FACULTAD DE POSTGRADO
TRABAJO FINAL DE GRADUACIÓN

PREDICCIÓN DE RIESGO DE IMPAGO EN INSTITUCIÓN FINANCIERA USANDO MODELOS DE MACHINE LEARNING

SUSTENTADO POR:

JOSE MANUEL GARCIA HERNANDEZ
WALTHER NAHUN TORRES MORENO

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN
ANALÍTICA DE NEGOCIOS

TEGUCIGALPA, FRANCISCO
MORAZÁN, HONDURAS, C.A.

SEPTIEMBRE, 2023

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTORA
ROSALPINA RODRÍGUEZ GUEVARA

PRORECTOR/ SECRETARIO GENERAL

ROGER MARTÍNEZ MIRALDA

VICERRECTOR ACADÉMICO
JAVIER ABRAHAM SALGADO LEZAMA

DIRECTORA NACIONAL DE POSTGRADO
ANA DEL CARMEN RETTALLY

PREDICCIÓN DE RIESGO DE IMPAGO EN INSTITUCIÓN FINANCIERA USANDO MODELOS DE MACHINE LEARNING

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE

MÁSTER EN

ANALÍTICA DE NEGOCIOS

ASESOR

HENRY ANTONIO OSORTO RUIZ

MIEMBROS DE LA TERNA:

CARLOS AMADOR
JOSUÉ MEJÍA
GERARDO LUJANO

DERECHOS DE AUTOR

AUTORIZACIÓN DEL AUTOR(ES) PARA LA CONSULTA, REPRODUCCIÓN PARCIAL O TOTAL Y PUBLICACIÓN ELECTRÓNICA DEL TEXTO COMPLETO DE TESIS DE POSTGRADO

Señores

CENTRO DE RECURSOS PARA **EL APRENDIZAJE Y LA INVESTIGACIÓN (CRAI)**

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA (UNITEC)

Estimados Señores:

Yo, José Manuel García y Walther Nahun Torres, de Tegucigalpa, autores del trabajo de postgrado titulado: Predicción de riesgo de impago en institución financiera usando modelos de machine learning, presentado y aprobado en octubre del 2023, como requisito previo para optar al título de máster en Analítica de Negocios y reconociendo que la presentación del presente documento forma parte de los requerimientos establecidos del programa de maestrías de la Universidad Tecnológica Centroamericana (UNITEC), por este medio autorizo a las Bibliotecas de los Centros de Recursos para el Aprendizaje y la Investigación (CRAI) de UNITEC, para que con fines académicos puedan libremente registrar, copiar o utilizar la información contenida en él, con fines educativos, investigativos o sociales de la siguiente manera:

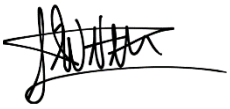
1) Los usuarios puedan consultar el contenido de este trabajo en las salas de estudio de la biblioteca y/o la página Web de la Universidad.
2) Permita la consulta y/o la reproducción a los usuarios interesados en el contenido de este trabajo, para todos los usos que tengan finalidad académica, ya sea en formato CD o digital desde Internet,

Intranet, etc., y en general en cualquier otro formato conocido o por conocer.

De conformidad con lo establecido en los artículos 9.2, 18, 19, 35 y 62 de la Ley de Derechos de Autor y de los Derechos Conexos; los derechos morales pertenecen al autor y son personalísimos, irrenunciables, imprescriptibles e inalienables. Asimismo, el autor cede de forma ilimitada y exclusiva a UNITEC la titularidad de los derechos patrimoniales. Es entendido que cualquier copia o reproducción del presente documento con fines de lucro no está permitida sin previa autorización por escrito de parte de UNITEC.

En fe de lo cual se suscribe el presente documento en la ciudad de Tegucigalpa, a los 26 días del mes de septiembre del año 2023

José Manuel García Hernández	11623063
Nombre completo	**Número de cuenta**



Walther Nahun Torres Moreno	12213071
Nombre completo	**Número de cuenta**

*** La autorización firmada se encuentra adjunta a mí expediente**

FACULTAD DE POSTGRADO

PREDICCIÓN DE RIESGO DE IMPAGO EN INSTITUCIÓN FINANCIERA USANDO MODELOS DE MACHINE LEARNING

JOSE MANUEL GARCIA HERNANDEZ
WALTHER NAHUN TORRES MORENO

Resumen

Esta investigación se centra en la identificación de las variables clave que influyen en el riesgo de impago en la cartera de consumo fiduciario de instituciones financieras en Honduras. A través de un análisis exhaustivo de datos históricos, se logró destacar las principales variables predictoras. Posteriormente, se desarrolló y aplicó un modelo de machine learning altamente preciso que utiliza estas variables para anticipar el riesgo de impago. Este modelo ha demostrado una capacidad excepcional al predecir con éxito aproximadamente el 60% de los clientes propensos a caer en impago. Además de mejorar la gestión del riesgo crediticio, la implementación de este enfoque promete reducir costos, optimizar recursos y fortalecer la toma de decisiones en el sector financiero hondureño.

Palabras claves: Cartera de consumo fiduciario, Machine learning, Modelos predictivos, Morosidad, Riesgo crediticio.

GRADUATE SCHOOL

PREDICTION OF DEFAULT RISK IN A FINANCIAL INSTITUTION USING MACHINE LEARNING MODELS

JOSE MANUEL GARCIA HERNANDEZ
WALTHER NAHUN TORRES MORENO

Abstract

X This research focuses on identifying the key variables that influence the default risk in the fiduciary consumer portfolio of financial institutions in Honduras. Through a comprehensive analysis of historical data, the main predictive variables were highlighted. Subsequently, a highly accurate machine learning model was developed and applied, using these variables to anticipate the default risk. This model has demonstrated exceptional capability by successfully predicting approximately 60% of clients prone to default. In addition to improving credit risk management, the implementation of this approach promises to reduce costs, optimize resources, and strengthen decision-making in the Honduran financial sector.

Keywords: Fiduciary consumer portfolio, Machine learning, Predictive models, Delinquency, Credit risk.

DEDICATORIA

A mi madre, un faro constante de amor y sabiduría; a mis hermanas, compañeras inquebrantables en este viaje; a los licenciados que generosamente compartieron su experiencia y conocimiento; y a mis compañeros de maestría, cuya colaboración y amistad han enriquecido mi camino académico Y a José Manuel García, mi compañero de tesis, cuya dedicación y esfuerzo compartidos han sido fundamentales para culminar este proyecto. Dedico esta tesis con sincero agradecimiento y cariño, reconociendo que cada uno de ustedes ha sido un pilar fundamental en mi búsqueda de conocimiento y éxito, dejando una huella imborrable en cada logro alcanzado.

Walther Nahun Torres

Dedicado a Dios, fuente infinita de sabiduría y guía, por iluminar mi camino en esta travesía académica y brindarme la fuerza para perseverar.

A mi amada familia, cuyo apoyo inquebrantable ha sido el pilar de mi éxito. A mi madre, por su amor incondicional y sacrificio. A mis hermanos, por su ánimo constante y comprensión.

Este logro es un tributo a su amor y apoyo inquebrantables. Gracias por estar a mi lado en este viaje.

José Manuel García

AGRADECIMIENTO

Quiero expresar mi profundo agradecimiento a todas las personas que contribuyeron de manera significativa a la realización de este trabajo de tesis. Sus esfuerzos y apoyo han sido fundamentales en este viaje académico.

En primer lugar, agradezco a Dios por brindarme la fortaleza y la perseverancia para completar este proyecto y por ser mi guía constante en la vida.

A mi familia, les dedico un agradecimiento especial. A mi madre, por su amor incondicional, valores y apoyo constante a lo largo de mi educación. A mis hermanos, por su ánimo y comprensión.

Agradezco sinceramente a mis profesores y asesores por su orientación, conocimientos y valiosos consejos expertos que fueron cruciales para la realización de este trabajo. Su dedicación a la enseñanza y la investigación ha sido una fuente constante de inspiración. Además, quiero expresar mi reconocimiento especial a mi compañero de tesis, Walther Torres, por su colaboración y su compromiso con este proyecto compartido.

José Manuel Garcia

En primer lugar, agradezco a Dios por su guía constante en mi trayecto académico, a mi amada familia por su inquebrantable apoyo y amor incondicional que han sido mi fuente de fortaleza, a la universidad por brindarme la oportunidad de crecer intelectualmente y por los recursos indispensables para llevar a cabo este proyecto, a los distinguidos licenciados que compartieron su valiosa experiencia y conocimientos conmigo, y a José Manuel, mi compañero de tesis, por su colaboración incansable, su dedicación y amistad inquebrantable, juntos superamos obstáculos y alcanzamos este logro significativo. Este trabajo es un testimonio del impacto positivo que cada uno de ustedes ha tenido en mi vida y en mi formación académica.

Walther Nahun Torres

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

La solidez de la cartera crediticia y la gestión efectiva del riesgo de impago son aspectos fundamentales para asegurar la estabilidad y rentabilidad de las instituciones financieras. La incertidumbre asociada a si un cliente caerá o no en mora representa un desafío significativo y más aún cuando son préstamos que no cuentan con una garantía tangible que respalde el crédito concedido al cliente, lo que incrementa el riesgo de incumplimiento de parte de ellos.

En esta investigación se plantea desarrollar un modelo de machine learning, diseñado para predecir la probabilidad de que un cliente, que en la actualidad mantiene su deuda al día, pueda caer en mora en el futuro. Para ello se emplearán datos históricos almacenados en las bases de dato de la institución, que incluyen variables financieras, demográficas y transacciones, con el objetivo de identificar a los clientes con mayor probabilidad de impago.

El propósito de este modelo predictivo es proporcionar una herramienta para fortalecer la gestión del riesgo crediticio, que permita a la institución enfocar de manera eficiente sus recursos e implementar estrategias preventivas dirigidas aquellos clientes con mayor probabilidad de impagos, facilitando la toma de decisiones y la adaptación de políticas de créditos que contribuyan a mantener una cartera sana y rentable.

1.2 ANTECEDENTES DEL PROBLEMA

La predicción de riesgo de impago en instituciones financieras es un desafío de suma importancia en el ámbito bancario, ya que permite evaluar la capacidad crediticia de los clientes y minimizar los riesgos asociados con los préstamos otorgados. En el caso específico de Honduras, un país con una economía en desarrollo y una industria financiera en constante evolución, la gestión adecuada del riesgo de crédito se convierte en una prioridad para asegurar la estabilidad del sector financiero y promover el acceso responsable al crédito.

Hasta hace poco, las instituciones financieras en Honduras se han basado principalmente en enfoques tradicionales, como análisis de crédito manual y modelos estadísticos lineales, para evaluar el riesgo de impago. Sin embargo, estos métodos pueden tener limitaciones para capturar patrones complejos y no lineales presentes en los datos crediticios, lo que puede llevar a una

evaluación inadecuada del riesgo y aumentar la exposición a pérdidas crediticias.

El uso de modelos de Machine Learning se presenta como una alternativa prometedora para mejorar la predicción de riesgo de impago en instituciones financieras hondureñas. Los modelos de Machine Learning, como Random Forest, Gradient Boosting y Redes Neuronales, entre otros, tienen la capacidad de aprender automáticamente a partir de grandes volúmenes de datos y capturar relaciones complejas entre variables, lo que los hace ideales para abordar el problema del riesgo crediticio.

Una investigación muy extensa titulada "Propuesta de un Modelo para Evaluar el Riesgo de Crédito utilizando Algoritmos de Predicción en la Cooperativa de Ahorro y Crédito LA" (Cuenca et al., 2019). En esta publicación, se afirma que la regresión logística demostró ser el mejor modelo debido a su excelente rendimiento y métricas en la predicción del riesgo. El modelo logró un accuracy de 0.6634, una sensibilidad de 0.6448, una especificidad de 0.6821 y una precisión de 0.6698. Además, se menciona que utilizaron herramientas como la curva de características operativas receptoras (ROC) y su área bajo la curva (AUC) para evaluar el modelo, lo cual les permitió descartar las modelaciones econométricas convencionales. Se reconoce que hubo posibles dificultades en el proceso de modelación, como la calidad de los datos, el filtrado de variables y la elección adecuada de algoritmos.

A pesar de los avances en los algoritmos de Machine Learning, se requiere una investigación específica para adaptar y aplicar estos modelos de predicción al contexto hondureño y evaluar su efectividad en la predicción de riesgo de impago en las instituciones financieras del país. Es importante considerar las particularidades del mercado financiero hondureño, así como la disponibilidad y calidad de los datos crediticios, para lograr resultados precisos y relevantes en esta área.

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 ENUNCIADO

Las entidades financieras por su modelo de negocio se encuentran expuestas a la posibilidad de que los clientes no cumplan con sus compromisos de pago, lo que representa un desafío para las instituciones financieras, por tal razón están constantemente desarrollando nuevas formas para medir y evaluar la posibilidad de incumplimiento de los clientes que solicitan créditos.

La gestión efectiva del riesgo crediticio es pilar fundamental para garantizar la estabilidad

y rentabilidad de las instituciones financieras. La cartera de préstamos de consumo fiduciaria se destaca por presentar un mayor riesgo, ya que no está respaldada por ninguna garantía tangible.

El análisis de los datos históricos de mora de la cartera de consumo de la institución bancaria adquiere una importancia para comprender la evolución del riesgo asociado a este producto, a diciembre del 2019, la tasa de morosidad de la cartera de consumo se situaba en un 7.7%, a diciembre del 2021 se situaba en un 14.6% y a diciembre 2022 la tasa se situaba en un 16.8%, reflejando un incremento en los casos de impago. Este aumento de la mora enfatiza la necesidad de plantear el riesgo crediticio de la cartera de consumo como un tema de suma importancia.

Haciendo uso de los datos históricos guardados en las bases de datos de la institución, se propone desarrollar un modelo de machine learning para predecir la probabilidad de que un cliente, que tiene sus deudas al día, pase a ser moroso en el futuro. Se utilizarán los datos históricos de clientes, incluyendo variables financieras, demográficas y transacciones para entrenar los modelos con el objetivo de identificar a los clientes con mayor riesgo de impago.

El contar con un modelo de predicción preciso será una herramienta muy valiosa para la gestión efectiva del riesgo crediticio, la aplicación de estrategias preventivas permitirá a la institución financiera enfocar sus recursos de manera más eficiente en los clientes con mayor probabilidad de caer en mora.

Además, durante esta investigación, se llevará a cabo un análisis exhaustivo de las variables utilizadas, mediante métodos gráficos, estadísticos y utilizando algoritmos de ML. Este análisis permitirá identificar de manera más precisa aquellas variables que tienen una influencia significativa en el comportamiento de los clientes en relación con sus pagos y deudas. Al destacar estas variables clave, la institución financiera estará mejor preparada para diseñar estrategias más específicas y efectivas, personalizadas para cada segmento de clientes, lo que contribuirá a una gestión del riesgo crediticio más precisa y a la toma de decisiones más informadas para mantener la salud financiera de la institución y la satisfacción de los clientes.

1.3.2 FORMULACIÓN DEL PROBLEMA

¿Cómo se pueden identificar y analizar las variables más relevantes que influyen en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario, mediante el desarrollo y aplicación de un modelo de machine learning, con el propósito de mejorar la gestión

del riesgo crediticio y reducir la morosidad en el banco?

1.3.3 PREGUNTAS DE INVESTIGACIÓN

¿Cuáles son las variables más relevantes para predecir si un cliente cae o no a mora?

¿Cómo crear un modelo de machine learning efectivo para predecir la probabilidad de impago en la cartera de consumo fiduciario, con el objetivo de reducir la morosidad y mejorar la gestión del riesgo crediticio en el Banco?

¿Cuál es el rendimiento de los diferentes modelos de machine learning evaluados, considerando las métricas más relevantes?

¿Cómo diseñar y desarrollar un dashboard dinámico para la visualización y supervisión de clientes con alta probabilidad de impago en la cartera de consumo fiduciario?

1.4 OBJETIVOS DEL PROYECTO

1.4.1 OBJETIVO GENERAL

Identificar y analizar las variables más relevantes que influyen en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario, con el propósito de mejorar la gestión del riesgo crediticio y reducir la morosidad del Banco.

1.4.2 OBJETIVOS ESPECÍFICOS

- Evaluar y seleccionar las variables más influyentes en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario mediante el uso de técnicas de análisis de datos y modelos de machine learning.
- Desarrollar y aplicar un modelo de machine learning precios y confiable que sea capaz de predecir la probabilidad de impago de los clientes de la cartera de consumo fiduciario, con el propósito de mejorar la gestión del riesgo crediticia y reducir la morosidad del Banco.
- Comparar y evaluar el rendimiento de diferentes modelos de ML, utilizando métricas como la precisión, el área bajo la curva ROC y la sensibilidad, para determinar el modelo que mejor desempeño tiene.

- Desarrollar un dashboard diseñado para visualizar y supervisar de manera dinámica a los clientes con alta probabilidad de caer en mora en la cartera de consumo fiduciaria.

1.5 JUSTIFICACIÓN

La implementación de un modelo de predicción de riesgo de impago en una institución financiera en Honduras surge como una respuesta esencial a una preocupación que afecta a todas las instituciones financieras, tanto en Honduras como en todo el mundo. La morosidad de los clientes representa una amenaza significativa para la estabilidad y sostenibilidad de estas entidades, y sus consecuencias pueden tener un impacto negativo en la economía nacional. En este contexto, la capacidad de anticipar y prevenir el incumplimiento de pagos se convierte en un objetivo crítico para salvaguardar la salud financiera de estas instituciones y, por ende, contribuir a la estabilidad económica del país.

El propósito de esta investigación es desarrollar y aplicar modelos de Machine Learning que permitan predecir la probabilidad de que un cliente, que actualmente mantiene sus deudas al día, se convierta en moroso en el futuro. La utilización de datos históricos de clientes, que incluyen variables financieras, demográficas y transacciones, permitirá construir modelos predictivos confiables y precisos.

Con la aplicación de estos modelos de predicción en las instituciones financieras se podrá identificar a clientes propensos a ser morosos, lo que ayudará a tomar decisiones informadas al otorgar créditos, establecer límites adecuados y aplicar medidas preventivas, también mejorará la evaluación de clientes potenciales, evitando otorgar créditos a personas con alto riesgo de incumplimiento, lo que optimizará la cartera de préstamos y reducirá pérdidas financieras. Todo esto generará un impacto en la reducción de riesgos financieros, la mejora de la toma de decisiones crediticias, el impulso al desarrollo económico y tecnológico del país, así como por la relevancia del conocimiento generado a nivel local. La aplicación de estos modelos predictivos contribuirá a una gestión crediticia más eficiente y responsable, beneficiando tanto a las instituciones financieras como a la economía hondureña en general.

CAPÍTULO II. MARCO TEÓRICO

2.1 ANÁLISIS DE LA SITUACIÓN ACTUAL.

2.1.1 SITUACIÓN ECONÓMICA Y FINANCIERA NIVEL GLOBAL

En el contexto actual de la economía global, la situación económica y financiera desempeña un papel crucial en la gestión del riesgo crediticio en instituciones financieras, tanto en Honduras como a nivel internacional. Elementos como la estabilidad económica, las tasas de interés, el crecimiento del producto interno bruto (PIB) y la dinámica del mercado laboral tienen un impacto significativo en la capacidad de los prestatarios para cumplir con sus compromisos crediticios. Dada la naturaleza cambiante de la economía, es esencial que las instituciones financieras sean capaces de anticipar y mitigar los riesgos de impago, particularmente en carteras de préstamos de consumo fiduciario, donde las variaciones en los ingresos y las condiciones económicas pueden afectar la capacidad de pago de los prestatarios.

En un informe elaborado por el sector de Centroamérica y llevado a cabo por la Sociedad Calificadora de Riesgo Centroamericana S.A. (SCRiesgo) , la cual es una entidad especializada en la evaluación del riesgo que posee sistemas de calificación para analizar emisiones de compañías no financieras, entidades financieras, fondos de inversión, acciones, titularizaciones, aseguradoras, administradoras de pensiones y estructuras financieras, se señala que hasta el final del año 2022, el sistema bancario de Honduras está compuesto por un total de 15 bancos, de los cuales 9 tienen alcance regional. Asimismo, durante el año pasado, Honduras experimentó una tasa de inflación anual del 9.8% (lo que representa una desaceleración del 0.64% en comparación con el mes previo). Gracias a las medidas implementadas para mitigar los impactos de la inflación y a la recuperación en la demanda de bienes y servicios, se logró un aumento en la concesión de préstamos(Riesgo, s. f.).

Dentro de este contexto, nuestra investigación se enfoca en desarrollar y aplicar modelos de machine learning para predecir el riesgo de impago en una institución financiera en Honduras. A través de la integración de indicadores macroeconómicos y técnicas avanzadas de análisis de datos, nuestro objetivo es proporcionar a la institución herramientas eficientes y actualizadas que les permitan anticipar y gestionar el riesgo crediticio de manera efectiva. Al hacerlo, contribuiré no solo a la optimización de la asignación de recursos, sino también al fortalecimiento de la estabilidad financiera en un entorno económico dinámico y desafiante.

2.1.2 RIESGOS FINANCIEROS EN INSTITUCIONES DE HONDURAS

La situación financiera en las instituciones de Honduras ha sido un tema recurrente en el ámbito bancario, dada la importancia de mantener la solidez de estas entidades. En un país marcado por su economía volátil y la exposición a factores externos, los riesgos financieros emergen como una preocupación fundamental. Entre estos riesgos, el riesgo crediticio se destaca por la morosidad y la falta de pago por parte de los clientes, lo cual puede afectar negativamente la salud financiera y la capacidad de préstamo de los bancos. Ante la carencia de un enfoque preventivo, el riesgo de asignación ineficiente de recursos y el aumento de la morosidad se convierten en una realidad que requiere atención.

La imperante necesidad de una gestión eficaz del riesgo es un hecho ampliamente reconocido, que está adquiriendo una creciente relevancia con el paso del tiempo. La expansión global y la integración de los mercados, en conjunción con la emergencia de productos financieros derivados, han brindado a las instituciones financieras oportunidades de negocio para las cuales no estaban preparadas previamente. Durante un largo período, el enfoque de la administración bancaria se ha centrado en la concepción y la implementación de sistemas de medición para diversas categorías de riesgo. Con los avances tecnológicos, estos sistemas han sido refinados

gradualmente, y en la actualidad, la problemática principal radica en el control y supervisión de las posiciones asumidas.

La actividad crediticia constituye la principal fuente de ingresos para una entidad bancaria, pero al mismo tiempo puede ser la causa determinante de su posible insolvencia. Esta aparente contradicción se origina en que un préstamo puede generar tanto ganancias para el estado financiero como también plantear desafíos significativos en términos de liquidez en caso de incumplimiento en los pagos de intereses y/o del capital.

Considerando el marco de riesgos financieros en las instituciones de Honduras y la necesidad de fortalecer la gestión del riesgo crediticio, la presente investigación cobra una relevancia inminente. A través de la implementación de metodologías avanzadas en análisis de datos y machine learning, se pretende brindar a las entidades financieras una herramienta eficaz para prever y mitigar la morosidad, optimizando la asignación de recursos y favoreciendo la toma de decisiones informada. La incorporación de enfoques innovadores basados en modelos predictivos no solo contribuirá al avance de las prácticas financieras en Honduras, sino que podría tener implicaciones más amplias para la estabilidad económica en Centroamérica. De esta manera, se abre una nueva vía de desarrollo y mejora de la confiabilidad en el sistema financiero regional, fundamentada en el análisis riguroso y la anticipación proactiva de riesgos.

2.1.3 PROBLEMÁTICA DEL IMPAGO EN INSTITUCIONES FINANCIERAS

En el ámbito actual de la industria financiera, la importancia de gestionar el riesgo crediticio de manera eficaz es innegable. Predecir y mitigar la morosidad se ha convertido en un desafío crucial para la sostenibilidad y rentabilidad de las instituciones financieras, especialmente en países como Honduras. El paso de clientes con un historial de cumplimiento a una situación de impago puede tener un impacto directo en la estabilidad financiera de estas entidades. En este

contexto, surge la cuestión fundamental de cómo desarrollar y aplicar modelos de machine learning capaces de anticipar con precisión la probabilidad de que un cliente, que hasta ahora ha cumplido con sus obligaciones, pueda incumplir en el futuro. El objetivo esencial de esta investigación es mejorar la gestión del riesgo crediticio en la cartera de consumo fiduciario del banco, permitiendo una distribución eficiente de recursos y contribuyendo a la reducción de la morosidad.

