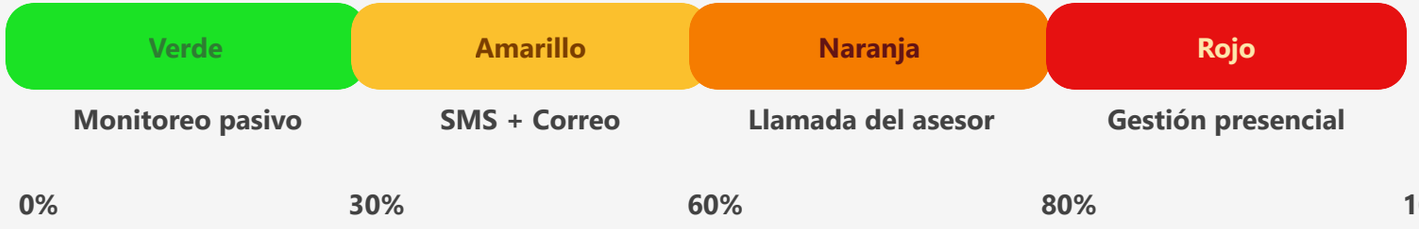
COMPARATIVO DE ENFOQUES

Modelo manual (revisión tradicional)
Detecta solo: **30 de cada 100 morosos**

Modelo UAO (Ciencia de Datos)
Detecta: **85 de cada 100 morosos**

El modelo identifica casi 3 veces más estudiantes en riesgo que la revisión tradicional, permitiendo intervenir antes de que ocurra el incumplimiento.

Cada estudiante recibe un puntaje entre 0% y 100% que indica su probabilidad de incumplir



El modelo fue construido analizando 167.275 créditos durante 10 años. Su confiabilidad fue validada en datos del año 2024 que nunca vio durante su entrenamiento.

ESTRATEGIA DE INTERVENCIÓN POR ZONAS

$9,034,000	$775,755,508	+8.487%	$85,90
Inversión cobranza	Mora evitada	Retorno	Retorno por $1

Costo de intervención vs beneficio estimado por zona

Zona	N° Estudiantes	N° de Créditos	Total Mora	Mora Real	Inversión cobranza	Retorno	Retorno por $1	Mora evit. (Conservador)	Mora evit. (Base)	Mora evit. (Optimista)
Rojo (≥80%)	175	3.307	$1.574.948.521	88,6%	$5.250.000	10400%	$105,0	$393.737.130	$551.231.982	$708.726.834
Amarillo (30–60%)	2.860	20.470	$878.038.343	5,5%	$1.430.000	9110%	$92,1	$43.901.917	$131.705.751	$219.509.586
Naranja (60–80%)	428	2.330	$371.271.095	16,4%	$2.354.000	3843%	$39,4	$55.690.664	$92.817.774	$129.944.883
Verde (<30%)	187	870	$45.290.919	4,8%	$0	0%	$0,0	$0	$0	$4.529.092
Total	**3.650**	**26.977**	**$2.869.548.878**	**10,7%**	**$9.034.000**	**8487%**	**$85,9**	**$493.329.711**	**$775.755.507**	**$1.062.710.395**

Por cada $1 invertido en cobranza preventiva, la UAO podría recuperar $85,9 en mora evitada. Retorno total: **+8.487%.**

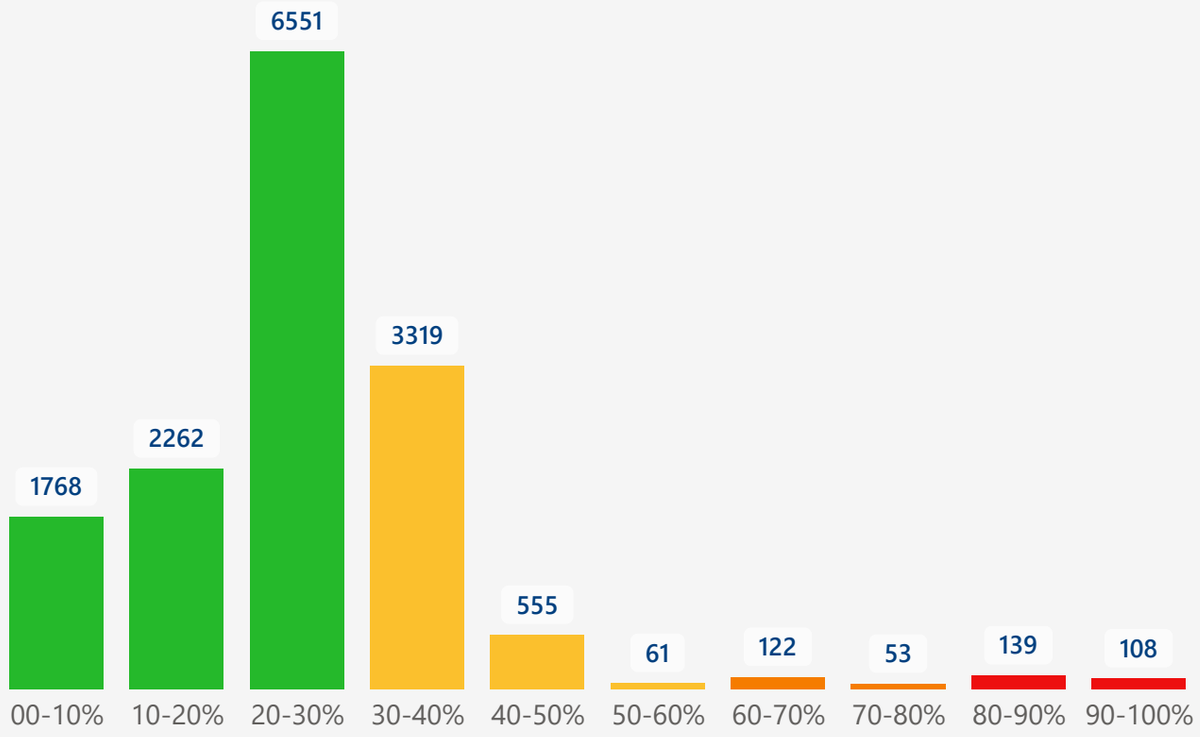
PLAN DE ACCIÓN 2025

1,385 Estudiantes — **9,998** MONITOREO PASIVO (Bajo)

789 Estudiantes — **4,453** INTERVENCIÓN PREVENTIVA (Medio)

88 Estudiantes — **487** ACCIÓN INMEDIATA (Alto)

Distribución de los 14.938 créditos 2025 por probabilidad de mora



Zona	Acción Recomendada	N° Estudiantes	N° Créditos	Costo Estudiante
ALTO	Gestion intensiva del asesor	88	487	$2.640.000
MEDIO	SMS y correo preventivo	789	4453	$394.500
BAJO	Monitoreo pasivo	1.385	9998	$0
Total		**2.262**	**14938**	**$3.034.500**

PRÓXIMOS PASOS PARA EL DEPARTAMENTO DE APOYO FINANCIERO

- **Semana 1:** Contactar a los 88 estudiantes de alto riesgo y agendar gestión presencial con el estudiante y su codeudor.
- **Semana 2:** Configurar campaña masiva de SMS y correos para los 789 estudiantes de riesgo medio.
- **Mes 1:** Activar el monitoreo automático para los 1385 estudiantes de bajo riesgo
- **Cada 2 semestres:** Recalibrar el modelo con datos recientes

Este es el plan de gestión para los 14.938 créditos otorgados durante 2025. El modelo se aplica antes del primer vencimiento, lo que permite intervenir preventivamente sobre los casos de mayor riesgo y reducir la mora antes de que ocurra.