En la presente investigación se pretende abordar una cuestión crítica para las instituciones financieras en Honduras, ofreciendo soluciones basadas en la innovación tecnológica y el análisis avanzado de datos. Al desarrollar y aplicar modelos de machine learning que puedan predecir con precisión el riesgo de impago de los clientes en el futuro cercano, se brinda a estas instituciones la oportunidad de prevenir y mitigar los efectos adversos de la morosidad. Al concentrarse en la cartera de consumo fiduciario del banco, esta investigación pretende no solo mejorar la asignación de recursos, sino también respaldar la eficiencia operativa y la estabilidad económica de la región. Más allá del contexto hondureño, este estudio podría sentar las bases para investigaciones futuras y aplicaciones en la gestión del riesgo crediticio en otros países centroamericanos, contribuyendo al avance de las prácticas financieras en toda la región.

2.2 CONCEPTUALIZACIÓN

La dinámica de la economía global junto con el crecimiento de la competencia, hacen que la flexibilidad y agilidad a la hora de conceder créditos sean un factor determinante en muchas instituciones financieras. No obstante, esta flexibilidad trae consigo una serie de riesgos inherentes en pos de mantener la competitividad. La falta de rigurosidad en la asignación de préstamos es uno de los aspectos que puede provocar situaciones de incumplimiento financiero y poner en peligro la estabilidad de la institución.

Es muy importante encontrar un equilibrio entre ser flexible y ser exhaustivo a la hora de evaluar la capacidad crediticia de los prestatarios. De esta manera, se garantiza la continuidad

operativa de la institución y la protección de los intereses de las partes involucradas.

2.2.1 DEFINICIÓN Y CONCEPTO DE RIESGO CREDITICIO Y SU RELEVANCIA EN EL CONTEXTO FINANCIERO

2.2.1.1 GESTIÓN DEL RIESGO FINANCIERO

A medida que la economía evoluciona, la concesión de créditos emerge como parte central para impulsar el crecimiento económico y sostenible de las instituciones financieras. La concesión de créditos ofrece a los individuos y empresas acceso a recursos financieros inmediatos, dotándolos de capacidad de inversión y ampliación de su consumo. Este proceso incide directamente en los prestatarios e impulsa el crecimiento económico al incitar oportunidades de inversión y estimular el consumo. La concesión de crédito requiere la adopción de procesos rigurosos de evaluación crediticia y la implementación de estrategias de gestión del riesgo.

2.2.1.2 CRÉDITO

"El crédito es un préstamo en dinero, donde la persona se compromete a devolver la cantidad solicitada en el tiempo o plazo definido según las condiciones establecidas por dichos préstamos, más los intereses devengados, seguros y costos asociados si los hubiese"(Morales Castro & Morales Castro, 2015, p. 23).

Los préstamos personas son una forma flexible de créditos que brindan a los individuos acceso a fondos para cubrir necesidades diversas, como gastos imprevistos, educación, viajes o mejoras en el hogar. (Morales Castro & Morales Castro) menciona que en estos créditos la firma es la garantía de pago y que para fijar la tasa de interés depende de cada prestatario, porque cada uno posee un riesgo especifico.

2.2.1.3 RIESGO

El riesgo constituye un elemento esencial en todas las actividades económicas. (Valle Carrascal, 2017) menciona "si definimos al crédito como la esperanza de obtener una devolución de un préstamo en un tiempo estipulado, el riesgo de crédito es que esa esperanza no se materialice".

(Ledesma Martinez & Sanchez Machado, 2007)M define el riesgo “como la oportunidad o probabilidad del surgimiento de algún evento desfavorable. Está ligado a la incertidumbre que rodea en general a cualquier hecho económico, en el sentido de contingencias que puedan ocasionar pérdidas”.

La gestión efectiva del riesgo es un pilar fundamental para garantizar la solidez de las instituciones financieras. Las instituciones en su rol de intermediarios se enfrentan a una diversidad de riesgos entre ellos el riesgo crediticio, de liquidez, de mercado, legal y riesgo operativo. El riesgo crediticio adquiere un mayor protagonismo debido a su relación implícita con los préstamos.

El riesgo crediticio ocupa un lugar de atención relevante en las instituciones, su importancia creciente se atribuye a los siguientes factores:

- Mayor deterioro de las carteras crediticias impulsado por un incremento entre prestamistas.
- La transformación del sistema financiero hacia una mayor desintermediación impulsa una mayor base de solicitantes de crédito, a la vez conlleva una reducción de la calidad crediticia.
- El riesgo crediticio impacta directamente la rentabilidad financiera. Una cartera deteriorada disminuye los ingresos y afecta la estabilidad financiera.
- Mayor competencia entre instituciones financiera conlleva disminución en los márgenes de ganancias.

Hay que recalcar un aspecto muy importante como es la adecuación del capital requerido para afrontar los riesgos crediticios, esto es una recomendación propuesta por el comité de Basilea, el objetivo de esta propuesta es establecer directrices que garanticen la estabilidad del sistema financiero global. Esta propuesta se basa en asegurar una gestión efectiva del riesgo y se busca

establecer niveles mínimos de capital que sean proporcionales a los riesgos asumidos.

2.2.1.4 PROBABILIDAD DE INCUMPLIMIENTO

La probabilidad de incumplimiento emerge como un factor de vital importancia en la gestión del riesgo crediticio, esta métrica no solo evalúa la posibilidad de que un cliente incumpla con su obligación crediticia, sino que también mide el grado de confianza que la institución financiera tiene sobre la capacitad de pago del cliente.

(Gómez Fernández-Aguado & Partal Ureña, 2010, p. 28) define que "La probabilidad de que un cliente no cumpla con sus compromisos financieros dentro de un periodo de tiempo específico se vincula con su deseo de paga. Esto también se ve influenciado por la situación financiera del deudor y por factores que podrían impactar su intención de pago en el futuro".

La esencia del mercado de crédito se basa en la confianza, las instituciones al conceder un préstamo confían en el compromiso de pago de parte del cliente. Sin embargo, no todos los clientes tienen el mismo perfil, es aquí donde la probabilidad de incumplimiento adquiere relevancia.

La estimación de la probabilidad de incumplimiento es fundamental para tomar decisiones informadas en el ámbito financiero. Su influencia es fundamental no solo en la aprobación de créditos o la estimación de la probabilidad de incumpliendo, también para tomar las siguientes decisiones estratégicas:

- Asignación de tasas de interés basadas en riesgo.
- Segmentación de clientes de acuerdo con su probabilidad de incumplimiento.
- Diseño de estrategias de cobranza que se adapten al perfil del deudor.

Ahora que entendemos la importancia y la influencia de la probabilidad de incumplimiento en la gestión del riesgo, vamos a profundizar en cómo las instituciones financieras cuantifican el

riesgo crediticio.

2.2.1.5 MEDICIÓN DEL RIESGO CREDITICIO

La medición del riesgo crediticio es crucial en el ámbito financiero debido a la natural incertidumbre asociada a los contratos de crédito. Este proceso aborda la volatilidad del mercado y se adapta a factores tales como el aumento de la exposición crediticia, la creciente competitividad y la desintermediación del sector. También hay que considerar las regulaciones impuestas por entes supervisores, los cuales demandan practicas más robustas y precisas para la gestión del riesgo. Frente a esto se hace imprescindible técnicas avanzadas e innovadoras para cuantificar y gestionar el riesgo crediticio. El sistema para evaluar el riesgo crediticio se centra en determinar los factores clave que influyen en el riesgo crediticio de los portafolios de cada entidad. El propósito es anticipar y evitar posibles pérdidas en las que se podría incurrir. Para este análisis se consideran criterios de calificación de las carteras de crédito, variables demográficas, macroeconómicas y el comportamiento histórico de la cartera. (García & García, 2010, p 299)

Los modelos de riesgo de crédito permiten identificar, medir y gestionar el riesgo de manera consistente y rigurosa. El interés inicial por la utilización de éstos surgió ante la necesidad de realizar estimaciones cuantitativas del capital económico.(Gómez Fernández-Aguado & Partal Ureña, 2010, p. 19)

Los modelos de medición del riesgo crediticio han evolucionado y mejorando constantemente, en esta investigación haremos referencia a los métodos tradiciones hasta los más modernos en donde se involucran técnicas estadísticas e inteligencia artificial.

2.2.1.5.1 METODOS TRADICIONALES

Las entidades financieras tradicionalmente basaban la medición del riesgo crediticio con métodos fundamentados en el juicio y experiencia de los analistas de crédito. Estos métodos

estudiaban al cliente y las interpretaciones de ratios financieros y entrevistas personales.

- Juicio o sistema expertos:

Son herramientas tecnológicas diseñadas para capturar el razonamiento y juicio de expertos humanos. Los sistemas expertos tratan de capturan la intuición de los expertos y sistematizarla aprovechando la tecnología, pues su campo de dominio es la inteligencia artificial, por medio de la cual intentan crear sistemas expertos y redes neuronales. (García & García, 2010, p 300). Los sistemas expertos se fundamentan en los siguientes conceptos:

Capacidad: Capacidad de pago del deudor.

Capital: análisis de la situación financiera del deudor, capacidad de endeudamiento, liquidez.

Colateral: Activos en posesión del deudor como pago de garantía.

Carácter: Información de hábitos de pago del deudor.

Condiciones: Factores externos que afectan los hábitos de pago del deudor.

El sistema contempla características simples del prestatario para conformar las bases del modelo posterior, los cuales a través de estos datos se realizan predicciones sobre la posibilidad de default. (Grau Álvarez, 2020, p 30)

- Sistemas de calificación:

"Los sistemas de calificación de crédito más antiguo lo desarrollo la Oficina de Control de Moneda de Estados Unidos para la evaluación de su deuda emitida y adecuación de las reservas para posibles casos de default"(García & García, 2010, p 303). Los sistemas de calificación son una herramienta valiosa, que proporciona una estructura metodológica para medir la probabilidad de incumplimiento de los deudores. El sistema establece 5 rangos para el portafolio de créditos:

Rangos de la Oficina de Control de Moneda

Reservas	%
Calificaciones de baja calidad	
Otros activos especialmente mencionados	0
Activos por debajo del estándar	20
Activos dudosos	50
Cartera vencida	100
Calificaciones de alta calidad	
Vencidos y vigentes	0

Figura 1: Rangos de la oficina de control de moneda

Fuente: (García & García, 2010, p 303)

Los sistemas de calificación se han ido adaptando e incorporando desde análisis cualitativo hasta evaluaciones cualitativas del comportamiento del prestatario.

Principales sistemas de calificación a considerar:

Scoring: Estos sistemas permiten decidir a quién se le concede un crédito, cuanto crédito conceder y estrategias para obtener un mayor beneficio de la operación. Estos sistemas funcionan mediante un algoritmo que otorgan una puntuación de la calidad crediticia de una solicitud, condicionando la decisión de conceder un crédito si o no.

Los sistemas de scoring clasifican la cartera minorista (consumo, hipotecario, tarjeta de crédito, pyme (Gómez Fernández-Aguado & Partal Ureña, 2010, p 40)

Por lo general, existen dos tipos de modelos de scoring: scoring de aprobación o de evaluación de solicitudes para créditos nuevos y scoring de gestión o de comportamiento, que realiza seguimiento a los clientes ya incorporados dentro de la EIF y permite generar y automatizar algunos procesos tales como: estimación de previsiones, acciones de cobranza, detección preventiva de fraude,

análisis de mercadeo para medir niveles de consumo, lealtad y deserción, proveer ratings que diferencien clientes. (Saunders & Allen, 2010)

Rating: Gómez, et al. (2010) menciona que los sistemas calificación fueron evolucionando, alcanzando fiabilidad en los resultados, se desarrollaron sistemas de clasificación más sofisticados, que atendían a otros segmentos del negocio (pymes, empresas, corporaciones, administración pública etc), es así como surge los sistemas internos de rating, que buscan categorizar a los clientes en grupos homogéneos de riesgo, esta metodología se centra en asignar una probabilidad especifica de impago, volviéndose más técnica, precisa y objetiva la evaluación de riesgo.

A medida que el mercado financiero evoluciona y se vuelven más complejo los sistemas de calificación se han ido adaptando para ofrecer análisis más precisos, su papel en la medición y gestión del riesgo sigue siendo fundamental para garantizar la estabilidad del sector financiero.

2.2.1.5.2 METODOS MODERNOS

Estos modelos son más innovadores y sofisticados e incluyen un mayor número de variables, buscan ofrecer una perspectiva más integral y rigurosa de los diferentes factores de incertidumbre.

Modelo KMV: El modelo KMV se basa en el modelo original de Robert Merton, es una técnica de simulación con información de los mercados para estimar la probabilidad de incumplimiento.

El procedimiento según García & García (2010) básico del modelo es:

- Estimar el valor y la volatilidad de los activos de la empresa
- Calcular el riesgo de los activos, en el que se incluye el riesgo del negocio y del sector en que trabaja la empresa.

- Estimar la probabilidad de incumplimiento (p 308-310)

2.3 TEORÍAS DE SUSTENTO

2.3.1 BASES TEÓRICAS

2.3.1.1 MACHINE LEARNING Y SU RELEVANCIA EN EL RIESGO CREDITICIO

El Machine Learning es una rama de la inteligencia artificial que emergió en la década de los 80, otorgando a las máquinas la capacidad de adaptar su comportamiento mediante la identificación de patrones y tendencias. Esta capacidad no solo enriquece su proceso analítico, sino que facilita la comprensión y procesamiento de datos, habilitando además la toma de decisiones automatizadas. En el contexto de una era digital saturada de datos, se presenta el desafío para empresas y organizaciones de extraer información significativa. Bobadilla (2020) resalta que, a diferencia de los métodos convencionales donde las soluciones se establecen paso a paso, el Machine Learning se define como una ciencia computacional que permite a las máquinas aprender directamente de los datos. Estos algoritmos, versátiles en naturaleza, son capaces de identificar patrones en diversos contextos (págs. 10). Es así como el Machine Learning se erige como una de las tecnologías más disruptivas e innovadoras de la última década, ofreciendo soluciones precisas, predictivas y adaptadas a las necesidades de un mundo en constante evolución digital.

El machine learning (ML) es una subdisciplina esencial de la inteligencia artificial. La IA se centra en la simulación de la inteligencia humana en máquinas, el ML se especializa en el uso de algoritmos y modelos estadísticos para realizar una tarea mediante la inferencia en lugar de las instrucciones.

Principales diferencias entre la inteligencia artificial y el machine learning

Tabla 1: Diferencias entre la inteligencia artificial y el machine learning

Inteligencia Artificial	Machine Learning
Aplicaciones que imitan la inteligencia humana, no todas las soluciones de IA son ML	Es una metodología de la Inteligencia Artificial. Todas las soluciones de ML son soluciones de IA.
IA es mejor para completar una tarea humana compleja con eficiencia	ML es mejor para identificar patrones en grandes conjuntos de datos.
Puede operar con o sin dato, y basarse en reglas predefinidas.	Esencialmente depende de los datos para su funcionamiento
Robótica, procesamiento de lenguaje natural, sistemas expertos entre otros.	Clasificación de imágenes, detención de anomalías, sistemas de recomendación.

Fuentes: Elaboración propia

En la presente investigación nos enfocaremos en como los modelos de machine learning pueden ser utilizados para predecir la probabilidad de incumplimiento de pago de los clientes de una institución financiera.

2.3.1.2 TIPOS DE MACHINE LEARNING

Dentro del ML existen diferentes enfoques según como se adquiera el aprendizaje, como se define la estructura y metodología que se empleara para procesar los datos. En el siguiente apartado profundizaremos en los principales tipos de ML, supervisado, no supervisado y por refuerzo.

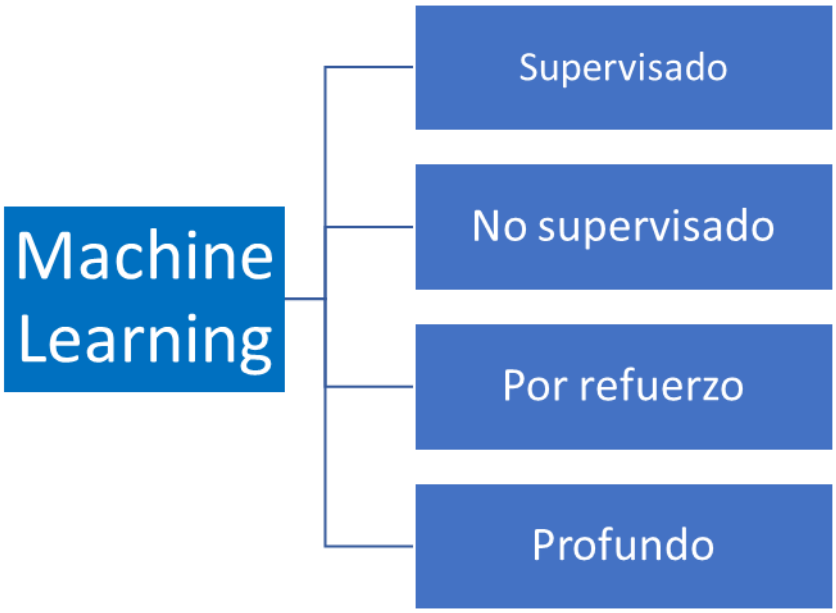


Figura 2: Clasificación de los principales tipos de aprendizajes machine learning

Fuentes: Elaboración propia

2.3.1.2.1 APRENDIZAJE SUPERVISADO

El aprendizaje supervisado es un enfoque esencial en el campo del aprendizaje automático. En este método de entrenamiento, se utiliza un conjunto de datos etiquetado para enseñar a un modelo de machine learning a realizar predicciones precisas. Cada ejemplo en el conjunto de datos de entrenamiento tiene una etiqueta que representa la respuesta deseada del modelo.

El aprendizaje estadístico supervisado es un conjunto de técnicas para deducir una función a partir de datos de entrenamiento y es una de las herramientas principales de la minería de datos y del aprendizaje automático. (Campo León & Alcalá Nalvaiz, 2016)

Este enfoque se aplica comúnmente en tareas de clasificación y regresión. La clasificación implica asignar categorías a los datos de entrada, mientras que la regresión se enfoca en predecir valores numéricos. Diversos algoritmos, como regresión lineal, árboles de decisión y redes

neuronales, se emplean para entrenar modelos supervisados.

Una vez entrenado, se evalúa el rendimiento del modelo utilizando un conjunto de datos de prueba independiente. Las métricas de rendimiento se utilizan para medir cuán precisas son las predicciones del modelo en comparación con las etiquetas reales.

2.3.1.2.2 APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado en machine learning es una rama en la que los algoritmos se entrenan para analizar datos sin utilizar etiquetas o información de salida predefinida. A diferencia del aprendizaje supervisado, donde se proporcionan ejemplos etiquetados para entrenar un modelo en función de la relación entre entradas y salidas, el aprendizaje no supervisado se centra en la exploración y descubrimiento de patrones y estructuras inherentes en los datos por sí mismos.

Una de las tareas principales del aprendizaje no supervisado es la agrupación (clustering), donde el algoritmo busca automáticamente dividir el conjunto de datos en grupos o clústeres de elementos similares. Esto es útil para segmentar datos en categorías no conocidas previamente, como la agrupación de clientes en segmentos de mercado con preferencias similares. Además, el aprendizaje no supervisado también aborda la reducción de dimensionalidad, que implica la simplificación de la representación de datos al eliminar características redundantes o ruidosas, lo que facilita la visualización y el procesamiento de datos de alta dimensionalidad.

2.3.1.2.3 APRENDIZAJE POR REFUERZO

El Aprendizaje por Refuerzo es un enfoque del aprendizaje automático en el que un agente, como un programa de computadora o un robot, toma decisiones en un entorno con el objetivo de maximizar una recompensa a lo largo del tiempo. Este proceso implica que el agente aprenda una política, es decir, una estrategia para tomar decisiones, a través de la exploración y la observación de las recompensas que recibe del entorno.

Los componentes clave del Aprendizaje por Refuerzo incluyen el agente, que toma decisiones; el entorno, en el que opera el agente; las acciones que el agente puede realizar; las recompensas que recibe después de cada acción; y la política que guía las decisiones del agente.

El objetivo principal del agente es aprender una política óptima que maximice las recompensas a lo largo del tiempo, lo que implica tomar decisiones inteligentes basadas en la experiencia acumulada. Algunos algoritmos populares en este campo son Q-Learning, SARSA, DDPG, PPO y A3C, y se utiliza en una amplia gama de aplicaciones, como juegos, robótica y recomendación de contenido personalizado.

2.3.1.2.4 APRENDIZAJE PROFUNDO

El Aprendizaje Profundo también es conocido como Deep Learning en inglés, se erige como un subcampo esencial dentro de machine learning, con su enfoque en la capacitación de redes neuronales artificiales para abordar tareas específicas. Su singularidad radica en su habilidad para extraer de forma automática características y representaciones complejas de datos a partir de extensos y complejos conjuntos de datos. En su núcleo, el Aprendizaje Profundo hace uso de redes neuronales artificiales, compuestas por capas interconectadas de unidades denominadas neuronas, que efectúan operaciones matemáticas en los datos de entrada y utilizan funciones de activación para transmitir información de una capa a la siguiente.

Una característica distintiva de las redes neuronales profundas es su arquitectura compuesta por múltiples capas ocultas, situadas entre la capa de entrada y la capa de salida. Esta estructura posibilita el aprendizaje de representaciones jerárquicas y abstractas de los datos, lo que resulta especialmente relevante en aplicaciones que involucran procesamiento de imágenes, voz, texto y otros datos complejos. Además, el Aprendizaje Profundo se caracteriza por su capacidad para realizar el aprendizaje automático de características, lo que significa que el modelo puede

identificar y utilizar las características más relevantes de los datos sin requerir una extracción manual.

El Aprendizaje Profundo ha tenido un impacto amplio y significativo en diversas áreas, incluyendo el reconocimiento de imágenes, el procesamiento del lenguaje natural, la traducción automática, la generación de texto y música, la conducción autónoma y la detección de fraudes, entre otras. Sin embargo, es importante destacar que su eficacia se ve potenciada por la disponibilidad de conjuntos de datos extensos y de alta calidad, así como por recursos computacionales considerables, como unidades de procesamiento gráfico (GPU) y unidades de procesamiento tensorial (TPU). En última instancia, el Aprendizaje Profundo continúa siendo un área de investigación activa y promete seguir desempeñando un papel esencial en el avance de la inteligencia artificial y sus aplicaciones prácticas.

2.3.2 TECNICAS DE APRENDIZAJES

2.3.2.1 MODELOS DE CLASIFICACIÓN

Los modelos de clasificación en ML intentan predecir con exactitud la categoría o etiqueta especifica a una entrada de datos dada basándose en datos de entrenamientos previos. Estos modelos no predicen un valor continuo, sino que asigna una etiqueta discreta (categórica) a partir de un conjunto predefinido. Por ejemplo, en un problema del ámbito financiero, un modelo de clasificación podría predecir si un cliente es de alto riesgo, medio riesgo o bajo riesgo.

2.3.2.1.1 REGRESIÓN LOGÍSTICA

La regresión logística está basada en la ecuación de una recta y = mx+b, este es un método estadístico usado para resolver problemas de clasificación binaria. Este modelo se ocupa en las situaciones donde la variable dependiente solo puede tener dos tipos posibles: 0,1. (Hosmer & Lemeshow, 2000, citado en Giraldo et al., 2021)

La regresión logística es una técnica que se basa en estimar la probabilidad de que ocurra un evento, como un cliente que cae en default, basándose en diversas variables asociadas a su comportamiento crediticio.

2.3.2.1.2 K-NEAREST NEIGHBORS (KNN)

Este algoritmo busca las k muestras más cercanas a la muestra que queremos clasificar y devuelve una etiqueta que clasifica a la muestra.

2.3.2.1.3 ARBOLES DE DECISIÓN

Son modelos ampliamente usados para tareas de clasificación y regresión, sus resultados pueden entenderse y explicarse fácilmente. Cada nodo interno denota una prueba en un atributo, cada rama representa el resultado de cada prueba y los nodos finales representan el resultado de la decisión.

Ventajas:

- ✓ Fácil interpretación
- ✓ Se requiere poca preparación de los datos

Desventajas:

- ✓ Propenso a sobreajuste
- ✓ Son inestables

El coeficiente o índice Gini es una métrica que se utiliza para medir la impureza de un conjunto de datos. Cuanto más alto es el valor de "Gini" más difícil será la clasificación de una muestra. El valor Gini es usado por el algoritmo para expandir el árbol eligiendo caminos que lleven a los nodos más puros. (Bobadilla, 2020, p 128)

2.3.2.1.4 RANDOM FOREST

Este es un algoritmo basado en los árboles de decisión. Los métodos ensemble combinan varios algoritmos de machine learning para mejorar la exactitud de los resultados. El algoritmo usa varios árboles de decisión, y cada uno de estos árboles es alimentado con un subconjunto aleatorio de los datos de entrenamiento. (Bobadilla, 2020, p 128). Los Random Forest tienen entre sus beneficios su flexibilidad y fácil de usar para resolver problemas, tanto en clasificación como en regresión. Una de sus desventajas es que requiere mucho poder de cómputo y puede hacer que el algoritmo se vuelva lento.

2.3.2.1.5 XGBoost

Son un algoritmo muy eficiente en la reducción del sesgo y de la varianza de un modelo, estos algoritmos implican crear muchos arboles de decisión a partir de datos de entrenamiento los cuales son creados de forma secuencial (James, 2017 citado en Soules et al., 2020).

2.3.2.1.6 Naive Bayes

Naive Bayes es un algoritmo de machine learning ampliamente utilizado en la clasificación de datos, especialmente en aplicaciones relacionadas con texto. Se basa en el teorema de Bayes y se caracteriza por su simplicidad.

Este algoritmo supone de manera simplificada que todas las características utilizadas para la clasificación son independientes, lo que es raramente cierto en situaciones reales. A pesar de esta simplificación, Naive Bayes es eficiente desde el punto de vista computacional y fácil de implementar, lo que lo hace popular en problemas de clasificación.

El proceso de clasificación con Naive Bayes implica recopilar datos de entrenamiento etiquetados, calcular las probabilidades de que una instancia de entrada pertenezca a cada clase utilizando el teorema de Bayes y, finalmente, asignar la clase con la probabilidad más alta como la predicción.

Este algoritmo destaca en tareas como la detección de spam, análisis de sentimientos o categorización de texto. Sin embargo, su suposición de independencia entre características puede limitar su eficacia en conjuntos de datos donde las relaciones entre las características son significativas. En tales casos, pueden ser preferibles otros algoritmos más avanzados.

2.3.2.2 PRINCIPALES PROBLEMAS EN LOS MODELOS DE CLASIFICACION

- ✓ Calidad de los datos: Los principales problemas que se pueden tener son los datos faltantes, incorrectos, duplicados, mal normalizados.
- ✓ Dimensionalidad: Los modelos trabajan bien con un conjunto de variables, si nos excedemos con un gran número de variables, podríamos no tener un rendimiento optimo en el modelo.
- ✓ Sobreajuste (Overfitting): El sobreajuste está relacionado con modelos que se ajustan a los datos de entrenamiento, estos modelos no son capaces de generalizar bien y por lo tanto no son capaces de predecir correctamente (Pineda Pertuz, 2022, p 44)
- ✓ Subajuste (Underfitting): Esto sucede cuando el modelo no se ha entrenado con los suficientes datos y no es capaz de representar adecuadamente la relación entre las variables descriptiva y las variables destino (Pineda Pertuz, 2022, p 44).

2.3.2.3 METRICAS PARA LA EVALUACION DEL DESEMPEÑO DE UN MODELO ML

- Matriz de confusión: Es una tabla de frecuencia para encontrar la precisión y corrección de algoritmo de aprendizaje automático.
    - ✓ VP: Verdadero Positivos, es el número correcto de predicciones de la instancia positiva.
    - ✓ FP: Falsos Positivos, es el número incorrecto de predicciones que la instancia es

positiva

- ✓ FN: Falsos Negativos, es el número incorrecto de predicciones que la instancia negativa.
- ✓ VN: Verdaderos Negativos, es el número correcto de predicciones de la instancia negativa.

- Accuracy (Exactitud): Es la proporción de predicciones correctas del total de predicciones realizadas.
- Precisión: Es la proporción de verdaderos positivos sobre el total de positivos (VP+FP).
- Recall: Es la proporción de verdaderos positivos sobre el total de casos positivos reales (VP+FN).
- Especificidad: Es la proporción de verdaderos negativos sobre el total de casos negativos reales (VN+FP).
- F1-Score: Combina la precisión y el recall en un solo indicador.
- Curva de ROC: Es una representación gráfica de como el modelo se comporta en todos los niveles de la clasificación.

2.3.2.4 APLICACIONES DE MACHINE LEARNING EN EL SECTOR FINANCIERO

Encontramos numerosas aplicaciones en el sector financiero, el cual ha estado en constante evolución en la última década, especialmente impulsado por la introducción de nuevas tecnologías. La oportunidad de procesar grandes cantidades de datos ha abierto las puertas al machine learning que surge como una herramienta que ofrece soluciones precisas en el análisis de datos financieros.

Detección de Fraude: Los métodos no supervisados son utilizados para analizar grandes

cantidades de datos. Los bancos utilizan modelos de ML para detectar cuando los gastos de un cliente no concuerdan con los gastos normales de los clientes.

Predecir riesgo de créditos: Los métodos supervisados son de gran utilidad para predecir el riesgo de incumplimiento futuro. El desarrollo de un modelo conlleva 3 fases, la recopilación de la información, la construcción del dataset, la aplicación del modelo y documentación.

Análisis de documentos: El análisis y reconocimiento de documentos tiene con objetivo la extracción automática de la información. Esta herramienta tiene la posibilidad de escanear los documentos y obtener la información más relevante convirtiéndose en uno de los principales beneficios de ML.

2.3.3 METODOLOGÍAS APLICADAS

2.3.3.1 RECOPILACIÓN Y PREPROCESAMIENTO DE DATOS

La evaluación del riesgo de crédito y la asignación de categorías de crédito a los consumidores pueden abordarse como una tarea de clasificación. En esta tarea, los individuos son etiquetados según su estado crediticio, que puede ser favorable o desfavorable, según los estándares establecidos por el prestamista y su historial de pagos. Esta evaluación puede llevarse a cabo a través de un análisis detallado del rendimiento y situación financiera de cada cliente, o de manera masiva utilizando métodos cuantitativos.

En la actualidad, los enfoques de aprendizaje automático están ampliamente disponibles y su implementación resulta conveniente en términos de accesibilidad, velocidad y confiabilidad. Estos enfoques pueden considerarse como alternativas a la regresión logística, que ha sido tradicionalmente la técnica predominante en la evaluación masiva de créditos.

Para que los modelos de Machine Learning sean exitosos, es importante seguir una serie de pasos que garanticen su calidad. Estos pasos incluyen comprender los datos, prepararlos para

el análisis y crear un modelo que se ajuste bien a los datos.

2.3.3.1.1 RECOPILACIÓN DE DATOS

La recopilación de datos se trata de la obtención de información relevante sobre los clientes, que pueda ser utilizada para identificar los factores que están asociados con el incumplimiento de pago, en este proceso se involucra la captura sistemática y organizada de información relevante relacionada con la actividad financiera de los clientes, comportamiento crediticio y variables contextualmente influyentes. La calidad y diversidad de los datos recopilados impactan directamente en la robustez y precisión de los modelos de machine learning empleados en la predicción de riesgo. La recopilación implica la identificación y selección de fuentes de datos confiables, así como la preparación, limpieza y transformación de los datos en formatos adecuados para el análisis posterior. Este proceso riguroso garantiza que los modelos de predicción puedan aprender patrones significativos y generar pronósticos certeros que respalden la toma de decisiones financieras informadas y estratégicas.

La recopilación de datos abarcará una variedad de dimensiones financieras y crediticias, como historiales de pagos, ingresos, deudas y detalles transaccionales, junto con factores externos como tasas de interés y condiciones económicas. La exhaustiva recopilación de estos datos permitirá construir conjuntos de entrenamiento y prueba que representen fielmente la complejidad del entorno financiero. Además, la metodología de recopilación deberá cumplir con los estándares de privacidad y seguridad de datos, garantizando la confidencialidad de la información sensible de los clientes. En última instancia, la recopilación de datos como parte integral de la metodología respalda la generación de modelos de predicción sólidos y fiables, capaces de anticipar con precisión los riesgos de impago y contribuir de manera significativa a la gestión proactiva y eficaz del riesgo crediticio en instituciones financieras.

2.3.3.1.2 PREPROCESAMIENTO DE DATOS

La fase de preprocesamiento de datos desempeña un papel crítico en el análisis de información y en la construcción de modelos de aprendizaje automático. Consiste en emplear un conjunto de técnicas y procedimientos dirigidos a los datos en su estado original y sin manipular, con el propósito de adaptarlos y remodelarlos en un formato más adecuado y funcional para su ulterior examen y modelado. Este proceso es esencial debido a que los datos en su forma primaria tienden a ser ruidosos, incompletos y en muchas ocasiones no idóneos para ser introducidos directamente en los modelos de machine learning.

El preprocesamiento de datos puede implicar las siguientes acciones:

- Limpieza de datos: La limpieza de datos es el proceso de identificar y corregir errores, inconsistencias y anomalías en los conjuntos de datos. Esto implica eliminar datos duplicados, valores atípicos (outliers) y registros erróneos que podrían afectar la calidad y confiabilidad de los resultados del análisis.

- Manejo de valores faltantes: Los valores faltantes son entradas de datos que no están presentes en ciertas observaciones. Estos pueden surgir debido a errores de entrada, problemas técnicos o simplemente falta de información. El manejo de valores faltantes implica tomar decisiones sobre cómo tratar esos valores, ya sea imputando valores estimados, eliminando registros con valores faltantes o utilizando técnicas más avanzadas para preservar la integridad de los datos.

- Codificación de variables categóricas: Las variables categóricas son aquellas que representan categorías o etiquetas en lugar de valores numéricos. Para utilizar estas variables en modelos de machine learning, es necesario codificarlas en un formato numérico. Esto puede hacerse mediante técnicas como la codificación one-hot (binaria), la

codificación ordinal o la codificación basada en la frecuencia.

- Normalización/estandarización: La normalización y estandarización son procesos para ajustar las escalas de las variables numéricas en un rango específico. La normalización generalmente escala los datos en un rango entre 0 y 1, mientras que la estandarización los transforma para que tengan una media de 0 y una desviación estándar de 1. Estos pasos son importantes para garantizar que las variables con diferentes escalas no afecten negativamente el rendimiento de los modelos de machine learning.

- Creación de características relevantes: La creación de características relevantes implica generar nuevas variables a partir de las características existentes o incorporar información externa que pueda mejorar la capacidad predictiva de los modelos. Esto podría incluir la combinación de variables, la extracción de características a partir de datos no procesados o la incorporación de conocimientos expertos.

- Balanceo de clases: En el contexto de la predicción de riesgo de impago, es probable que las clases estén desequilibradas, es decir, una clase (por ejemplo, "no impago") puede ser significativamente más grande que la otra ("impago"). El balanceo de clases implica técnicas para igualar la representación de las clases en el conjunto de datos, como el submuestreo de la clase mayoritaria, el sobre muestreo de la clase minoritaria o el uso de técnicas como SMOTE (Synthetic Minority Over-sampling Technique).

- División de datos: La división de datos es el proceso de separar el conjunto de datos en subconjuntos para entrenar, validar y probar los modelos de machine learning. Esto se hace típicamente en tres conjuntos: entrenamiento, validación y prueba. El conjunto de entrenamiento se utiliza para entrenar el modelo, el conjunto de validación se utiliza para

ajustar hiperparámetros y tomar decisiones de modelado, y el conjunto de prueba se utiliza para evaluar el rendimiento general del modelo en datos no vistos.

CAPÍTULO III. METODOLOGÍA

3.1 CONGRUENCIA METODOLÓGICA

Es fundamental establecer una estructura metodológica sólida para abordar la predicción de impagos en la cartera fiduciaria mediante técnicas de machine learning. En el siguiente apartado, desplegaremos una matriz metodológica que proporcionara claridad y cohesión a nuestro tema de investigación.

3.1.1 MATRIZ METODOLÓGICA

Tabla 2: Matriz de congruencia metodológica

Formulación del problema	Objetivo General	Objetivos Específicos	Preguntas de Investigación	Variables Dependientes	Variables Independientes
¿Cómo se pueden identificar y analizar las variables más relevantes que influyen en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario, mediante el desarrollo y aplicación de un modelo de machine learning, con el propósito de mejorar la gestión del riesgo crediticio y reducir la morosidad en el banco?	Identificar y analizar las variables más relevantes que influyen en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario, con el propósito de mejorar la gestión del riesgo crediticio y reducir la morosidad del Banco.	Evaluar y seleccionar las variables más influyentes en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario mediante el uso de técnicas de análisis de datos y modelos de machine learning.	¿Cuáles son las variables más relevantes para predecir si un cliente cae o no a mora?	Impago	Edad, agencia, zona, tipo de persona, tiene tarjeta débito, tiene banca en línea, tiene APP, seguros, tiene celular, tiene correo, genero, departamento, nivel educativo, profesión, cargo, nivel de ingreso, ingresos, frecuencia de ingresos, años de trabajo, nombre empresa, estado laboral, origen recursos, estado civil, dependientes, cantidad préstamos, saldos préstamos, tipos de garantía, tiene refinanciamiento, tasa, cuotas pagadas,
		Desarrollar y aplicar un modelo de machine learning precios y confiable que sea capaz de predecir la probabilidad de impago de los clientes de la cartera de consumo fiduciario, con el propósito de mejorar la gestión del riesgo crediticia y reducir la morosidad del Banco.	¿Cómo crear un modelo de machine learning efectivo para predecir la probabilidad de impago en la cartera de consumo fiduciario, con el objetivo de reducir la morosidad y mejorar la gestión del riesgo crediticio en el Banco?		
		Comparar y	¿Cuál es la precisión de un		

		evaluar el rendimiento de diferentes modelos de ML, utilizando métricas como la precisión, el área bajo la curva ROC y la sensibilidad, para determinar el modelo que mejor desempeño tiene.	sistema de score de crédito desarrollado con ML al predecir el riesgo de impago de los clientes de la cartera de crédito de consumo fiduciario?		plazo, forma de pago, mora al cambio, días en mora, saldo de mora, Conteo cuentas ahorro, Conteo Cuentas cheque+CDT, Saldos Cuentas ahorro, Saldos Cuentas cheque+CDT, Conteo veces en mora en semestre I, Conteo veces en mora en semestre II, Cantidad Promedio Pagos mensuales, Conteo pagos promedios mensuales.
		Desarrollar un dashboard diseñado para visualizar y supervisar de manera dinámica a los clientes con alta probabilidad de caer en mora en la cartera de consumo fiduciaria.	¿Cómo diseñar y desarrollar un dashboard dinámico para la visualización y supervisión de clientes con alta probabilidad de impago en la cartera de consumo fiduciario?		

Fuente: Elaboración propia

3.1.2 ESQUEMA DE VARIABLES DE ESTUDIO

Dentro de nuestra investigación sobre la gestión crediticia y el riesgo de impago, es muy importante entender las variables que incluyen en este comportamiento. Estas variables no solo sirven para analizar la situación actual, sino que son fundamentales para modelar predicciones futuras. En esta sección presentamos un diagrama sagital categorizado para visualizar como estas variables se relacionan con la variable dependiente el “Impago”.

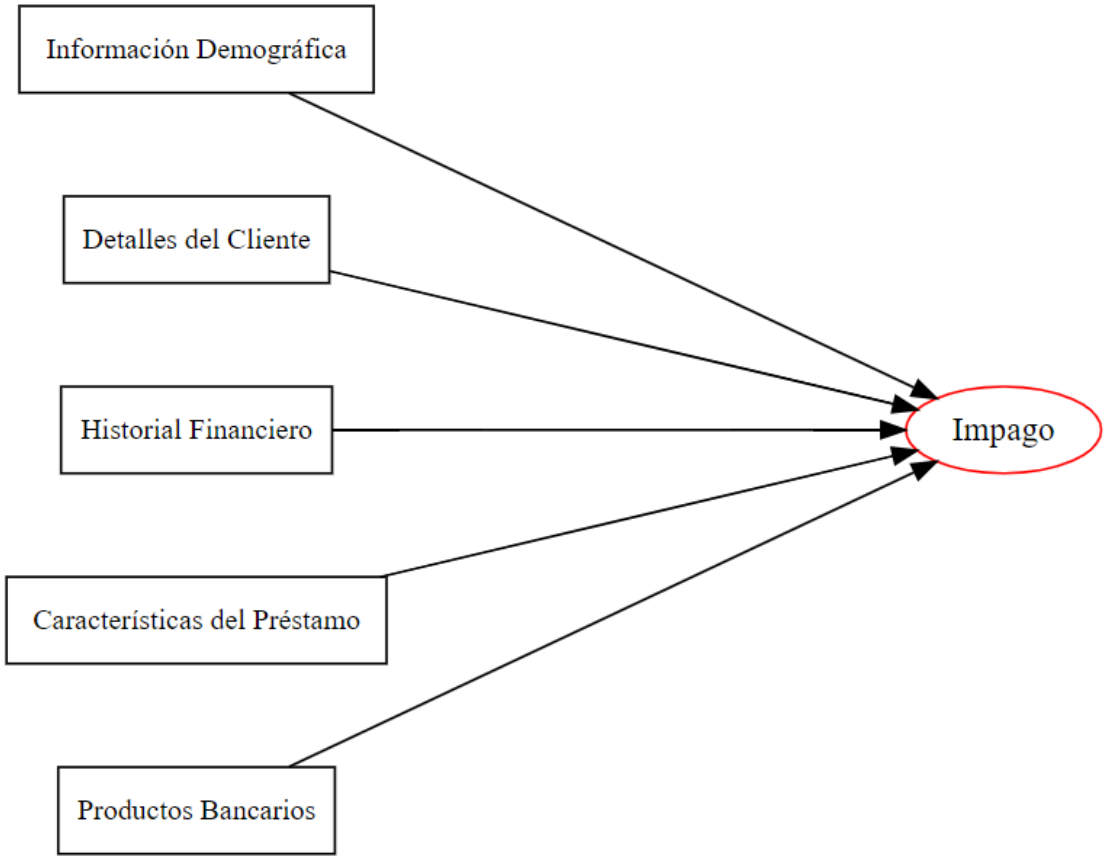


Figura 3: Esquema de variables de estudio

Fuentes: Elaboración propia

3.1.3 OPERACIONALIZACIÓN DE LAS VARIABLES

Tabla 3: Caracterización de las variables

Variable	Descripción	Operacional	Dimensiones	Indicador
Edad	Años de vida del cliente.	Ayuda a determinar la etapa de vida del cliente.	Intervalo numérico (años)	Edad promedio de los clientes. Rango de edades más común.
Género	Sexo del cliente.	Diferencia entre clientes masculinos y femeninos.	Masculino, Femenino	Porcentaje de clientes masculinos y femeninos.
Estado Civil	Situación conyugal del cliente.	Informa sobre la situación familiar del cliente.	Soltero, Casado, Divorciado, etc.	Distribución porcentual de los estados civiles.

Departamento	Región o área geográfica de residencia del cliente.	Ayuda a identificar la ubicación del cliente.	[Lista de Departamentos]	Departamento con mayor número de clientes.
Tipo de persona	Categorización del cliente según su naturaleza.	Diferencia entre personas naturales y jurídicas.	Natural, Jurídica	Porcentaje de clientes naturales vs jurídicas.
Profesión	Ocupación o actividad laboral principal del cliente.	Permite identificar la diversidad de ocupaciones de los clientes.	[Lista de Profesiones]	Top 5 profesiones más comunes entre los clientes.
Cargo	Posición laboral del cliente en una organización.	Indica la jerarquía o nivel de responsabilidad del cliente.	[Lista de Cargos]	Distribución porcentual de cargos entre clientes.
Nivel Educativo	Grado de estudios alcanzado por el cliente.	Ofrece insights sobre la educación de los clientes.	Primaria, Secundaria, Técnico, Universitario, Postgrado, etc.	Porcentaje de clientes por nivel educativo.
Nombre Empresa	Organización donde el cliente trabaja o es dueño.	Brinda contexto sobre el ámbito laboral del cliente.	[Lista de Empresas]	Empresas con mayor cantidad de clientes asociados.
Estado Laboral	Situación laboral actual del cliente.	Ayuda a identificar la estabilidad laboral del cliente.	Empleado, Autónomo, Desempleado, Retirado, etc.	Distribución porcentual de la situación laboral de los clientes.
Origen de recursos	Fuente principal de los ingresos del cliente.	Indica la procedencia de los fondos del cliente.	Salario, Negocio propio, Inversiones, Herencia, Otros.	Principales orígenes de recursos entre los clientes.
Cantidad de préstamos	Número total de préstamos que un cliente tiene.	Refleja la actividad crediticia del cliente.	Número total de préstamos	Promedio de préstamos por cliente.
Saldos préstamos	Monto total pendiente en préstamos del cliente.	Indica la cantidad de deuda que el cliente aún tiene pendiente.	Monto en moneda local	Promedio de saldo pendiente por cliente.
Conteo veces en mora en semestre I	Número de veces que un cliente estuvo en mora durante el primer semestre.	Refleja la conducta de pago del cliente durante el primer semestre.	Número total de veces	Porcentaje de clientes con al menos 1 mora en el semestre I.
Conteo veces en mora en semestre II	Número de veces que un cliente estuvo en mora durante el segundo semestre.	Refleja la conducta de pago del cliente durante el segundo semestre.	Número total de veces	Porcentaje de clientes con al menos 1 mora en el semestre II.
Cantidad Promedio Pagos mensuales	Promedio de pagos que un cliente realiza al mes.	Indica el comportamiento de pago mensual del cliente.	Cantidad en moneda local	Cantidad promedio de pagos mensuales por cliente.

Conteo pagos promedios mensuales	Número promedio de veces que un cliente realiza pagos en un mes.	Ofrece una visión sobre la frecuencia con la que el cliente realiza pagos mensuales.	Número total de pagos mensuales	Promedio de frecuencia de pagos mensuales por cliente.
Tipos de garantía	Clasificación de las garantías que respaldan un préstamo.	Indica el nivel de seguridad de recuperación del préstamo.	Hipotecaria, Prendaria, Fiduciaria, etc.	Distribución porcentual de tipos de garantía en los préstamos.
Tiene refinanciamiento	Si un cliente ha refinanciado algún préstamo.	Refleja la capacidad del cliente de pagar sus obligaciones a tiempo.	Sí, No	Porcentaje de clientes que han refinanciado.
Tasa	Interés aplicado al préstamo.	Indica el costo de financiamiento para el cliente.	Porcentaje anual	Tasa de interés promedio de los préstamos.
Cuotas pagadas	Número de cuotas que el cliente ha pagado hasta la fecha.	Indica el comportamiento de pago del cliente.	Número de cuotas	Promedio de cuotas pagadas por cliente.
Plazo	Duración total del préstamo en meses o años.	Indica el compromiso de tiempo del cliente con el préstamo.	Meses, Años	Duración promedio de los préstamos.
Forma de pago	Método por el cual el cliente paga sus cuotas.	Refleja la preferencia o facilidad del cliente para realizar pagos.	Débito automático, Cheque, Transferencia, etc.	Distribución porcentual de las formas de pago utilizadas.
Mora al cambio	Si el cliente ha tenido moras al cambiar de forma de pago o refinanciar.	Indica la adaptabilidad del cliente ante cambios en las condiciones del préstamo.	Sí, No	Porcentaje de clientes con mora al cambio.
Días en mora	Número de días que un préstamo ha estado en mora.	Refleja el grado de incumplimiento del cliente.	Número de días	Promedio de días en mora por cliente.
Saldo de mora	Monto total que el cliente tiene pendiente por mora.	Indica el monto que el cliente debe por retrasos en los pagos.	Monto en moneda local	Monto promedio de saldo en mora por cliente.
Tiene tarjeta débito	Si el cliente posee una tarjeta débito del banco.	Indica la vinculación del cliente con el banco.	Sí, No	Porcentaje de clientes con tarjeta débito.
Tiene banca en línea	Si el cliente utiliza la banca en línea del banco.	Refleja la adaptabilidad digital del cliente.	Sí, No	Porcentaje de clientes que usan banca en línea.
Tiene APP	Si el cliente ha descargado y utiliza la app del banco.	Indica el nivel de interacción digital del cliente.	Sí, No	Porcentaje de clientes que usan la APP del banco.
Seguros	Si el cliente tiene seguros contratados con el banco.	Refleja el nivel de servicios adicionales del cliente.	Sí, No	Porcentaje de clientes con seguros contratados.

Tiene celular	Si el banco tiene registrado un número de celular del cliente.	Indica posibilidad de contacto directo.	Sí, No	Porcentaje de clientes con número de celular registrado.
Tiene correo	Si el banco tiene registrado un correo electrónico del cliente.	Facilita la comunicación y envío de información.	Sí, No	Porcentaje de clientes con correo electrónico registrado.
Conteo cuentas ahorro	Número de cuentas de ahorro que tiene un cliente.	Indica el nivel de ahorro del cliente.	Número de cuentas	Promedio de cuentas de ahorro por cliente.
Conteo Cuentas cheque+CDT	Número de cuentas corriente y CDTs de un cliente.	Refleja la diversidad de servicios bancarios del cliente.	Número de cuentas	Promedio de cuentas corriente y CDTs por cliente.
Saldos Cuentas ahorro	Monto total en cuentas de ahorro de un cliente.	Indica el volumen de ahorro del cliente.	Monto en moneda local	Saldo promedio en cuentas de ahorro por cliente.
Saldos Cuentas cheque+CDT	Monto total en cuentas corriente y CDTs de un cliente.	Refleja el volumen de servicios bancarios del cliente.	Monto en moneda local	Saldo promedio en cuentas corriente y CDTs por cliente.

Fuentes: Elaboración propia

3.1.4 HIPÓTESIS

Hi: Existe una relación significativa entre las variables analizadas y la probabilidad de incumplimiento de los clientes en la cartera de consumo fiduciario de la institución financiera.

Ho: No existe una relación significativa entre las variables analizadas para evaluar el riesgo de impago y la probabilidad de incumplimiento de los clientes en la cartera de consumo fiduciario de la institución financiera.

HO: La inclusión de variables socioeconómicas y demográficas no mejoran la capacidad predictiva de los modelos de machine learning para el riesgo de impago en la cartera de consumo fiduciario.

H1: La inclusión de variables socioeconómicas y demográficas mejoran la capacidad predictiva de los modelos de machine learning para el riesgo de impago en la cartera de consumo fiduciario.

HO: No existe una relación significativa entre las variables financieras y el riesgo de incumplimiento en la cartera de consumo fiduciario de la institución financiera.

H1: Existe una relación significativa entre al menos una variable financiera y el riesgo de

incumplimiento en la cartera de consumo fiduciario de la institución financiera.

3.2 ENFOQUE Y MÉTODOS

El enfoque de esta investigación es cuantitativo, se centra en la aplicación rigurosa de métodos numéricos y estadísticos para analizar datos financieros y desarrollar modelos de machine learning. Se busca cuantificar y medir de manera precisa la probabilidad de impago en una institución financiera utilizando técnicas de análisis cuantitativo de datos para identificar los factores que influyen en el riesgo de impago de los clientes. El enfoque cuantitativo permitirá la utilización de datos históricos, ratios financieras, y otras variables relevantes que se analizarán utilizando modelos de machine learning para desarrollar un modelo predictivo que pueda estimar la probabilidad de que un cliente incumpla con sus obligaciones crediticias.

3.2.1 ALCANCE

El alcance de esta investigación es correlacional, se enfoca en analizar las relaciones y las interdependencias entre diversas variables financieras y económicas que puedan influir en el riesgo de impago en una institución financiera. Se explorará cómo estas variables están relacionadas entre sí y cómo sus correlaciones pueden afectar la capacidad de los modelos de machine learning para predecir el riesgo de impago.

El alcance correlacional implica la identificación de las variables clave que pueden estar correlacionadas con el riesgo de impago, como tasas de interés, nivel de endeudamiento, índices económicos, entre otros. Se realizarán análisis de correlación para determinar la fuerza y la dirección de estas relaciones, lo que ayudará a comprender mejor los factores que contribuyen al riesgo de impago en una institución financiera.

3.3 DISEÑO DE LA INVESTIGACIÓN

El diseño de la investigación es no experimental, no longitudinal, de tipo descriptivo. Se trata de un estudio observacional en el que se recopilan datos longitudinales para predecir el riesgo de

impago en una institución financiera.

3.3.1 POBLACIÓN

Los datos utilizados en este estudio son proporcionados por la institución financiera en cuestión. Esto significa que se utilizarán registros históricos y actuales de la institución financiera, que incluyen información detallada sobre los clientes, sus transacciones financieras, historiales de préstamos, comportamiento de pago y otros datos relevantes, de la cual se ha realizado una selección rigurosa de clientes que cumplen con ciertos criterios específicos. Se han escogido aquellos clientes que poseen un préstamo de consumo fiduciario, es decir, préstamos que no cuentan con una garantía fiduciaria como respaldo. Este criterio se ha establecido para analizar el comportamiento de un segmento particular de la cartera de préstamos. Además, se ha limitado la selección a clientes que presentan menos de 15 días de mora en sus pagos al final del periodo de observación, lo que indica un nivel relativamente bajo de retraso en sus obligaciones financieras.

También se determinó que la condición impuesta es que los préstamos seleccionados deben tener una antigüedad mínima de 12 meses. Este período de tiempo se ha considerado necesario para comprender de manera adecuada la evolución de los préstamos de consumo fiduciario en la cartera de la institución financiera. La elección de esta muestra se basa en la observación de un aumento significativo en la mora de estos préstamos a lo largo del tiempo, pasando del 7% en 2019 al 16% en diciembre de 2022. Por lo tanto, se consideró fundamental analizar a fondo a este grupo de clientes durante un período específico.

El período de observación se llevó a cabo desde el 31 de mayo del 2021 hasta el 30 de abril del 2022, (figura 4) y durante este lapso se aplicaron todos los filtros mencionados anteriormente. Luego, se procedió a evaluar el desempeño de estos clientes en el período comprendido entre el 1 de mayo de 2022 y el 31 de julio de 2022, clasificándolos en clientes morosos y no morosos. La variable dependiente para esta clasificación se basó en aquellos clientes que tenían una mora mayor a 30 días, y a partir de esta categorización se determinó si un cliente pagó (menos de 30 días de mora) o no pagó (30 días o más de mora). Como resultado de este proceso, se obtuvo una muestra que representa aproximadamente el 20% de la población total de clientes del banco que tienen un crédito. Esta selección de datos permitirá realizar un análisis más detallado y específico sobre el comportamiento de este grupo en particular en relación con los préstamos de consumo fiduciario y su historial de morosidad.

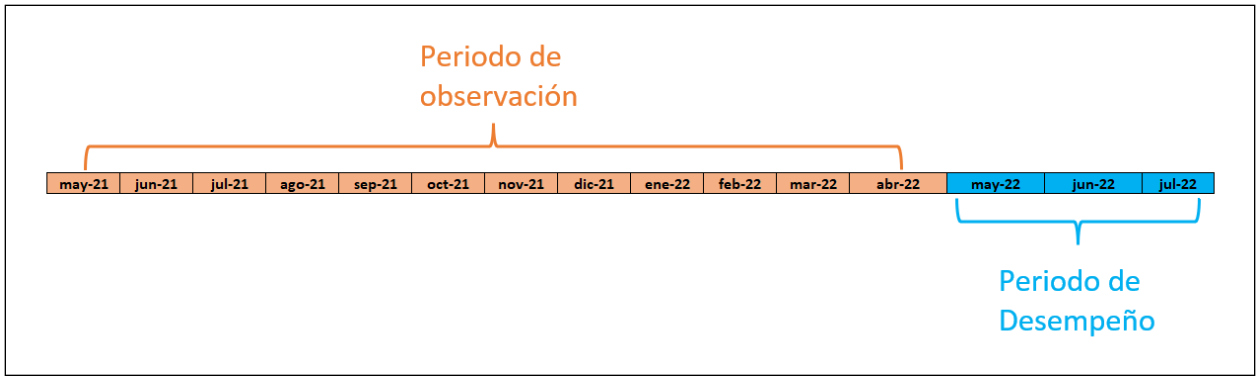


Figura 4: Periodo de Observación / Periodo de Desempeño

Fuentes: Elaboración propia

3.4 TÉCNICAS, INSTRUMENTOS Y PROCEDIMIENTOS APLICADOS

Figura 5: Esquema de Trabajo

Fuentes: Elaboración propia

3.4.1 TÉCNICAS

3.4.1.1 Recolección de datos:

Se recolectaron datos a partir de una base de datos proporcionada por la institución financiera, cuyo contenido es de suma relevancia para el estudio del riesgo de impago en préstamos de consumo sin garantía fiduciaria. La base de datos utilizada abarca tanto los datos demográficos como ser el género, edad, estado civil, ocupación e ingreso de los clientes; datos Financieros que abarcan los historiales crediticios y comportamientos de pagos y por último los datos transaccionales, como ser los saldos y movimientos de los clientes, brindando un panorama

completo y detallado de la cartera de clientes de la institución. Además, la base de datos proporciona acceso a los historiales de préstamos de los clientes, lo que permite un análisis profundo del comportamiento de pago a lo largo del tiempo. Este conjunto de datos se erige como una herramienta esencial para comprender y evaluar con precisión el historial crediticio de los clientes y, por ende, para el desarrollo y la validación de modelos de predicción de riesgo de impago en la institución financiera.

3.4.1.2 Reprocesamiento de datos:

Para llevar a cabo la aplicación de un modelo machine learning que permita predecir con éxito riesgo de impago en la institución financiera se requieren niveles óptimos de calidad de datos y volúmenes apropiados de observaciones para funcionar eficazmente. Para asegurar este requisito, es esencial llevar a cabo una adecuada etapa de preprocesamiento de datos que asegure la coherencia y la integridad de estos. En función de este principio, se detallan a continuación las metodologías empleadas en el proceso de preprocesamiento de datos.

3.4.1.3 Tratamiento de variables categóricas

El tratamiento de variables categóricas desempeña un rol fundamental tanto en el análisis de datos como en la creación de modelos estadísticos y de aprendizaje automático. Las variables categóricas son aquellas que representan grupos o categorías discretas en lugar de valores numéricos continuos. Ejemplos de estas variables incluyen el género (masculino o femenino), la ubicación geográfica (norte, sur, este u oeste), etc.

El tratamiento de las variables categóricas implica la conversión de estas categorías en una forma adecuada para su utilización en algoritmos de machine learning y los modelos estadísticos puedan trabajar de manera efectiva con ellas.

El concepto principal en el tratamiento de variables categóricas es la creación de variables ficticias o "variables dummy". Las variables ficticias son variables binarias que se utilizan para representar las diferentes categorías de una variable categórica. A cada categoría se le asigna una variable ficticia, que toma el valor 1 si la observación pertenece a esa categoría y 0 en caso contrario.

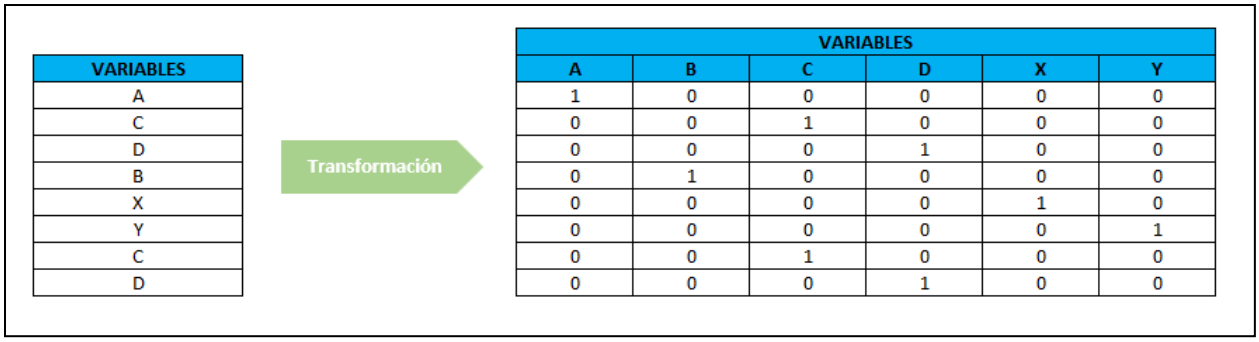


Figura 6: Ejemplo variables dummy

Fuentes: Elaboración propia

En la ilustración anterior, tenemos una serie de variables categóricas (A, B, C, D, X, Y), para tratar estas variables categóricas, debemos asignar el valor 1 si la observación pertenece a esa categoría y 0 en caso contrario.

Estas variables ficticias permiten incorporar la información de la variable categórica en modelos de regresión o clasificación. Algunos algoritmos de machine learning requieren que las variables categóricas se conviertan en numéricas para funcionar adecuadamente. Algunas de las ventajas de tratar convertir las categóricas es mejorar el rendimiento del modelo, evitar sesgos, preservar información útil, brindar flexibilidad en la elección del algoritmo y reducir la dimensionalidad de los datos.

3.4.1.4 Tratamiento de los valores atípicos:

Los valores atípicos, son datos en un conjunto de datos que se desvían significativamente de la mayoría de las otras observaciones, también conocidos como outliers en inglés. Estos valores son inusualmente altos o bajos en comparación con el patrón general o la tendencia de los datos y pueden ser resultado de errores de medición, eventos raros o inesperados, o simplemente representar casos excepcionales.

El tratamiento de los valores atípicos se refiere a las técnicas y estrategias utilizadas para identificar, analizar y, en algunos casos, corregir o gestionar observaciones que son significativamente diferentes de la mayoría de los datos en un conjunto de datos. Estos valores atípicos son puntos de datos que se desvían sustancialmente de la tendencia general o patrón de los datos y pueden tener un impacto negativo en el análisis estadístico o en la toma de decisiones si no se manejan adecuadamente.

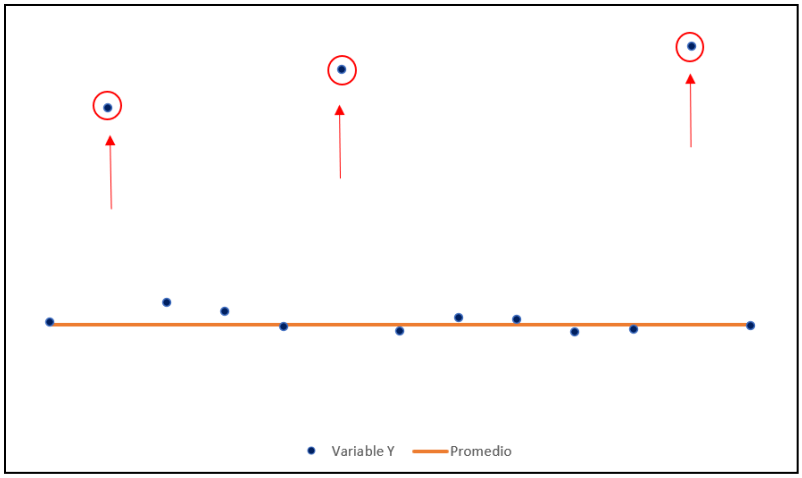


Figura 7: Ejemplo valores atípicos

Fuentes: Elaboración propia

Existen varios métodos para tratar los valores atípicos, entre los métodos más comunes están:

- Eliminar los valores atípicos. Este es el método más simple y directo. Sin embargo, es importante tener cuidado al eliminar datos, ya que se puede perder información importante.
- Corregir los valores atípicos. Este método consiste en reemplazar los valores atípicos con valores que sean más probables que sean verdaderos.
- Utilizar métodos estadísticos robustos. Estos métodos son menos sensibles a los valores atípicos que los métodos tradicionales.

Para efectos de esta investigación utilizamos la media para tratar los valores atípicos, la cual consiste en calcular la media del conjunto de datos determinando un límite superior y un límite inferior para identificar los valores atípicos, utilizando la media más o menos un múltiplo de la desviación estándar, una vez identificados, se eliminan los datos que se encuentran fuera de estos límites para tratar los valores atípicos.

3.4.1.5 Tratamiento de los datos faltantes:

El tratamiento de datos faltantes es también conocido como imputación de datos faltantes, constituye un procedimiento fundamental, su propósito radica en abordar la ausencia de valores en un conjunto de datos, fenómeno que puede derivar de distintas causas, como errores al introducir datos, fallos durante la recolección de información. El objetivo central de este proceso

es completar o estimar los valores que faltan de manera que se permita llevar a cabo análisis estadísticos o modelados de datos con mayor precisión y representatividad.

La ausencia de datos puede distorsionar los resultados, lo que implica que estos pueden ofrecer una representación imprecisa de la realidad, existen dos enfoques principales para abordar la gestión de datos faltantes:

1. Eliminación de datos: Esta estrategia implica la eliminación de los registros que presentan valores faltantes. Puede ser una solución efectiva cuando los datos faltantes son escasos. No obstante, esta acción puede reducir el tamaño de la muestra y, por ende, la capacidad de las pruebas estadísticas.
2. Imputación de datos: Este enfoque se basa en reemplazar los datos faltantes por valores estimados. Se conocen métodos de imputación los cuales se pueden categorizar en dos grupos fundamentales:

- Imputación por sustitución: Esta técnica consiste en sustituir los datos faltantes con el valor de una variable de referencia, como la media, la mediana o el valor más probable.
- Imputación por modelo: Este método emplea un modelo estadístico para calcular los valores faltantes de manera estimada.

En esta investigación se optó por utilizar la técnica de imputación por sustitución utilizando la media para compensar los datos faltantes que se presentaron en el conjunto de datos proporcionados por la entidad financiera.

3.4.1.6 Selección de variables:

La selección de variables es el proceso el cual consiste en elegir un subconjunto de las variables o características disponibles en los datos para usar en un modelo de machine learning. Esto se puede hacer para mejorar el rendimiento del modelo, reducir la complejidad del modelo y mejorar la interpretabilidad del modelo. Consiste en elegir un subconjunto de las variables o características disponibles en los datos para usar en el modelo, con el objetivo de mejorar la precisión del modelo, reducir la redundancia, eliminar el ruido de los datos y reducir la complejidad, lo que puede conducir a modelos más simples y precisos. En esta investigación utilizamos los modelos Random Forest y XGBoost utilizando el lenguaje de programación Python.

Random Forest: La selección de variables con Random Forest se realiza de forma implícita y

automática a través de dos mecanismos principales: muestreo aleatorio de características y evaluación de la importancia de las variables.

- Muestreo aleatorio de características: En cada uno de los árboles de decisión que conforman el Random Forest, se elige un conjunto de características (también llamadas variables predictoras) de manera aleatoria para su entrenamiento. Este procedimiento se denomina "muestreo aleatorio de características" o "bagging de características". Esto implica que cada árbol se entrena utilizando un conjunto único de variables, lo que aporta diversidad a los árboles individuales.
- Evaluación de la importancia de las variables: Después completar el proceso de entrenamiento del Random Forest, es posible realizar un análisis para determinar la relevancia de cada variable en relación con las predicciones resultantes. La evaluación de la importancia de las variables se basa en cuánto aportan al desempeño general del modelo. Esta información resulta útil para identificar cuáles variables son las más significativas en el conjunto de datos.

El modelo Random Forest tiene tres características principales:

- Reducción del sobreajuste: Evita que los árboles se especialicen demasiado en un conjunto específico de variables, mejorando así la capacidad del modelo para hacer predicciones precisas en nuevos datos y disminuyendo el riesgo de sobreajuste.
- Manejo de características irrelevantes: Da menor importancia a las variables que tienen un impacto limitado en las predicciones, lo que simplifica la identificación y eliminación de características que no son significativas para el modelo.
- Robustez ante datos faltantes: Puede tratar de manera efectiva con conjuntos de datos que presentan valores faltantes sin necesidad de realizar una imputación previa, lo que lo hace adecuado para abordar información incompleta.

XGBoost: La selección de variables en modelos XGBoost, consiste en un algoritmo de aumento de gradiente, el cual es utilizado en el proceso de elegir las características más relevantes o importantes para entrenar el modelo. La selección de variables con XGBoost funciona así:

- Entrenamiento inicial del modelo: Comienza entrenando un modelo XGBoost utilizando todas las características disponibles en los datos de entrenamiento.

- Evaluación de la importancia de las variables: Durante el proceso de entrenamiento, XGBoost calcula la importancia de cada característica basada en su contribución a la ganancia de información y su frecuencia de uso en la división de nodos en el árbol de decisión.
- Selección de características: Después de entrenar el modelo, se seleccion las características más relevantes utilizando la puntuación de importancia.
- Entrenamiento del modelo final: Se entrena un nuevo modelo XGBoost utilizando solo las características seleccionadas en el paso anterior. Este modelo reducido debería tener un mejor rendimiento en términos de precisión y velocidad en comparación con el modelo original, ya que se eliminan las características menos importantes.

El modelo XGBoost cuneta con las siguientes características principales:

- Robustez: XGBoost es resistente a la presencia de características irrelevantes o ruidosas, lo que lo hace apropiado para la selección de variables.
- Selección automática: XGBoost puede generar puntuaciones de importancia de variables y se integra fácilmente con algoritmos de selección de características para automatizar el proceso.
- Manejo de datos mixtos: XGBoost es versátil y puede manejar datos mixtos, incluyendo características numéricas y categóricas, lo que lo hace adecuado para diversos tipos de conjuntos de datos.

Para complementar la selección de variables utilizadas con los modelos Random Forest y XGBoost se utilizaron las técnicas estadísticas Pruebas Chi-cuadrado, Pruebas t de Student y Pruebas de correlación.

- Pruebas Chi-cuadrado: La prueba Chi-cuadrado se usa para evaluar la independencia entre dos variables categóricas y determinar si existe una relación significativa entre ellas. Se calcula comparando las frecuencias observadas con las esperadas bajo la hipótesis nula de independencia. Si el valor de Chi-cuadrado es significativo, se rechaza la hipótesis nula, lo que indica una relación entre las variables. Su principañ característica es su utilidad para variables categóricas, no para variables numéricas. Cuanto mayor sea la estadística de Chi-cuadrado, mayor será la evidencia de dependencia entre las variables.

- Pruebas t de Student: La prueba t de Student se utiliza para comparar las medias de dos grupos y determinar si existe una diferencia significativa entre ellos. Se calcula la estadística t al comparar las diferencias entre las medias observadas con las diferencias esperadas bajo la hipótesis nula de que no hay diferencia significativa. Si el valor t resulta significativo, se rechaza la hipótesis nula, lo que sugiere la presencia de una diferencia entre las medias. Una de sus características principales es utilizada para comparar variables numéricas entre dos grupos. Se puede realizar tanto en grupos emparejados como no emparejados.
- Pruebas de correlación: Las pruebas de correlación evalúan la relación lineal entre dos variables numéricas y determinan si existe una asociación estadística significativa entre ellas. Se calcula un coeficiente de correlación, como el coeficiente de correlación de Pearson o el coeficiente de correlación de Spearman, para medir la fuerza y la dirección de la relación entre las variables. Un valor cercano a 1 indica una correlación positiva, cerca de -1 indica una correlación negativa y cerca de 0 indica una correlación débil. Su característica principal es ayudar a identificar relaciones lineales entre variables numéricas. El coeficiente de correlación de Pearson se utiliza cuando las variables tienen una distribución normal, mientras que el coeficiente de correlación de Spearman se utiliza cuando no se cumple esta suposición

3.4.2 INSTRUMENTOS

Para la selección de variables y el entrenamiento de un modelo predictivo, se utilizaron varias herramientas claves para este proceso:

- SQL (Recolección de datos y Transformaciones de datos - Preprocesamiento):
    - Utilizado para extraer datos de fuentes de bases de datos.
    - Realiza transformaciones de datos, como la limpieza, filtrado y agregación de datos.
    - Preprocesa los datos para que sean aptos para el modelado.
- Knime (Modelamiento):
    - Utilizado para construir y entrenar modelos de predicción.

- Proporciona una interfaz gráfica para el flujo de trabajo de análisis de datos y modelado.
- Permite la configuración de algoritmos de aprendizaje automático y evaluación de modelos.

- Python (Selección de características con modelos Random Forest y XGBoost):
  - Empleado para la selección de características utilizando algoritmos de aprendizaje automático como Random Forest y XGBoost.
  - Estas técnicas fueron utilizadas para identificar las variables más relevantes para el modelo de predicción.
- Excel (Cálculos y Visualizaciones estadísticas):
  - Utilizado para realizar cálculos estadísticos adicionales y análisis exploratorio de datos (EDA).
  - Crea visualizaciones gráficas para comprender mejor los datos y los resultados del modelo.

3.4.3 PROCEDIMIENTOS APLICADOS

3.4.3.1 MODELOS APLICADOS

Una vez realizados estos ajustes, se planteó la estimación mediante el uso de diversos modelos de Machine Learning, con el propósito de llevar a cabo una comparación exhaustiva de su eficiencia en la predicción de clientes propensos a entrar en mora. De entre el conjunto inicial de modelos evaluados, se seleccionaron aquellos que demostraron un rendimiento superior para nuestro proyecto, los cuales son los siguientes: Naive Bayes, Decision Tree Learner y Logistic Regression Learner. Estos modelos destacaron por su capacidad para proporcionar estimaciones precisas y fiables en la detección de posibles morosos. La elección de estos modelos se basó en criterios rigurosos de evaluación, validación y pruebas, y sus resultados sobresalientes los convierten en herramientas esenciales para mejorar la gestión de riesgos en la institución financiera objeto de este estudio.

- Naive Bayes: es un algoritmo de aprendizaje supervisado que se basa en el teorema de

Bayes y hace una suposición "naive" o ingenua de independencia condicional entre las características (variables predictoras). Se utiliza principalmente para tareas de clasificación y categorización, como la clasificación de spam en correos electrónicos, análisis de sentimientos en textos y diagnóstico médico, entre otros. El funcionamiento del modelo implica calcular la probabilidad condicional de que una instancia pertenezca a una clase específica, dadas sus características. Luego, el modelo clasifica la instancia en la clase que tiene la probabilidad más alta.

Las características principales de Naive Bayes incluyen:

- Fácil de implementar y entender: El modelo Naive Bayes es relativamente sencillo de implementar y comprender, lo que lo hace adecuado para aplicaciones donde se necesita una solución rápida y efectiva.
- Buen rendimiento en datos con muchas características categóricas: Es especialmente útil cuando se trabaja con conjuntos de datos que contienen muchas características categóricas o discretas, como palabras en un texto.
- Suposición de independencia condicional: Esta suposición simplifica el cálculo de las probabilidades condicionales, lo que facilita la aplicación del modelo. Sin embargo, esta suposición puede no ser válida en todas las situaciones, lo que se conoce como "naivety".

- Decision Tree Learner: es también conocido como aprendizaje de árboles de decisión, es un modelo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Su función principal consiste en dividir un conjunto de datos en subconjuntos más pequeños basándose en características específicas, con el fin de tomar decisiones.

Este modelo se utiliza en tareas de clasificación y regresión, y se destaca por su capacidad para interpretar y visualizar cómo se toman las decisiones. Su funcionamiento implica dividir el conjunto de datos en nodos utilizando reglas basadas en características, continuando esta división hasta que se cumpla un criterio de parada, como la pureza de un nodo. Posteriormente, clasifica las instancias basándose en las hojas del árbol.

Las características clave del Decision Tree Learner son:

- Fácil de interpretar y visualizar: Los árboles de decisión son modelos que se pueden

entender y representar de manera intuitiva, lo que los hace útiles cuando se necesita comprender cómo se toman las decisiones en un problema.

- Puede manejar características categóricas y numéricas: Este modelo es versátil y puede manejar tanto características categóricas como numéricas, lo que lo hace adecuado para una variedad de tipos de datos.
- Puede ser propenso al sobreajuste: Si no se controla adecuadamente, los árboles de decisión pueden ajustarse demasiado a los datos de entrenamiento y no generalizar bien a nuevos datos, lo que se conoce como sobreajuste.
- Requiere ajuste de hiperparámetros: Para lograr un rendimiento óptimo, es necesario ajustar los hiperparámetros del árbol de decisión, como la profundidad máxima del árbol o el número mínimo de muestras por hoja.

- Logistic Regression Learner: La Regresión Logística es un modelo de clasificación que se utiliza para estimar la probabilidad de que una instancia pertenezca a una de dos clases binarias. Se utiliza principalmente en tareas de clasificación binaria, como predecir si un cliente es moroso o no, o diagnosticar si un paciente tiene una enfermedad basándose en pruebas médicas. El funcionamiento de este modelo implica el uso de la función logística para transformar la suma ponderada de las características en una probabilidad. Luego, se emplea un umbral para clasificar las instancias en una de las dos clases.

Las características clave de la Regresión Logística son:

- Simple de implementar y entender: La Regresión Logística es un modelo relativamente sencillo de implementar y comprender, lo que la hace adecuada para aplicaciones donde se requiere una solución rápida y efectiva.
- Útil para problemas de clasificación binaria: Su utilidad principal radica en problemas de clasificación con dos clases, donde se necesita predecir la pertenencia a una de las dos categorías.
- No es adecuado para problemas de múltiples clases sin modificaciones: La Regresión Logística se diseñó originalmente para problemas de clasificación binaria y requiere adaptaciones para manejar problemas de múltiples clases.

- No maneja automáticamente la interacción de características: La Regresión Logística no captura automáticamente interacciones complejas entre características, lo que puede requerir ingeniería de características adicional en ciertos casos.

3.4.3.2 EVALUCION DE CALIDAD DE LOS MODELOS

Para evaluar la calidad de los modelos entrenados se utilizaron los siguientes indicadores:

- Accuracy: También conocido como "exactitud", es un indicador utilizado en la evaluación de modelos de clasificación, como algoritmos de aprendizaje automático, para medir qué tan bien el modelo clasifica las observaciones correctamente. Sirve para evaluar la calidad del modelo al calcular la proporción de predicciones correctas en relación con el total de predicciones realizadas. Para calcularlo, simplemente se divide el número de predicciones correctas por el número total de predicciones, y se expresa como un porcentaje. Por ejemplo, si un modelo de clasificación hizo 90 predicciones correctas de un total de 100, su Accuracy sería del 90%. El Accuracy es una medida fundamental para comprender la efectividad de un modelo en términos de clasificación.

  Entre sus principales destacan las siguientes:

  - Fácil de entender: El Accuracy es un indicador simple y fácil de comprender, ya que se expresa como un porcentaje que va del 0% al 100%.
  - No considera el desequilibrio de clases: Una limitación del Accuracy es que puede ser engañoso cuando se trata de conjuntos de datos desequilibrados, donde una clase es mucho más frecuente que la otra. En tales casos, un modelo que predice siempre la clase mayoritaria podría tener un alto Accuracy, pero ser ineficaz.
  - Dependencia del umbral de decisión: El Accuracy no tiene en cuenta los falsos positivos y falsos negativos de un modelo, lo que significa que no considera las consecuencias reales de las predicciones erróneas. En algunos casos, es importante considerar más el impacto de los errores en una clase específica, y para eso, se utilizan otras métricas como la sensibilidad y la especificidad.

- Matriz de confusión: La matriz de confusión es una herramienta en estadística y

aprendizaje automático que evalúa el rendimiento de un modelo de clasificación. Sirve para medir la eficiencia y precisión al comparar las predicciones del modelo con los valores reales, ayudando a comprender la clasificación y detectar errores y se utiliza de la siguiente manera:

- La matriz de confusión se construye comparando las predicciones del modelo con las etiquetas verdaderas (valores reales) en un conjunto de datos de prueba. Se divide en cuatro partes:
  - Verdaderos positivos (VP): Observaciones positivas correctamente clasificadas.
  - Verdaderos negativos (VN): Observaciones negativas correctamente clasificadas.
  - Falsos positivos (FP): Observaciones negativas incorrectamente clasificadas como positivas (error Tipo I).
  - Falsos negativos (FN): Observaciones positivas incorrectamente clasificadas como negativas (error Tipo II).
- A partir de estos valores, se pueden calcular diversas métricas, como la precisión, la sensibilidad (recall), la especificidad y la F1-score, para evaluar la eficiencia del modelo.

La matriz de confusión es esencial en la evaluación de modelos de clasificación sus características principales permiten visualizar y cuantificar aciertos y errores de manera intuitiva. Identifica tipos de errores (falsos positivos/negativos) y se usa para calcular métricas como precisión, sensibilidad, especificidad y F1-score. Es crucial en conjuntos de datos desequilibrados para detectar sesgos hacia una clase dominante.

- Precission-Recall:
  - Precision: Precision es un indicador utilizado en problemas de clasificación que mide la proporción de verdaderos positivos (TP) respecto al total de predicciones positivas (TP + falsos positivos, FP). Sirve para evaluar cuán confiables son las predicciones positivas. Indica la proporción de predicciones positivas que son

realmente correctas.

- Recall (también conocido como Sensibilidad o Tasa de Verdaderos Positivos): Recall mide la proporción de verdaderos positivos (TP) respecto al total de casos positivos reales (TP + falsos negativos, FN). Sirve para evaluar cuán efectivo es un modelo en la identificación de todos los casos positivos en un conjunto de datos. Es útil cuando es importante no perder ningún caso positivo.

Se utiliza de la siguiente manera:

- Para calcular Precision: Precision = TP / (TP + FP)
- Para calcular Recall: Recall = TP / (TP + FN)
- Ambos indicadores varían entre 0 y 1, donde un valor más alto indica un mejor rendimiento. Generalmente, se busca un equilibrio entre Precision y Recall según las necesidades específicas del problema.

Entre sus principales características podemos encontrar las siguientes:

- Precision se enfoca en la precisión de las predicciones positivas, mientras que Recall se enfoca en la cantidad de casos positivos reales detectados.
- Los falsos positivos afectan la Precision, mientras que los falsos negativos afectan el Recall.
- En problemas desequilibrados, es común considerar ambos indicadores para una evaluación completa del modelo.

- Curvas de Roc: La Curva ROC (Receiver Operating Characteristic) es una herramienta gráfica utilizada en estadísticas y análisis de datos para evaluar la capacidad de un modelo de clasificación binaria, como un modelo de aprendizaje automático, para distinguir entre dos clases o categorías. Su principal función es medir y visualizar el rendimiento de dicho modelo en términos de su capacidad para discriminar entre verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Esto ayuda a seleccionar el umbral de clasificación óptimo y a comprender cómo se comporta el modelo en diferentes puntos de corte, lo que es esencial para evaluar su eficacia en la tarea de clasificación binaria.

  Para construir una Curva ROC, primero obtienes las probabilidades de clasificación

del modelo en lugar de las predicciones binarias. Luego, varías el umbral de decisión y calculas la tasa de verdaderos positivos y la tasa de falsos positivos en cada punto de corte. Estas tasas se grafican en un gráfico, donde el eje x representa la tasa de falsos positivos y el eje y representa la tasa de verdaderos positivos. La diagonal representa un modelo aleatorio, y el área bajo la curva (AUC) se usa para resumir la eficiencia del modelo: cuanto mayor sea el AUC, mejor será el rendimiento del modelo. En resumen, la Curva ROC evalúa la capacidad de un modelo para clasificar en función de diferentes umbrales de decisión utilizando probabilidades en lugar de predicciones binarias.

Las principales características son las siguientes:

- Gráfica de dos dimensiones: La Curva ROC es una representación bidimensional que muestra la relación entre las tasas de verdaderos positivos y falsos positivos.
- AUC (Área bajo la curva): El AUC es una métrica comúnmente asociada con la Curva ROC. Cuanto mayor sea el AUC, mejor será el rendimiento del modelo.
- Evaluación de clasificación binaria: Se utiliza específicamente para evaluar modelos de clasificación binaria y su capacidad para distinguir entre dos clases.
- Ayuda en la selección del umbral: Permite ajustar el umbral de decisión del modelo según los requisitos específicos de la aplicación, optimizando el equilibrio entre las tasas de falsos positivos y verdaderos positivos.

3.5 FUENTES DE INFORMACIÓN

3.5.1 FUENTES PRIMARIAS

Las fuentes primarias en esta investigación se centran en los datos proporcionados directamente por la institución financiera objeto de estudio. La base de datos utilizada en este trabajo es una fuente primaria fundamental, ya que contiene información detallada sobre la cartera de clientes de la institución financiera. Esta base de datos fue proporcionada por la institución financiera en cuestión y se compone de registros financieros históricos, que incluyen datos de transacciones, perfiles de clientes y métricas de rendimiento.

3.5.2 FUENTES SECUNDARIAS

Libros y artículos sobre la predicción de riesgo de impago. Estos libros y artículos

proporcionan información sobre los diferentes métodos y técnicas utilizados para predecir el riesgo de impago, así como sobre las limitaciones de estos métodos.

Artículos sobre machine learning. Estos artículos proporcionan información sobre los diferentes algoritmos de machine learning utilizados para la predicción de riesgo de impago.

CAPÍTULO IV. RESULTADOS Y ANÁLISIS

4.1 INFORME DE PROCESO DE RECOLECCIÓN DE DATOS

4.1.1 ANALISIS EXPLORATORIO DE LOS DATOS VARIABLES CATEGORICAS

En el ámbito financiero, la toma de decisiones efectiva basada en datos en crucial para garantizar la estabilidad y el correcto crecimiento de las instituciones. El riesgo del crédito es uno de los desafíos más significativos a los que se enfrentan las instituciones financieras, ya que la concesión de créditos es parte central del negocio de intermediación.

En análisis exploratorio de los datos (EDA) es una herramienta esencial para comprender y abordar el problema del riesgo crediticio. El EDA implica la exploración y el análisis de los datos disponible, su objetivo principal es revelar patrones, conocer tendencias, y relaciones en los datos.

En diciembre de 2019, la tasa de morosidad en la cartera de consumo de nuestra institución financiera se encontraba en un nivel del 7.7%. Sin embargo, en diciembre de 2022, esta tasa había aumentado significativamente, llegando al preocupante 16.8%. Este aumento representa un desafío financiero que requiere un análisis profundo y estrategias efectivas para revertir la tendencia.

Con el objetivo de realizar un análisis más enfocado y específico en el contexto de la gestión del riesgo crediticio, se ha procedido a seleccionar una muestra particular de la cartera de crédito al 30 de abril. Esta selección se centra en los préstamos de consumo con garantía fiduciaria que cumplen con dos criterios específicos: presentan una morosidad menor a 15 días, que no pertenecen y que su préstamo tenga una antigüedad mínima de 12 meses.

Esta selección cuidadosa de la muestra nos permitirá examinar detenidamente un subconjunto de préstamos que, en ese momento, presentaban un nivel de riesgo aún gestionable. Nos aseguramos de que la muestra sea representativa de la cartera de préstamos a clientes externos,

lo que es fundamental para nuestra evaluación de riesgo crediticio. A lo largo de este análisis, exploraremos en profundidad los factores que influyen en el comportamiento de esta muestra específica, lo que proporcionará una base sólida para el desarrollo de estrategias de gestión de riesgos efectivas y enfocadas en la mejora de la calidad de la cartera de préstamos de consumo con garantía fiduciaria.

Para iniciar nuestro análisis, es crucial categorizar a los clientes en dos grupos principales:

Grupo "Pago" (95.7%): Esta categoría representa la gran mayoría de los clientes, lo que indica que la mayoría cumple con sus obligaciones de pago puntualmente. Esto es un indicio positivo de la solidez de la cartera crediticia en términos de cumplimiento de pagos.

Grupo "No Pago" (4.3%): Aunque este grupo es numéricamente pequeño en comparación con el grupo de "Pago", su relevancia radica en que, después de tres meses, estos clientes caen en mora mayor a 30 días. Esta transición representa un indicio claro de riesgo crediticio y subraya la necesidad apremiante de comprender las causas subyacentes de esta morosidad y de desarrollar estrategias que permitan tanto identificar como predecir a los clientes que podrían caer en mora en el futuro.

En resumen, la mayoría de los clientes se mantienen al día con sus pagos, lo que es una señal positiva. Sin embargo, el pequeño porcentaje de clientes que cae en mora mayor a 30 días merece una atención especial. El análisis posterior podría centrarse en identificar los factores que contribuyen a la morosidad en este grupo y en desarrollar estrategias para prevenir y gestionar el riesgo crediticio.

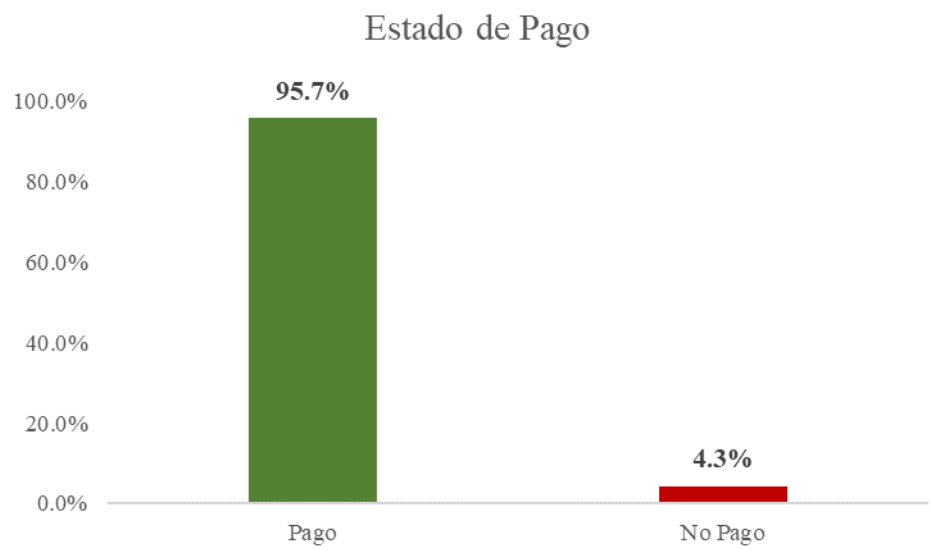


Figura 8: Estado de Pago (Pago o No Pago).

Fuentes: Elaboración propia

Para un análisis más completo, a continuación, exploraremos en detalle las variables esenciales, incluyendo aspectos demográficos de los clientes, con el fin de adquirir una comprensión más profunda de los elementos que ejercen influencia en el comportamiento de pago de nuestros clientes

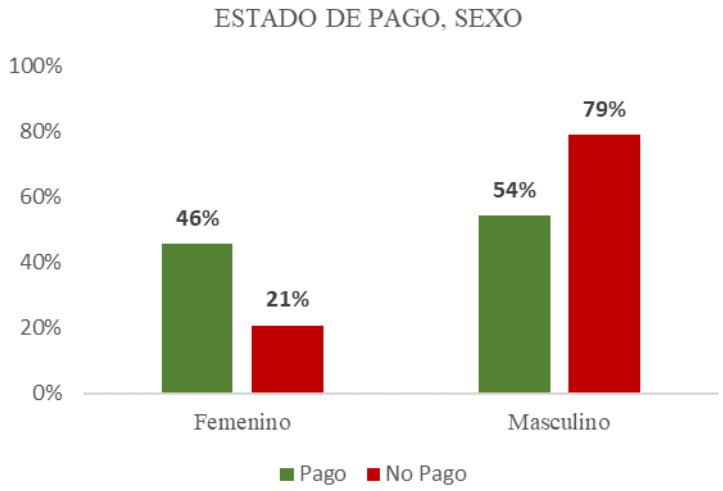


Figura 9: Estado de Pago y Género

Fuentes: Elaboración propia

La figura 9 desglosan el comportamiento de los clientes en términos de género (Femenino y Masculino) y su relación con el cumplimiento de pagos. Aquí se destacan algunas observaciones clave:

Género y Cumplimiento de Pagos: Se observa que, en general, las clientes femeninas tienen una tasa de cumplimiento de pagos ligeramente más alta (54%) en comparación con los clientes masculinos (46%). Sin embargo, es importante contextualizar estos datos, el 55% fueron otorgados al género masculino, mientras que el 45% corresponden al género femenino.

Género y Riesgo de Morosidad: Por otro lado, los clientes masculinos presentan una tasa de incumplimiento (79%) mucho más alta que las clientes femeninas (21%).

Al analizar los datos relacionados con el género y el comportamiento de pago de los clientes, se destaca una diferencia significativa en la tasa de incumplimiento. En general, los clientes masculinos muestran una tasa de incumplimiento del 79%, mientras que las clientes femeninas tienen una tasa de incumplimiento del 21%. Es importante tener en cuenta que, a pesar de esta brecha en las tasas de incumplimiento, el 55% de los préstamos se otorgaron a clientes masculinos, en comparación con el 45% otorgado a clientes femeninas. Esta observación nos recuerda la necesidad de considerar no solo las tasas de incumplimiento, sino también la distribución de género en la cartera de préstamos al evaluar el riesgo crediticio y desarrollar estrategias de gestión adecuadas.

Ahora, nos centraremos en analizar la variable estado civil de los clientes y su comportamiento de pago. Analizaremos cómo las diferentes categorías de estado civil se correlacionan con las tasas de incumplimiento y cumplimiento de pagos en la cartera de préstamos

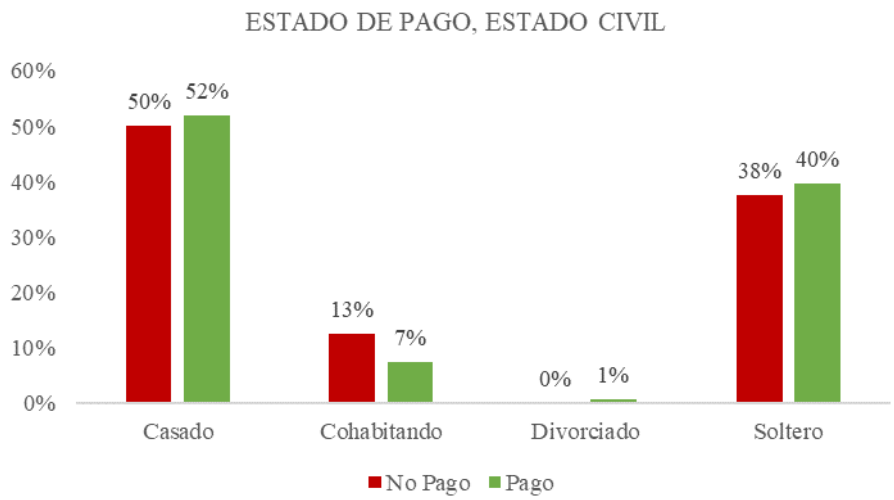


Figura 10: Estado de Pago, Estado Civil.

Fuentes: Elaboración propia

Clientes Casados: Muestran una tasa de incumplimiento del 50% y una tasa de cumplimiento del 52%. La diferencia entre estas tasas es también pequeña, pero en este grupo, la tasa de incumplimiento es ligeramente más alta que la tasa de cumplimiento.

Clientes Solteros: Presentan una tasa de incumplimiento del 38% y una tasa de cumplimiento del 40%. Esta diferencia es relativamente pequeña, lo que sugiere un comportamiento de pago similar entre los clientes solteros.

Al enfocarnos en los clientes solteros y casados, los datos indican que ambos grupos tienen tasas de incumplimiento y cumplimiento de pagos muy similares. La diferencia entre estas tasas en ambos grupos es pequeña, lo que sugiere que el estado civil, en este caso, parece tener un impacto limitado en el comportamiento de pago de los clientes en la cartera de préstamos.

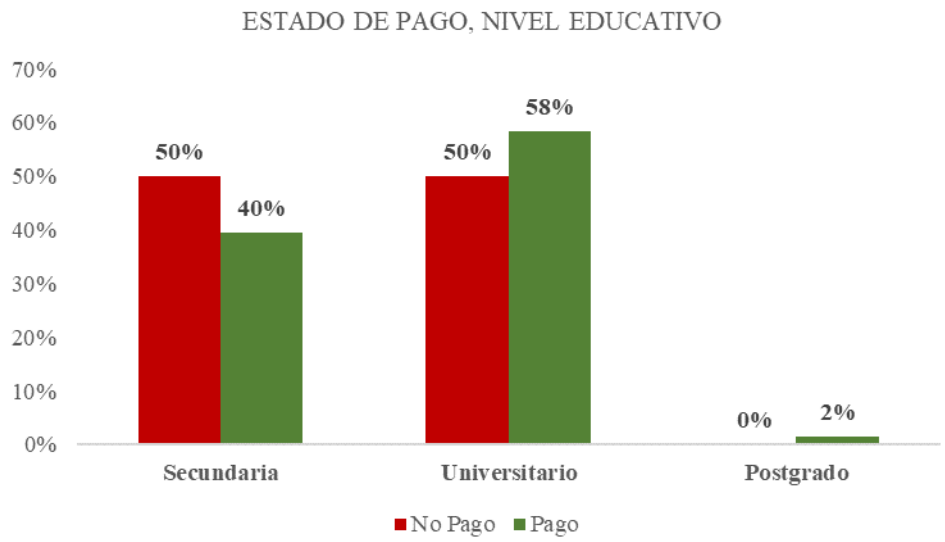


Figura 11: Estado de Pago y Nivel Educativo

Fuentes: Elaboración propia

Dentro del grupo de clientes que no cumplieron con sus pagos, se observa que tanto los clientes con nivel educativo de secundaria como los universitarios muestran una tasa de incumplimiento del 50%. Esto indica que, en este conjunto de datos, no hay una diferencia significativa en las tasas de incumplimiento entre estos dos grupos. Por otro lado, en el grupo de clientes que cumplieron con sus pagos, el 40% tiene un nivel educativo de bachillerato y el 58% tiene un nivel educativo universitario. Esto sugiere que, aunque hay una representación significativa de clientes con niveles educativos más bajos en el grupo de cumplimiento de pagos, la mayoría aún cumple con sus pagos. Además, es relevante mencionar que un pequeño porcentaje de clientes con nivel educativo de maestría (2%) también forma parte de este grupo de cumplimiento de pagos.

Después de analizar el comportamiento de pago en función del nivel educativo, nos centrémonos en cómo el impago se relaciona con el departamento de origen de los clientes que se les concedió un crédito. Esto nos permitirá comprender mejor si existen patrones regionales que puedan influir en las tasas de incumplimiento en nuestra cartera crediticia.

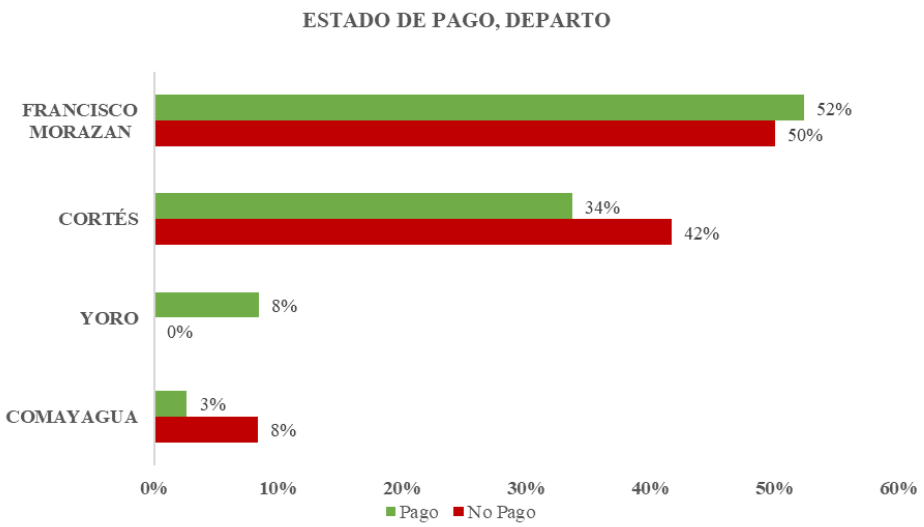


Figura 12 Gráfico de Barras Apiladas por Estado de Pago y Departamento

Fuentes: Elaboración propia

El departamento de Francisco Morazán tiene una alta incidencia de no cumplimiento de sus obligaciones crediticias, representando el 52% de los casos de incumplimiento. Es importante destacar que el 52% del total de préstamos que otorga la institución son otorgados en este departamento. Si consideramos el 100% de los préstamos que cayeron en mora mayor a 30 días, el 92% de estos casos se concentra en los departamentos de Francisco Morazán y Cortés. Estos dos departamentos representan en conjuntamente el 86% de los préstamos de consumo fiduciario otorgados por el Banco.

Es importante destacar que, dentro de la cartera de consumo, el departamento de Yoro representa el 8% de los préstamos otorgados. Sin embargo, lo que resalta aún más es que el 100% de los préstamos otorgados en Yoro se encuentran en cumplimiento, lo que significa que ninguno de los clientes de este departamento ha caído en mora mayor a 30 días.

Después de analizar el impacto de la ubicación geográfica en las tasas de impago, exploraremos cómo el nivel de ingreso de los clientes puede influir en su comportamiento de pago y en las tasas de impago. Esto nos permitirá comprender mejor si existen patrones relacionados con los ingresos que puedan contribuir al impago en nuestra cartera crediticia.

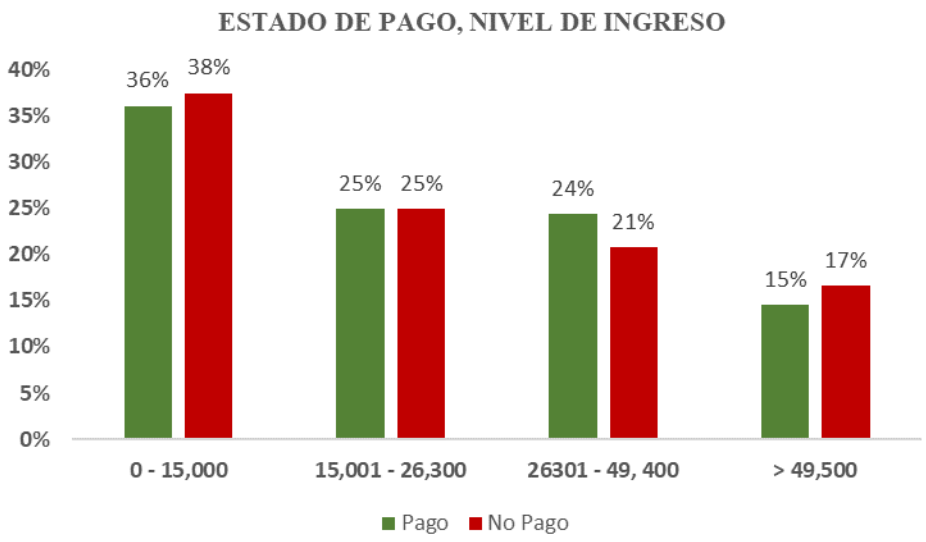


Figura 13: Estado de Pago y Nivel de Ingreso

Fuentes: Elaboración propia

Los datos de la Figura 13 sugieren que el nivel de ingreso no parece ser un factor determinante en el comportamiento de pago de los clientes en la cartera de consumo. Aunque existen algunas diferencias en las tasas de cumplimiento y no cumplimiento entre los grupos de ingresos, estas diferencias no son significativas. Los datos revelan que no existe una relación lineal clara entre el nivel de ingreso y el comportamiento de pago de los clientes.

4.1.2 ANALISIS INFERENCIAL PRUEBA DE INDEPENDENCIA VARIABLE CATEGORICAS UTILIZANDO LA PRUEBA CHI-CUADRADO

El análisis de independencia de variables categóricas desempeña un papel fundamental en la exploración de relaciones significativas entre diferentes atributos o características y la variable de interés, en nuestro caso, Estado de Pago. En este contexto, el objetivo principal es evaluar si existe una asociación estadísticamente significativa entre las variables categóricas consideradas y la ocurrencia de impagos. La prueba de Chi-cuadrado se convierte en una herramienta esencial para llevar a cabo esta evaluación, permitiendo determinar si las variables categóricas (edad, sexo, niel de ingresos, departamento, agencia) e Estado de Pago están relacionadas de manera independiente o si existe una dependencia estadística significativa entre ellas.

Tabla 4: Asociación de las variables categóricas con la variable dependiente Estado de Pago.

Variable	Chi-Cuadrado	Valor P	Significativo
Agencia	26.6533	0.0634	No
Zona	0.9710	0.3244	No
Departamento	7.5717	0.6706	No
Nivel Educacion	8.5024	0.3860	No
Ingresos	0.2154	0.9751	No
Estado Civil	1.0670	0.8995	No
Genero	4.8016	0.0284	Si

Fuentes: Elaboración propia

El análisis Chi-Cuadrado nos revela que no existe asociación significativa entre las variables categóricas agencia, zona, departamento, nivel educativo, ingresos y estado civil con la variable dependiente. No obstante, la variable genero puede desempeñar un papel en la predicción del impago. Es importante considerar que, en un modelo predictivo, las variables se consideran en un contexto más complejo es posible que algunas variables que no tienen asociación con la variable dependiente sean relevantes cuando se combinan con otras variables.

4.1.3 ANALISIS EXPLORATORIO DE LOS DATOS VARIABLES CONTINUAS

Después de haber realizado un análisis exhaustivo de las variables categóricas, continuaremos nuestro estudio enfocándonos en las variables continuas. En esta etapa, exploraremos en profundidad estas variables cuantitativas para obtener una comprensión detallada de su comportamiento y distribución en nuestros datos. Esto nos proporcionará una visión completa de todas las dimensiones de nuestro conjunto de datos, permitiéndonos evaluar cómo se relacionan estas variables con nuestra variable dependiente, "Estado de Pago".

Para llevar a cabo este análisis, utilizaremos una variedad de métodos gráficos y estadísticas descriptivas. Estas herramientas nos ayudarán a identificar tendencias, patrones y diferencias significativas entre los grupos de clientes que cumplieron y los que cayeron en mora en sus pagos. De esta manera, podremos obtener información valiosa que respalde la toma de

decisiones informadas en el ámbito financiero y crediticio.

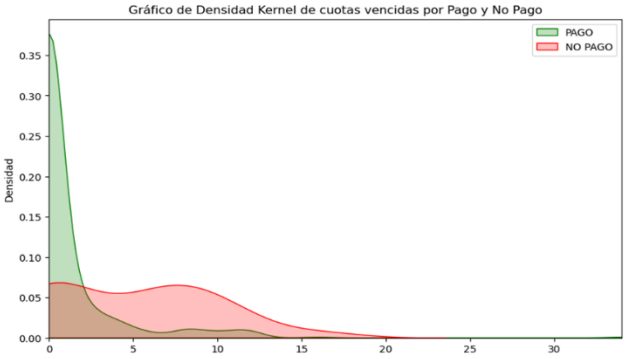


Figura 15: Cuotas vencidas consumo fiduciario promedio grafico de Kernel

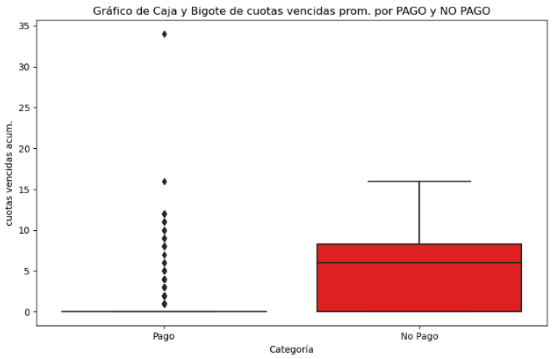


Figura 14: Cuotas vencidas consumo fiduciario promedio boxplot

Fuentes: Elaboración propia

La variable número de cuotas vencidas promedio año, proporciona información valiosa sobre el comportamiento de pago de los clientes y su relación con la variable "Estado de Pago". En la gráfica de densidad de kernel, observamos que la mayoría de los clientes que cumplen con sus pagos tienen cero cuotas vencidas, lo que sugiere que estos clientes tienden a estar al día con sus obligaciones crediticias. Por otro lado, entre los clientes que no cumplen con sus pagos, se observa una distribución diferente, con una presencia significativa de casos que tienen una o más cuotas vencidas.

Esta variable puede ser un indicador importante para predecir si un cliente cae o no en mora mayor a 30 días. Los clientes que acumulan cuotas vencidas pueden estar en mayor riesgo de caer en mora en el futuro, por lo tanto, esta variable puede ser relevante para la predicción y la gestión del impago de la cartera de crédito fiduciario. Puede ser útil en la identificación temprana en clientes con signos de dificultades en el cumplimiento de sus obligaciones crediticias y en la aplicación de estrategias preventiva para reducir la morosidad.

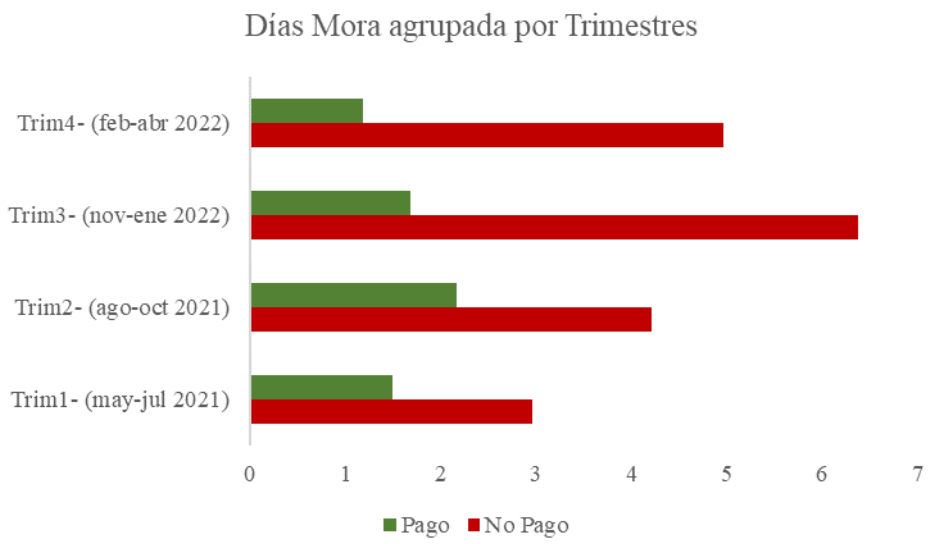


Figura 16: Días mora promedio agrupadas por trimestre

Fuentes: Elaboración propia

Los datos de la Figura 16 muestran claramente la importancia del análisis del comportamiento pasado en la predicción del riesgo crediticio. Por ejemplo, los clientes marcados que no pagaron en julio del 2022, el promedio de días mora fue de 5 días, en relación con aquellos clientes que pagaron el promedio de días mora fue de 1.6 días.

Esto respalda la idea de que el análisis del comportamiento histórico de morosidad de los clientes puede ser un factor importante para predecir su probabilidad de impago en el futuro. Los clientes que han demostrado un historial de morosidad persistente tienden a tener un mayor riesgo de impago en comparación con aquellos que han mantenido un historial de morosidad bajo. Por lo tanto, considerar el historial de morosidad en la evaluación del riesgo crediticio puede ser una práctica valiosa para tomar decisiones más precisas y eficientes en la gestión de carteras de crédito.

Siguiendo con el análisis sobre la relevancia del historial de morosidad, nos enfocaremos ahora en el número de meses que el cliente no cayó en mora durante el año. Esta métrica nos brinda información crucial sobre la capacidad de los clientes para mantener un historial de pagos sin interrupciones a lo largo del año.

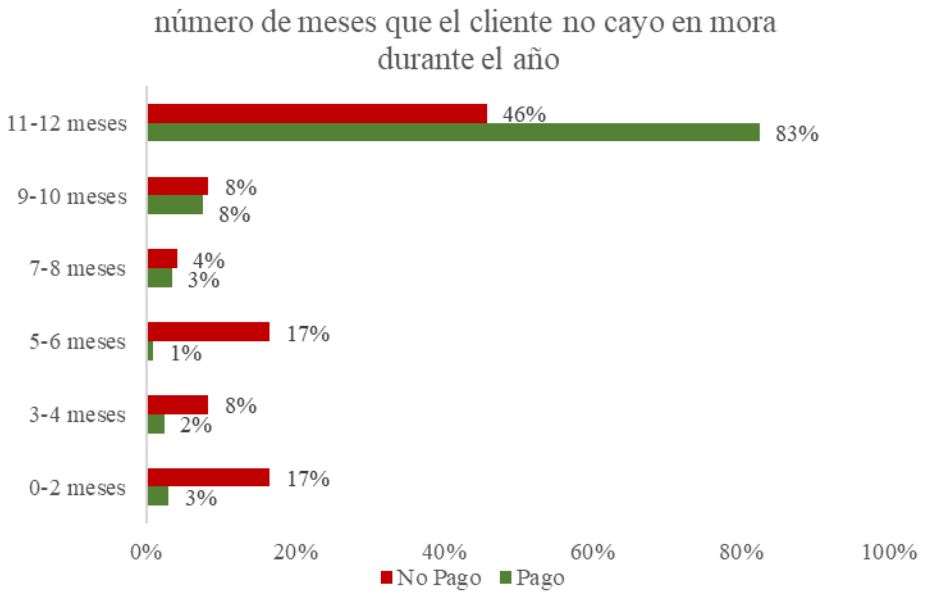


Figura 17: Número de meses que el cliente no cayó en mora.

Fuentes: Elaboración propia

Se observar una interesante relación entre el número de meses en los que un cliente no cayó en mora durante el año y su comportamiento de pago. En el grupo de clientes que cumplieron con sus pagos, destaca que el 83% de ellos logró mantener un historial limpio durante todo el año, estos se ubican dentro del rango de (11-12 veces). Este hecho sugiere una fuerte asociación entre la consistencia en el cumplimiento de pagos a lo largo del año y la puntualidad en las obligaciones crediticias. En contraste, en el grupo de clientes que no pagaron, solo el 46% pudo mantenerse sin morosidad durante todos los meses del año, mientras que un 17% cayó en mora entre 5-6 veces. Estos resultados subrayan la importancia de mantener una trayectoria de pagos constante como un indicador relevante para predecir el impago de los clientes y respaldan la idea de que el historial de morosidad es un factor fundamental en la evaluación del riesgo crediticio.

El 6% de los clientes que cumplieron con sus pagos pasaron a mora de 31-60 días ver gráfico 17, mayormente de manera ocasional y breve. Por otro lado, de los clientes que no pagaron, el 21% enfrentó esta morosidad, siendo más frecuente pero aún limitada en número de ocasiones, con un 80% de ellos en ese rango durante un máximo de 2 veces. Este análisis sugiere que la morosidad de 31-60 días es menos común entre los clientes que pagaron y, cuando ocurre, tiende a ser de corta duración. Por otro lado, los clientes que no pagaron mostraron una mayor incidencia

de morosidad de 31-60 días, aunque la mayoría de ellos también limitaron su mora a un número reducido de ocasiones

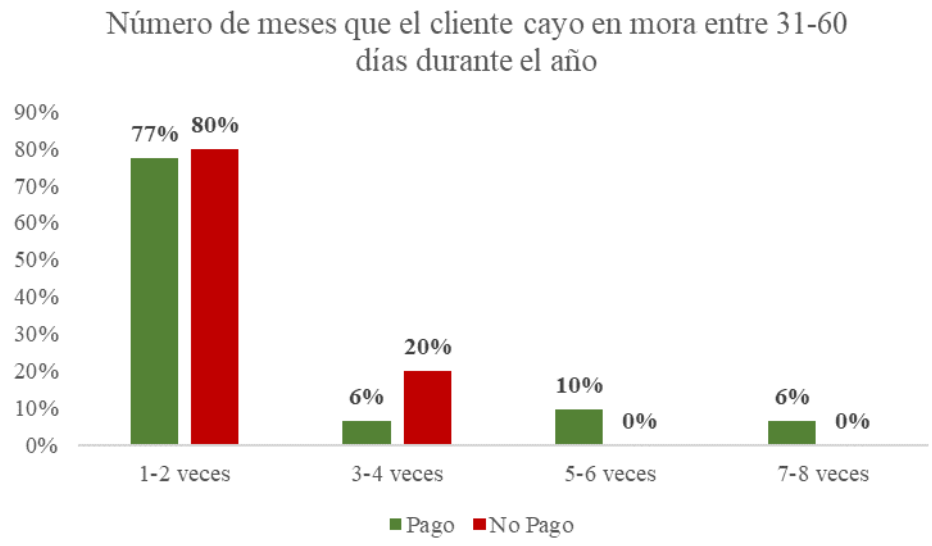


Figura 18: Número de meses que el cliente cayo en mora entre 31-60 días durante el año.

Fuentes: Elaboración propia

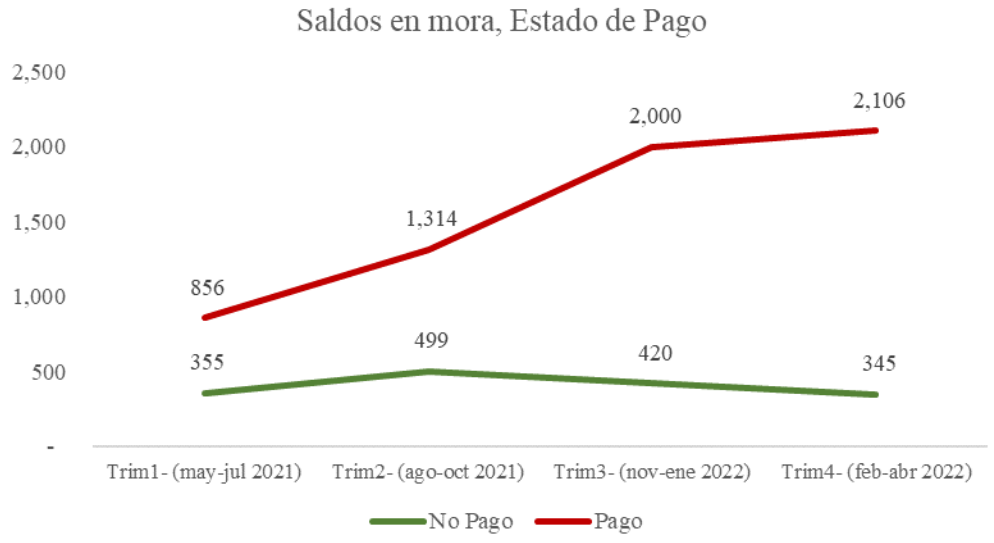


Figura 19: Saldos en mora por trimestre.

Fuentes: Elaboración propia

La tendencia mostrada en la Figura 19, nos revela que los clientes que “No Pagan” tienen 4 veces más en promedio saldo en mora que los clientes que cumplieron con sus obligaciones. Esta diferencia resalta una disparidad significativa en los costos financieros asumidos por ambos grupos. Es importante destacar que el grupo de "clientes que no pagan" enfrentara un mayor gasto en intereses, lo que podría indicar dificultades financieras o una mayor exposición a tasas de interés más altas.

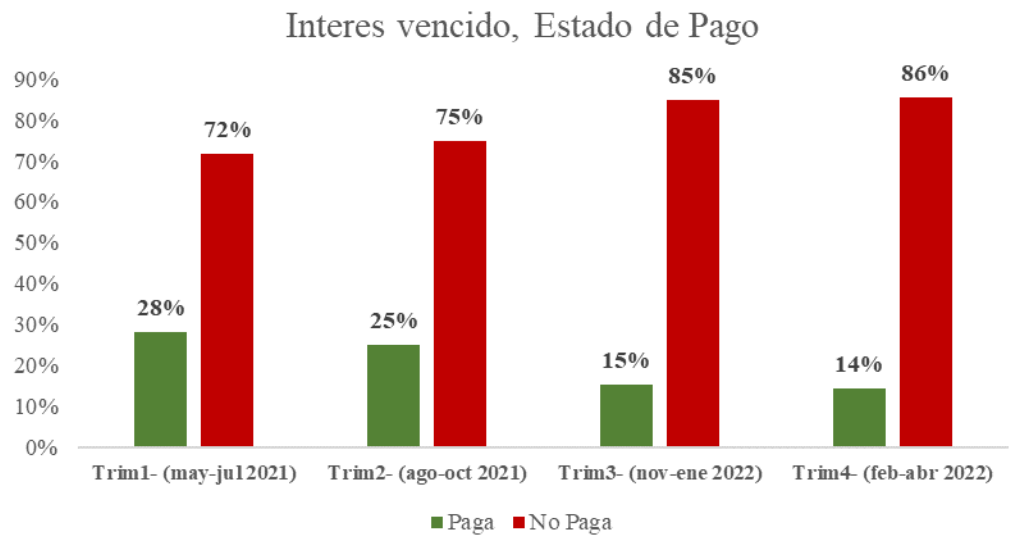


Figura 20: Interés vencido.

Fuentes: Elaboración propia

El análisis de los datos revela una disparidad significativa en el comportamiento de pago entre los dos grupos de clientes, en términos de intereses vencidos. A lo largo de los cuatro trimestres, los clientes que cumplen con sus obligaciones de pago han acumulado intereses vencidos sustancialmente más bajos en comparación con los clientes que no lo hacen. En total, los clientes “No Pago” han pagado aproximadamente 4 veces más en promedio en intereses vencidos que sus contrapartes del grupo “Pago”. Estos hallazgos resaltan la urgencia de abordar de manera efectiva el incumplimiento de pago, especialmente entre los clientes del grupo “No Pago”, y respaldan la necesidad de estrategias diferenciadas para la gestión de riesgos y la recuperación de deudas en función del historial de pago de los clientes.

Tabla 5: Resumen de Estadísticas Descriptivas de Variables Relevantes

Variables	Media		Desviación Estándar		Mediana	
	Pago	No Pago	Pago	No Pago	Pago	No Pago
Edad	42.31	40.00	10.56	8.42	40.00	37.50
Ingresos	29,015.09	28,805.04	19,546.02	18,563.13	20,001.00	20,250.00
Numero Dependientes	1.05	1.21	1.14	1.41	1.00	1.00
Años Empleo	10.11	7.08	7.12	3.31	8.00	6.50
Saldo cons-Fid.	173,985.12	157,918.83	406,046.55	121,975.14	103,454.20	112,680.92
Tasa Cons-Fid.	17.84	21.19	6.44	5.58	15.00	20.00
Dias mora al cierre permitidos	0.78	3.25	2.93	4.64	-	-
trim-1 dias mora prom.	1.49	2.96	7.07	5.48	-	-
trim-2 dias mora prom.	2.17	4.21	10.29	5.35	-	1.00
trim-3 dias mora prom.	1.67	6.38	10.08	9.58	-	3.50
trim-4 dias mora prom.	1.18	4.96	5.22	6.94	-	2.50
conteo meses nivel de mora = 0	11.02	8.04	2.44	4.21	12.00	9.00
cuotas vencidas acum.	1.11	5.04	2.93	4.75	-	6.00
pago capital promed año	4,077.08	3,976.11	3,722.73	2,764.63	3,058.39	2,829.67
cuota_lps_promed_año	8,264.21	10,154.74	8,935.45	8,867.72	5,618.27	5,227.39

Fuentes: Elaboración propia

La Tabla 5 presenta un resumen de las principales métricas de las variables analizadas, desglosadas por los grupos de pago y no pago. Entre las variables más destacadas se encuentran el número de cuotas vencidas y los días de mora promedio trimestrales, donde los clientes que no pagaron muestran valores considerablemente más altos. Asimismo, las variables relacionadas con el saldo total del préstamo y la tasa promedio evidencian diferencias significativas, con los clientes que no pagaron mostrando valores superiores.

Por otro lado, las variables demográficas y financieras, como la edad, años de empleo y número de dependientes parecen tener valores medios bastante similares entre ambos grupos. Esta similitud sugiere que, en el contexto de esta cartera de consumo fiduciario, estas variables pueden no ser determinantes en la predicción del impago. Aunque las medias sean parecidas, aún pueden existir relaciones y patrones interesantes en los datos que se deben explorar con mayor profundidad mediante análisis más avanzados y modelos predictivos.

4.1.4 ANALISIS INFERENCIAL DE LAS VARIABLES NÚMERICAS CONTINUAS

Tabla 6: Resultados de Pruebas de Hipótesis para Variables Independientes Continuas en Relación con Estado de Pago

Variable continua	Estadístico t	Valor p	Conclusión
Edad	- 1.2977	0.2056	No se rechaza la hipótesis nula
Ingresos	- 0.0541	0.9573	No se rechaza la hipótesis nula
Numero Dependientes	0.5437	0.5916	No se rechaza la hipótesis nula
Años Empleo	- 4.0701	0.0003	Se rechaza la hipótesis nula
Saldo cons-Fid.	- 0.5264	0.6009	No se rechaza la hipótesis nula
Tasa Cons-Fid.	2.8525	0.0084	Se rechaza la hipótesis nula
Dias mora al cierre permitidos	2.5835	0.0163	Se rechaza la hipótesis nula
trim-1 dias mora prom.	1.2675	0.2159	No se rechaza la hipótesis nula
trim-2 dias mora prom.	1.7290	0.0937	No se rechaza la hipótesis nula
trim-1 intereses prom.	1.1855	0.2464	No se rechaza la hipótesis nula
trim-2 intereses prom.	1.4501	0.1591	No se rechaza la hipótesis nula
Trim-1 Saldo en mora prom	1.5017	0.1455	No se rechaza la hipótesis nula
Trim-2 Saldo en mora prom	1.9055	0.0673	No se rechaza la hipótesis nula
Trim-1 Interes por mora prom	0.6919	0.4941	No se rechaza la hipótesis nula
Trim-2 Interes por mora prom	0.1186	0.9058	No se rechaza la hipótesis nula
Trim-1 Interes vencido prom	1.4122	0.1700	No se rechaza la hipótesis nula
Trim-2 Interes vencido prom	1.8437	0.0768	No se rechaza la hipótesis nula
conteo meses en mora consumo-fid	4.2432	0.0003	Se rechaza la hipótesis nula
conteo meses nivel de mora = 0	- 3.4413	0.0022	Se rechaza la hipótesis nula
conteo meses nivel de mora = 1-30	4.0173	0.0005	Se rechaza la hipótesis nula
conteo meses nivel de mora = 31-60	1.4409	0.1622	No se rechaza la hipótesis nula
cuotas vencidas acum.	4.0227	0.0005	Se rechaza la hipótesis nula
cuotas vencidas prom.	3.7548	0.0010	Se rechaza la hipótesis nula

Fuentes: Elaboración propia

En el análisis de las variables en relación con la variable dependiente “Estado de Pago”, se identificaron variables significativas que demostraron tener un impacto estadísticamente significativo en la predicción de impago. Estas variables, como la tasa de consumo fiduciario, los años de empleos son críticas para comprender y predecir los patrones de impago. Por otro lado, algunas variables no mostraron diferencias estadísticas significativas, pero su exclusión no es automática. Variables no significativas, como saldo del préstamo pueden desempeñar un papel importante debido a consideraciones teóricas, interacciones con otras variables o efectos prácticos. En el proceso de construcción del modelo de aprendizaje automático, estas variables no significativas merecen una evaluación más detallada antes de tomar decisiones definitivas sobre su inclusión o exclusión.

Tabla 7: Tabla de Correlaciones entre Variables Independientes y Estado de Pago

Variables	conteo meses nivel de mora = 1-30	conteo meses en mora consumo-fid	cuotas vencidas acum.	conteo meses nivel de mora = 0	Trim-4 Saldo en mora prom	Trim-4 Interes vencido prom	cuotas vencidas prom.	Estado de Pago
conteo meses nivel de mora = 1-	1.0000	0.8838	0.9534	0.8524	0.4756	0.6163	0.7591	0.2809
conteo meses en mora consumo-	0.8838	1.0000	0.9236	0.9868	0.6052	0.6587	0.8430	0.2769
cuotas vencidas acum.	0.9534	0.9236	1.0000	0.9000	0.5766	0.6837	0.9064	0.2564
conteo meses nivel de mora = 0	0.8524	0.9868	0.9000	1.0000	0.6004	0.6557	0.8292	0.2329
Trim-4 Saldo en mora prom	0.4756	0.6052	0.5766	0.6004	1.0000	0.6632	0.7387	0.2155
Trim-4 Interes vencido prom	0.6163	0.6587	0.6837	0.6557	0.6632	1.0000	0.6565	0.2138
cuotas vencidas prom.	0.7591	0.8430	0.9064	0.8292	0.7387	0.6565	1.0000	0.1958

Fuentes: Elaboración propia

El análisis de las correlaciones entre las variables independientes y la variable dependiente "Estado de Pago" es fundamental para identificar factores que influyen en el incumplimiento de pagos. Hemos encontrado correlaciones significativas, destacando la importancia de analizar detenidamente estas variables predictoras. Estos hallazgos sugieren que estas variables pueden desempeñar un papel relevante en la predicción de incumplimientos, crucial para desarrollar estrategias de gestión de riesgos y decisiones financieras efectivas

La alta correlación observada entre las variables "conteo meses nivel de mora = 1-30", "conteo meses en mora consumo-fid" y "conteo meses nivel de mora = 0" se puede atribuir principalmente a su origen común: la variable original "Días Mora". Dado que estas variables representan los días de mora categorizada por nivel, es natural esperar una fuerte relación entre ellas. Esta correlación elevada se justifica por la influencia que el tiempo ejerce sobre los datos, donde los patrones de mora en un trimestre están relacionados con los trimestres anteriores y posteriores.

4.1.5 METODOS DE MACHINE LEARNING PARA SELECCIÓN DE CARACTERISTICAS

Los métodos de selección de características son técnicas que se basan en ML para determinar cuáles de las múltiples variables disponible en un conjunto de datos contribuyen de manera efectiva a la predicción de la variable objetivo. Los métodos de machine learning permiten optimizar y simplificar los modelos al reducir la cantidad de variables utilizadas, lo que a la vez puede mejorar la capacidad de generalizar y reducir el sobreajuste.

El método Random Forest es una poderosa técnica de aprendizaje automático que se utiliza

para resolver problemas de clasificación. Se basa en la construcción de múltiples árboles de decisión y combina sus resultados para obtener predicciones más precisas y robustas. Una de las ventajas clave de Random Forest es su capacidad para evaluar la importancia de las características en la toma de decisiones del modelo.

Como resultado, se ha encontrado que las características Cuotas vencidas con-fid. Promed Año, Días mora al cierre permitidos y Trim-2 días mora son algunas de las principales variables influyentes en la predicción de nuestro variable objetivo, Estado de Pago. Estas características desempeñarán un papel crucial en nuestra evaluación de riesgos crediticios y en la toma de decisiones financieras.

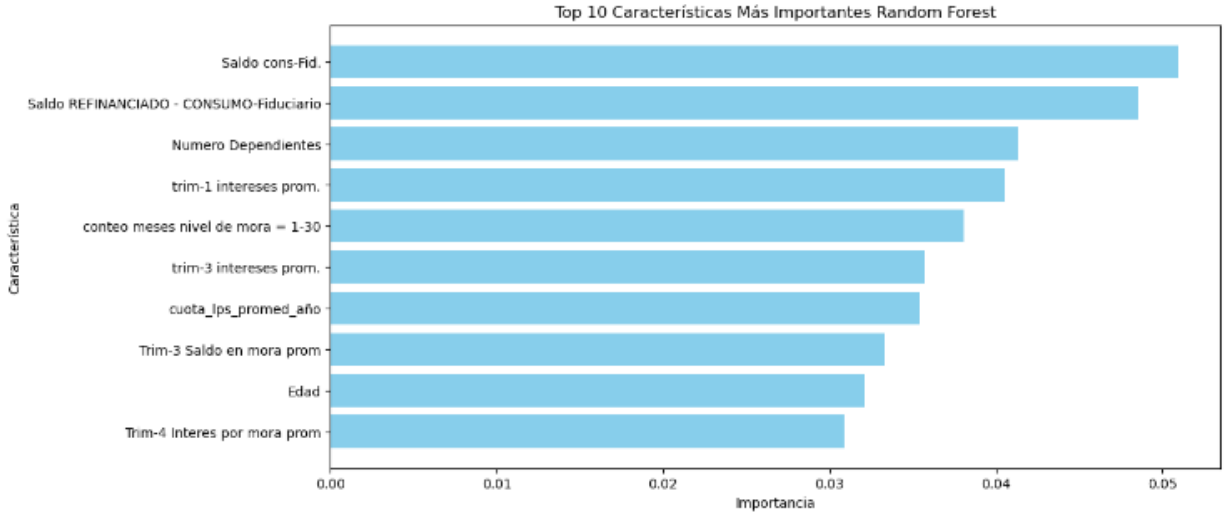


Figura 20: Principales características seleccionadas por el método Random Forest

Fuentes: Elaboración propia

Luego de examinar las principales características reveladas por el método Random Forest, continuaremos nuestro análisis exploratorio de selección de variables con el método XGBoost. Al igual que Random Forest, es ampliamente reconocido por su eficacia en la modelización de conjuntos de datos complejos y su capacidad para identificar variables influyentes. Ambos métodos nos brindarán una perspectiva valiosa sobre las características más relevantes relacionadas con nuestra variable objetivo Estado de Pago. Este método utiliza enfoques optimizados de gradientes para mejorar la precisión del modelo. Ambos métodos tanto Random Forest como el XGBoost se basan en la construcción de múltiples arboles de decisión.

Es importante observan que ambos métodos seleccionaron un conjunto similar de

características, aunque el orden o la importancia de la variable puede varían entre los métodos. Esta selección sugiere que estas variables pueden ser importante en la predicción del comportamiento de pago de los clientes en el fututo.

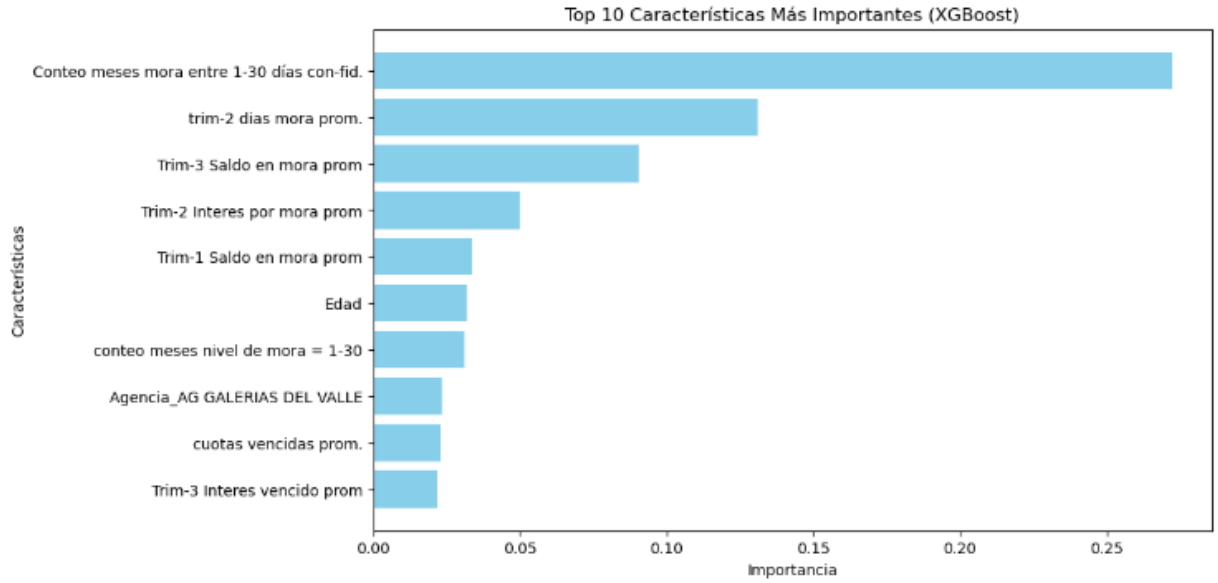


Figura 21: Principales características seleccionadas por el método XGBoost

Fuentes: Elaboración propia

4.1.6 SELECCIÓN DE LAS CARACTERISTICAS RELEVANTES

En resumen, el proceso de selección de las variables principales en nuestro estudio ha sido completo. Hemos aplicado métodos avanzados de selección de características, como Random Forest y XGBoost, que han identificado un conjunto clave de variables predictoras. Sin embargo, no nos hemos limitado únicamente a estas técnicas automatizadas. También hemos aprovechado un enfoque multidimensional que incorpora análisis gráficos, descriptivos y pruebas estadísticas específicas, como el análisis de chi-cuadrado para variables categóricas y la prueba t de Student para variables continuas.

Este enfoque nos ha permitido no solo identificar las variables más influyentes desde una perspectiva técnica, sino también comprender su contexto y relevancia en el ámbito financiero y crediticio. Como resultado, hemos establecido una base sólida y confiable para la toma de decisiones estratégicas en la gestión de riesgos crediticios.

Tabla 8: Selección de las características más relevantes

Variable	Correlación	Métodos de Estadística Inferencial		Metodos de Machine Learning	
		Chi-Cuadrado	t de student	Random Forest	XGBoost
Ingresos	Media	N/A	No	No	No
Años Empleo	Baja	N/A	Si	Si	-
Saldo cons-Fid.	Baja	N/A	No	Si	Si
Tasa Cons-Fid.	Media	N/A	Si	Si	-
Dias mora al cierre permitidos	Media	N/A	Si	-	Si
trim-1 dias mora prom.	Baja	N/A	No	Si	-
trim-2 dias mora prom.	Baja	N/A	No	-	Si
trim-3 dias mora prom.	Baja	N/A	Si	-	Si
trim-4 dias mora prom.	Media	N/A	Si	-	Si
Trim-1 Saldo en mora prom	Baja	N/A	No	-	Si
Trim-2 Saldo en mora prom	Baja	N/A	No	Si	-
Trim-3 Saldo en mora prom	Baja	N/A	Si	Si	Si
Trim-4 Saldo en mora prom	Media	N/A	Si	Si	Si
Trim-1 Interes vencido prom	Baja	N/A	No	Si	-
Trim-2 Interes vencido prom	Baja	N/A	No	Si	-
Trim-3 Interes vencido prom	Media	N/A	Si	Si	Si
Trim-4 Interes vencido prom	Media	N/A	Si	Si	-
Total días Mora Año Consumo Fid.	Baja	N/A	Si	-	-
Conteo meses mora = 0 con-fid.	Media	N/A	Si	-	-
Conteo meses mora entre 1-30 días con-fid.	Media	N/A	Si	-	Si
conteo meses en mora consumo-fid	Media	N/A	Si	-	-
cuotas vencidas acum.	Media	N/A	Si	-	-
pago capital promed año	Baja	N/A	No	Si	Si
cuota_lps_promed_año	Baja	N/A	No	Si	Si

Fuentes: Elaboración propia

Las variables presentadas en la Tabla 8 han sido cuidadosamente seleccionadas tras un riguroso proceso que involucró varios métodos de selección. Estas variables destacadas no solo son las más relevantes, sino que también desempeñan un papel fundamental en los resultados de nuestro modelo de machine learning. Los detalles y el impacto de estas variables en nuestro modelo de ML se analizarán en profundidad en el capítulo 6, donde se presentarán los resultados y se evaluará cómo estas variables influyen en la predicción del cumplimiento o incumplimiento del pago por parte de los clientes.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- Nuestra investigación demuestra que el comportamiento de pago de los clientes es fundamental para predecir el incumplimiento de pagos. Dentro de las variables analizadas, encontramos que el número promedio de cuotas vencidas por año, el historial de morosidad en términos de días de mora, el saldo en mora, los intereses vencidos, entre otras variables. Estas son indicadores fundamentales para identificar a los clientes que podrían estar en mayor riesgo de caer en mora y son esenciales para la toma de decisiones estratégicas en la gestión de riesgos crediticios y desempeñan un papel fundamental en la evaluación y mitigación de riesgos financieros

- Basado en el análisis gráfico y la prueba chi-cuadrado se ha constatado que la mayoría de estas variables demográficas como ser agencia, zona, departamento, nivel educativo, ingresos y estado civil entre otras, no presentan una asociación estadísticamente significativa con la variable 'Estado de Pago' en nuestra cartera de consumo fiduciario. Sin embargo, cabe resaltar la variable género, donde hemos identificado una relación estadísticamente significativa. Es esencial la tasa de incumplimiento es más alta entre los clientes masculinos, pero también representan una proporción mayor de préstamos otorgados en comparación con las clientes femeninas.

- Durante esta investigación, hemos analizado exhaustivamente las variables relacionadas con la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario en una entidad financiera. Se utilizaron diversas técnicas de análisis de datos y modelos de machine learning, evaluamos y seleccionamos las variables más influyentes para predecir el impago de los clientes.

- Con los resultados obtenidos por los modelos ML aplicados, se considera la creación de un dashboard diseñado específicamente para visualizar y supervisar de manera dinámica a los clientes con alta probabilidad de caer en mora en la cartera de consumo fiduciaria es una medida estratégica fundamental para la institución financiera. Esta herramienta proporciona una visión más clara y actualizada de los riesgos asociados a los préstamos y permite tomar decisiones informadas de manera proactiva. Al utilizar datos reales y análisis predictivos, el dashboard facilita la identificación temprana de clientes en riesgo, lo que a su vez permite implementar acciones preventivas y estratégicas para reducir la morosidad y proteger los activos financieros.
- En el capítulo 6 se define el desarrollo, la aplicabilidad y la evaluación de un modelo ML para predecir el impago.

5.2 RECOMENDACIONES

- En vista de los patrones y tendencias identificados en nuestro análisis financiero, es necesario que la institución financiera tome medidas proactivas y basadas en datos para abordar el riesgo crediticio en la cartera de consumo fiduciario. Además, se debe considerar la implementación de un sistema de monitoreo continuo de los clientes en riesgo, lo que permitiría una intervención temprana para evitar el incumplimiento, teniendo en cuenta especialmente las variables clave que hemos identificado, como los días de mora promedio, las cuotas vencidas y los intereses pagados. Estas medidas, respaldadas por un enfoque en la prevención y la toma de decisiones basada en datos, contribuirán significativamente a mejorar la gestión del riesgo crediticio y a reducir la morosidad en el Banco. Asimismo, se recomienda una revisión periódica y actualización de estas políticas y estrategias, considerando la evolución de los datos y el comportamiento de los clientes a lo largo del tiempo.

- Es de vital importancia abordar los problemas relacionados con la calidad de los datos y el acceso a la información. La falta de consistencia y la falta de conciliación en los datos, así como cualquier desafío en el acceso a información relevante, deben ser tratados con urgencia. Se recomienda realizar una revisión exhaustiva de la calidad de los datos, identificar y rectificar cualquier discrepancia, y garantizar un acceso fluido a los datos necesarios. Estas medidas no solo mejorarán la precisión de nuestro modelo predictivo, sino que también fortalecerán nuestra capacidad para gestionar eficazmente el riesgo crediticio.

CAPÍTULO VI. APLICABILIDAD

6.1 NOMBRE DE LA PROPUESTA

Implementación de un modelo de Machine Learning para predecir la probabilidad de impago de la cartera de cliente de consumo fiduciario

6.2 JUSTIFICACIÓN DE LA PROPUESTA

El análisis de los datos históricos de mora de la cartera de consumo adquiere una importancia fundamental para comprender la evolución del riesgo asociado a este producto. A diciembre de 2019, la tasa de morosidad de la cartera de consumo se situaba en un 7.7%. Sin embargo, al llegar a diciembre de 2021, esta tasa había aumentado significativamente a un 14.6%. El incremento continuó en diciembre de 2022, cuando la tasa alcanzó un preocupante 16.8%. Estos datos reflejan un deterioro progresivo en la calidad de la cartera de consumo.

Simultáneamente, los datos históricos de pagos y morosidad de la cartera de consumo revelan una tendencia crítica. A medida que avanzamos desde julio de 2022 hasta enero de 2023, notamos un patrón de deterioro en la cartera del 4%. A pesar de que el número de clientes que realiza pagos se ha mantenido constante, el deterioro también es constante en el número de clientes que no han pagado o tienen una mora superior a 30 días.

La implementación de esta herramienta permitirá una gestión más precisa y proactiva del riesgo crediticio, lo que resultará en una reducción sustancial de la morosidad. La identificación temprana de los clientes con mayor probabilidad de impago permitirá implementar medidas preventivas y personalizadas. Este enfoque basado en datos mejorará la calidad de la cartera de consumo fiduciario, también fortalecerá la rentabilidad del banco al reducir las pérdidas asociadas con la morosidad y gestionar de manera más efectiva y focalizada los gastos asociados con la gestión de la mora.

6.3 ALCANCE DE LA PROPUESTA

- Desarrollar y aplicar un modelo de machine learning preciso y confiable que sea capaz de predecir la probabilidad de impago de los clientes de la cartera de consumo fiduciario.
- Comparar y evaluar el rendimiento del modelo utilizando métricas como la precisión, el

área bajo la curva ROC y la sensibilidad. También se verifica la capacidad predictiva utilizando datos de validación que representen un período futuro en relación con el período de entrenamiento.

6.4 DESCRIPCIÓN Y DESARROLLO

6.4.1 DESCRIPCIÓN

Recopilación y tratamiento de los datos: En la fase inicial, se recolectarán datos cruciales de la cartera de consumo fiduciario, abarcando el período del 31 de mayo de 2021 al 30 de abril de 2022. Estos datos comprenderán información relevante como la edad de los clientes, la antigüedad de los préstamos, años de empleo, número de dependientes, saldos de préstamos, cuotas atrasadas y días de mora, intereses por mora, saldos vencidos entre otras variables. Se incluyen variables demográficas, financieras y transacciones.

Una vez ya definidas las variables a considerar, los datos son sometidos a limpieza y preprocesamiento. Esto incluye la identificación y gestión de valores atípicos, la imputación de datos faltantes o nulos, en caso necesario, la conversión de variables categóricas a numéricas. El objetivo principal es garantizar la consistencia de los datos para su utilización en el modelado.

Análisis de Datos Exploratorio: Se llevó a cabo un análisis de datos exploratorio para comprender mejor las características de la cartera de consumo fiduciario. Esto incluirá visualizaciones de datos y la identificación de patrones que puedan influir en el impago de los clientes.

6.4.1.1 VALIDACION CRUZADA DE LOS MODELOS PREDICTIVOS

La validación cruzada es una metodología crucial en la construcción y evaluación de modelos predictivos. Estamos trabajamos con datos secuenciales, como registros financieros a lo largo del tiempo, es esencial garantizar que nuestro modelo sea capaz de hacer predicciones precisas en situaciones futuras. Este enfoque de validación se divide en tres etapas fundamentales:

Entrenamiento del Modelo: Durante esta fase, el modelo aprende utilizando datos históricos que abarcan un período de tiempo específico. En nuestro estudio, hemos empleado registros desde el 31 de mayo de 2021 hasta el 30 de abril de 2022. Durante el entrenamiento, el modelo aprende patrones, relaciones y tendencias dentro de estos datos históricos. Esta etapa es

crucial, ya que sienta las bases para las predicciones futuras. Para llevar a cabo el entrenamiento, se utiliza el 70% de los datos disponibles, mientras que el 30% restante se reserva para su validación, garantizando así la robustez y eficacia del modelo.

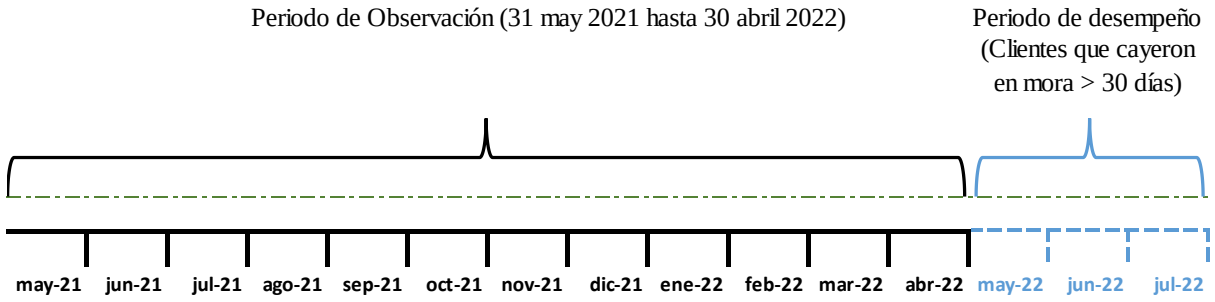


Figura 22: Periodo de tiempo analizado datos de entrenamiento

Fuentes: Elaboración propia

Prueba del Modelo: La prueba se realiza utilizando conjuntos de datos que representan un período futuro en relación con el período de entrenamiento. En nuestro caso, contamos con 3 conjuntos de validación: uno que abarca desde el 31 de agosto de 2021 hasta el 31 de julio de 2022, otro desde el 30 de noviembre de 2021 hasta el 31 de octubre de 2022 y el último que abarca desde 28 de febrero del 2022 hasta 31 de enero del 2023. Estos conjuntos de validación simulan situaciones futuras en las que el modelo debe realizar predicciones basadas en su conocimiento adquirido durante el entrenamiento.

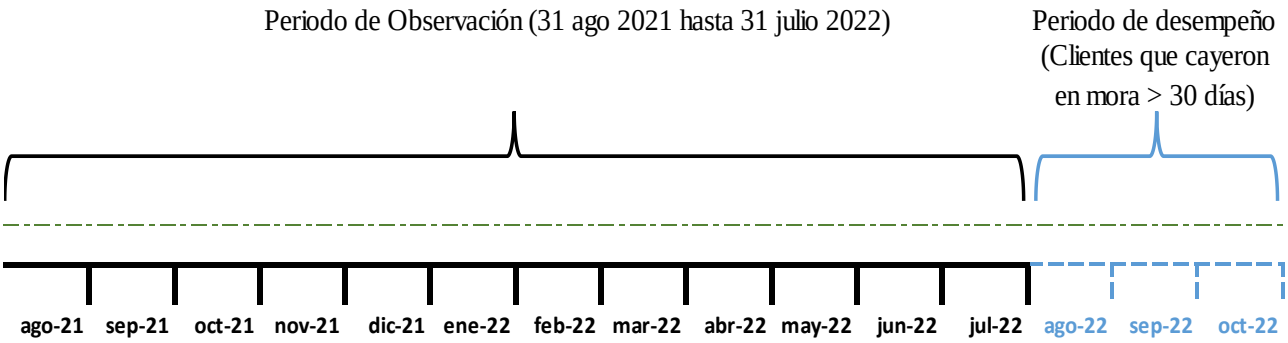


Figura 23: Periodo de tiempo analizado datos de validación

Fuentes: Elaboración propia

Evaluación del Modelo: La etapa final consiste en evaluar el rendimiento del modelo

mediante la comparación de sus predicciones con los resultados reales en los datos de validación. Medimos su capacidad de generalización en situaciones del mundo real mediante métricas como precisión, sensibilidad, especificidad y el área bajo la curva ROC. Esta evaluación nos proporciona información crucial sobre la confiabilidad y la precisión del modelo al hacer predicciones en contextos futuros.

La validación cruzada garantiza que nuestro modelo pueda realizar predicciones confiables en escenarios de la vida real, especialmente cuando se enfrenta a datos futuros para los cuales no tenemos información de resultados reales. Su aplicación es fundamental para demostrar la robustez y el valor de nuestro modelo en la toma de decisiones, la gestión de riesgos y la reducción de la morosidad en el sector financiero.

6.4.2 DESARROLLO

El proceso de desarrollo del modelo predictivo para evaluar la probabilidad de impago de los clientes que cuentan con un préstamo de consumo fiduciario implica una serie de pasos claves. Se presentará de manera resumida las herramientas usadas en cada etapa del proceso. La herramienta principal para la extracción y procesamiento de los datos es SQL, posteriormente se utilizó KNIME para el desarrollo del modelo de Machine Learning y Power BI se implementó para visualizar los resultados de manera efectiva.

6.4.2.1 EXTRACCION Y TRANSFORMACIÓN DE LAS VARIABLES INDEPENDIENTES

En este proceso se utilizó la herramienta SQL para extraer y manipular las variables claves necesarias para el desarrollo del modelo, estas variables incluyen datos históricos y esenciales del comportamiento del cliente como ser los días mora, intereses mora, intereses pagados por la deuda, principal vencido (mora al cambio), entre otras variables financieras.

Para mejorar la capacidad del modelo para capturar patrones y tendencias a lo largo del tiempo, se realizó una transformación significativa en estas variables. En lugar de trabajar con valores individuales, se crearon trimestres de estas variables. Por ejemplo, la variable "Intereses LPS" se transformó en "Trim1 Intereses Lps", "Trim2 Intereses Lps”, "Trim3 Intereses Lps" y "Trim4 Intereses Lps", representando así los cuatro trimestres del año. Esta transformación permitió una visión más completa de cómo estas variables evolucionaron a lo largo del año.

Se aplicaron filtros que garantizan que los clientes seleccionados no tuvieran más de 15 días de mora al cierre de abril de 2022, también se implementó en las corridas con datos futuros para asegurar que los clientes seleccionados para el análisis no tuvieran más de 15 días de mora al final de cada período de observación. Este filtro de no tener más de 15 días de mora al cierre de abril de 2022 desempeñó un papel fundamental en la selección de los datos utilizados para el modelo. Su propósito radica en enfocar el análisis en aquellos clientes que aún no habían caído en mora significativa al final del período histórico. Este filtro permitió identificar a los clientes que podrían estar en riesgo de impago en un futuro cercano, excluyendo a aquellos que ya habían acumulado un historial prolongado de morosidad. Otros filtros aplicados la antigüedad del préstamo mayor o igual a 12 meses y que el cliente no sea empleado del Banco.

Estos pasos se resumen en una consulta que estandariza la preparación de datos. Esto significa que, para cada corrida, solo se ajusta el período de observación, evitando la repetición de procesos manuales de limpieza y filtrado. Esto ahorra tiempo y permite ejecutar el modelo con diferentes conjuntos de datos históricos o futuros al modificar el rango de fechas.

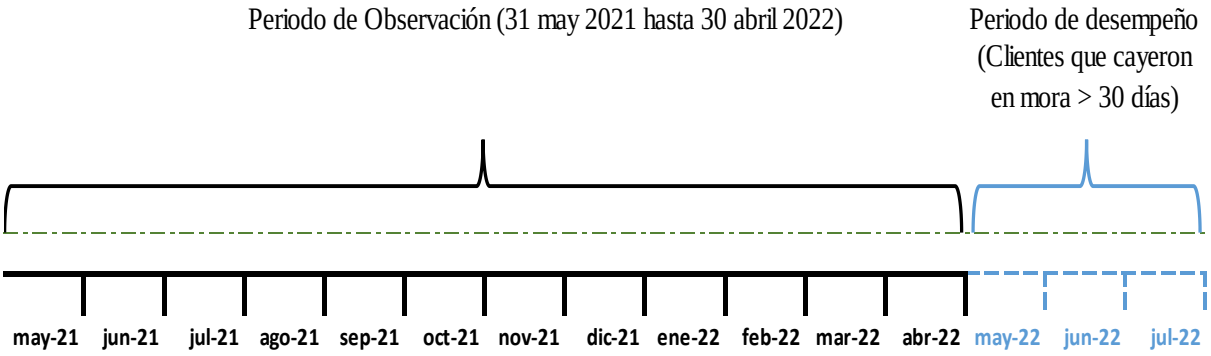
Tabla 9 Selección de variables independiente para el modelo

Variable	Variable Transformada	Tipo	Explicación
Años Empleo	N/A	Continua	Años de empleo del cliente en su empleador actual
Saldo cons-Fid.	N/A	Continua	Saldo en cuentas de consumo o fiduciarias del cliente
Tasa Cons-Fid.	N/A	Continua	Representa la tasa de interes del prétamo
Dias mora al cierre permitidos	N/A	Continua	Días mora del cliente al cierre de cada periodo observado (menor 15 días)
dias mora	Días mora trimestral promedio (ejemplo: trim-1 dias mora prom.)	Continua	Representa los días mora promedio del cliente en forma trimestral
Saldo en mora	Saldo en mora trimestral promedio (ejemplo: trim-1 saldo mora prom.)	Continua	Monto de la cuota que el deudor no ha pagado en la fecha de vencimiento acordada
Interes vencido	Intereses vencidos trimestral promedio (ejemplo: Trim-1 Interes vencido prom)	Continua	Intereses de la cuota que el deudor no ha pagado en la fecha de vencimiento acordada
Total días Mora Año Consumo Fid.	N/A	Continua	Total de días de mora acumulados en el año para la cartera de consumo fiduciario
Conteo meses mora = 0 con-fid.	N/A	Categórica	Número de meses que el cliente estuvo sin tener al menos un día en mora
Conteo meses mora entre 1-30 días con-fid.	N/A	Continua	Número de meses en mora con un nivel de 1 a 30 días para la cartera de consumo fiduciario
Conteo meses en mora consumo-fid	N/A	Continua	Número total de meses en mora para la cartera de consumo fiduciario
Cuotas vencidas acum.	N/A	Continua	Cantidad acumulada de cuotas vencidas
Pago capital promed año	N/A	Continua	Promedio anual de pagos realizados hacia el capital
Cuota_lps_promed_año	N/A	Continua	Promedio anual de cuotas pagadas en la moneda local (LPS)

Fuentes: Elaboración propia

6.4.2.2 CONTRUCCIÓN DE LA VARIABLE DEPENDIENTE

Para construir nuestra variable “Estado de Pago”, se evaluó el comportamiento de estos clientes seleccionados en un período de desempeño adicional de 3 meses. Durante este lapso, se monitoreó si los clientes experimentaron una mora mayor a 30 días. Si no la tenían, se consideró que estaban al día con sus pagos y se etiquetaron como 0 (Pago). En caso de que incurrieran en una mora mayor a 30 días, se catalogaron como 1 (No Pago). Esta definición de la variable dependiente se basó en el objetivo de evaluar la probabilidad de impago de los clientes y su idoneidad para la gestión de riesgos crediticios.



Volvemos a nuestra figura 22, en donde observamos un periodo de desempeño, que es utilizado para evaluar a los clientes seleccionados, si la mora del cliente al final del periodo de desempeño es mayor de 30 días, entonces se marca como 1, de lo contrario significa que el cliente no cayó en mora mayor a 30 día y corresponde marcarlo como 0, de esta forma se construye nuestra variable dependiente “Estado de Pago”

6.4.2.3 ENTRENAMIENTO

Una vez que hemos preparado los datos mediante la ejecución de la consulta, que contiene todo el proceso de extracción, transformación y limpieza, el siguiente paso es el entrenamiento del modelo, el cual se llevara a cabo en la herramienta KNIME, donde configuramos y entrenamos nuestro modelo de machine learning con el conjunto de datos preparados.

- Carga de Datos: En primer lugar, se carga el archivo que contiene el conjunto de datos necesario para el entrenamiento del modelo.

- Manejo de Valores Nulos: Se realiza una corrección de los valores nulos o missing values, asignándoles un valor de cero (0) para garantizar la integridad de los datos.
- Filtro de Empleados: Se aplica un filtro para eliminar a los clientes que son empleados del banco, ya que su comportamiento podría ser diferente al de los clientes regulares.
- Normalización de Datos: Los datos se normalizan para llevarlos a una escala común y facilitar el proceso de entrenamiento del modelo.
- Partición de Datos: Se divide el conjunto de datos en dos partes: un 70% se utiliza para entrenar el modelo y un 30% se reserva para validar su desempeño.
- Configuración del Modelo: Se configuran los nodos y parámetros necesarios para crear y entrenar el modelo de machine learning en Knime.
- Nodos adicionales: empleamos nodos adicionales en Knime para optimizar el proceso. Utilizamos el nodo ROWID para asignar identificadores únicos a los datos antes de su procesamiento, y posteriormente, el nodo JOINER se utiliza para reintroducir estos identificadores en el conjunto de datos, facilitando la identificación de predicciones con los registros originales de los clientes. Además, aplicamos el nodo Desnormalizer para revertir la normalización de datos realizada anteriormente, lo que nos permite obtener los resultados en su forma original y comprender mejor su significado.

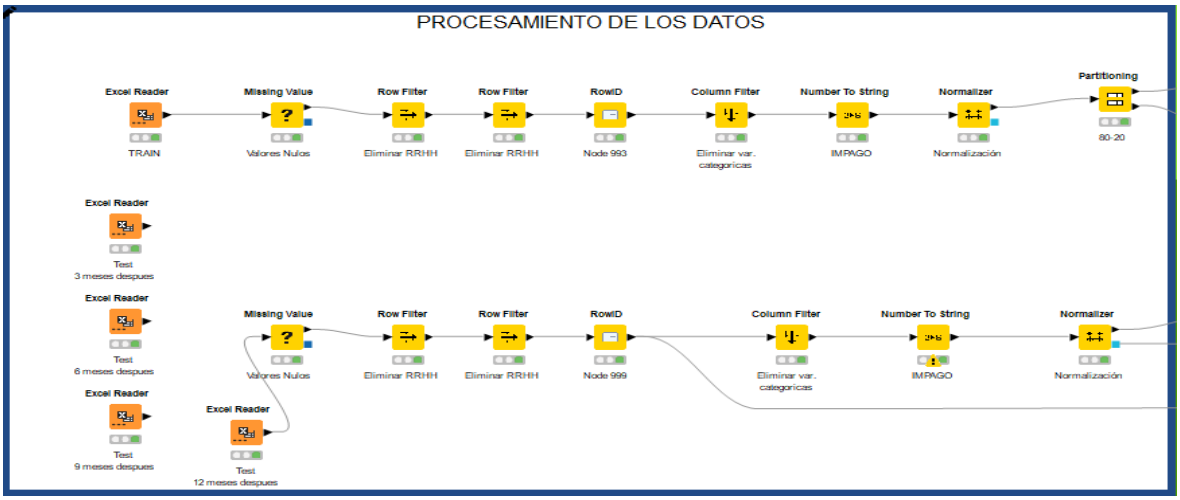


Figura 24: Carga y procesamiento de los datos Knime

Fuentes: Elaboración propia

Figura 25: Entrenamiento del modelo

Fuentes: Elaboración propia

6.4.2.4 VALIDACIÓN

La validación se lleva a cabo utilizando tres conjuntos de datos que representan períodos futuros con respecto al período de entrenamiento. Estos conjuntos abarcan desde agosto de 2021 hasta julio de 2022, desde noviembre de 2021 hasta octubre de 2022, y desde febrero de 2022 hasta enero de 2023. Estos escenarios de validación permiten evaluar la capacidad del modelo para hacer predicciones en situaciones futuras basadas en su entrenamiento previo.

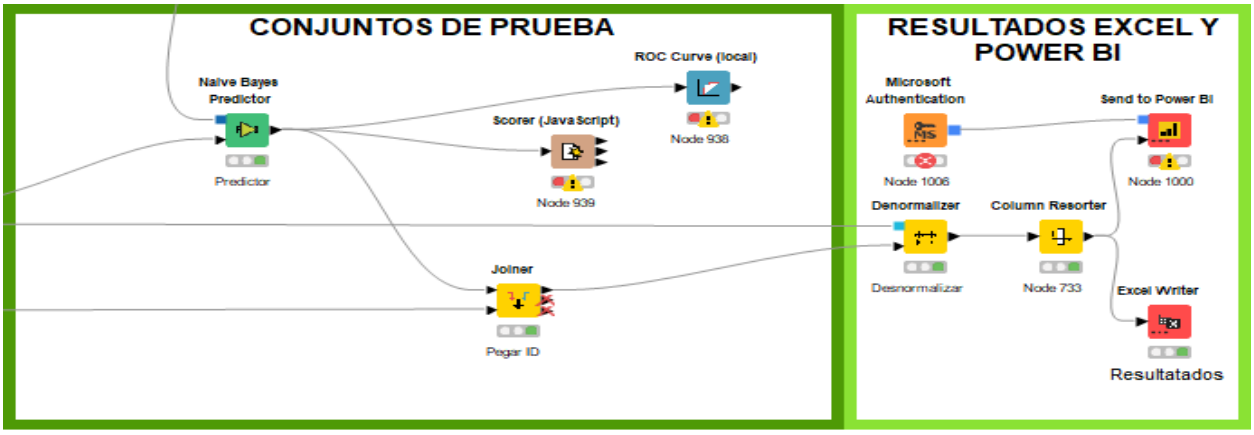


Figura 26: Validación cruzada

Para almacenar los resultados de las predicciones, empleamos el nodo Excel Writer en Knime. Este nodo permite guardar un archivo en formato Excel que contiene información detallada sobre las predicciones realizadas por el modelo.

6.5 MEDIDAS DE CONTROL

6.5.1.1 METRICAS CLAVES DE VALIDACION

Con el fin de mejorar la gestión del riesgo crediticio y reducir la morosidad en la cartera de consumo fiduciario, se ha desarrollado un análisis exhaustivo utilizando diferentes algoritmos como herramientas centrales. Este análisis se llevó a cabo en diferentes momentos posteriores a un período de tiempo específico, con el propósito de evaluar la eficacia de nuestro modelo de predicción en la identificación de posibles incumplimientos crediticios.

Cada prueba representa un escenario realista en el que el modelo fue aplicado para prever el comportamiento de los clientes. Estos resultados se han organizado de manera clara en la tabla 10, que incluye métricas clave como la exactitud, la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC (ROC Auc).

Tabla 10: Evaluación de desempeños de los modelos

Algoritmo	Pruebas	Exactitud	Precisión	Especificidad	Sensibilidad	ROC Auc
Naive Bayes	**Prueba (30%)**	**87.05%**	**96.64%**	**89.15%**	**60.00%**	**77.83%**
	Prueba 1 (3 meses déspues)	**82.61%**	**97.79%**	**83.71%**	**53.33%**	**76.07%**
	Prueba 2 (6 meses déspues)	**83.33%**	**98.40%**	**84.11%**	**61.54%**	**77.28%**
	Prueba 3 (9 meses déspues)	**83.48%**	**98.19%**	**84.21%**	**68.75%**	**82.37%**
Regresión Logística	Prueba (30%)	92.55%	97.64%	94.43%	57.14%	76.19%
	Prueba 1 (3 meses déspues)	87.68%	97.28%	89.72%	33.33%	72.48%
	Prueba 2 (6 meses déspues)	85.45%	97.55%	87.12%	38.46%	66.95%
	Prueba 3 (9 meses déspues)	87.32%	96.05%	90.40%	25.00%	72.58%
SVM	Prueba (30%)	75.00%	95.40%	77.75%	6.25%	68.75%
	Prueba 1 (3 meses déspues)	68.12%	96.84%	69.17%	40.00%	64.37%
	Prueba 2 (6 meses déspues)	68.52%	96.59%	69.86%	30.77%	55.12%
	Prueba 3 (9 meses déspues)	66.08%	96.02%	67.18%	43.75%	54.50%

En cada una de estas pruebas realizadas, que representa un escenario realista en el que el modelo se aplicó a datos futuros para evaluar su desempeño en situaciones diversas. Es importante destacar que las pruebas se llevaron a cabo en tres momentos diferentes, cada una con un conjunto de datos correspondiente a 3, 6 y 9 meses posteriores a la fecha de observación original. Este enfoque permite una evaluación exhaustiva del modelo en condiciones que simulan con precisión el comportamiento de los clientes a lo largo del tiempo. Naive Bayes ha demostrado ser la opción más equilibrada en términos de precisión y especificidad, lo que lo posiciona como la elección más confiable en la gestión del riesgo crediticio y la reducción de la morosidad en la institución financiera.

6.5.1.2 HERRAMIENTA DE VISUALIZACION: DASHBOARD DE SEGUIMIENTO DE CLIENTES

Después de haber realizado un análisis y evaluación de nuestro modelo de predicción de riesgo crediticio, llegamos al momento de la presentación de los resultados. Se desarrollo un tablero de seguimiento que refleja los resultados de nuestro modelo de predicción, donde se muestran las características clave y se analizan tendencias, identificaremos riesgos potenciales y demostraremos cómo nuestro modelo de aprendizaje automático impulsa la toma de decisiones eficaces en nuestra institución financiera.

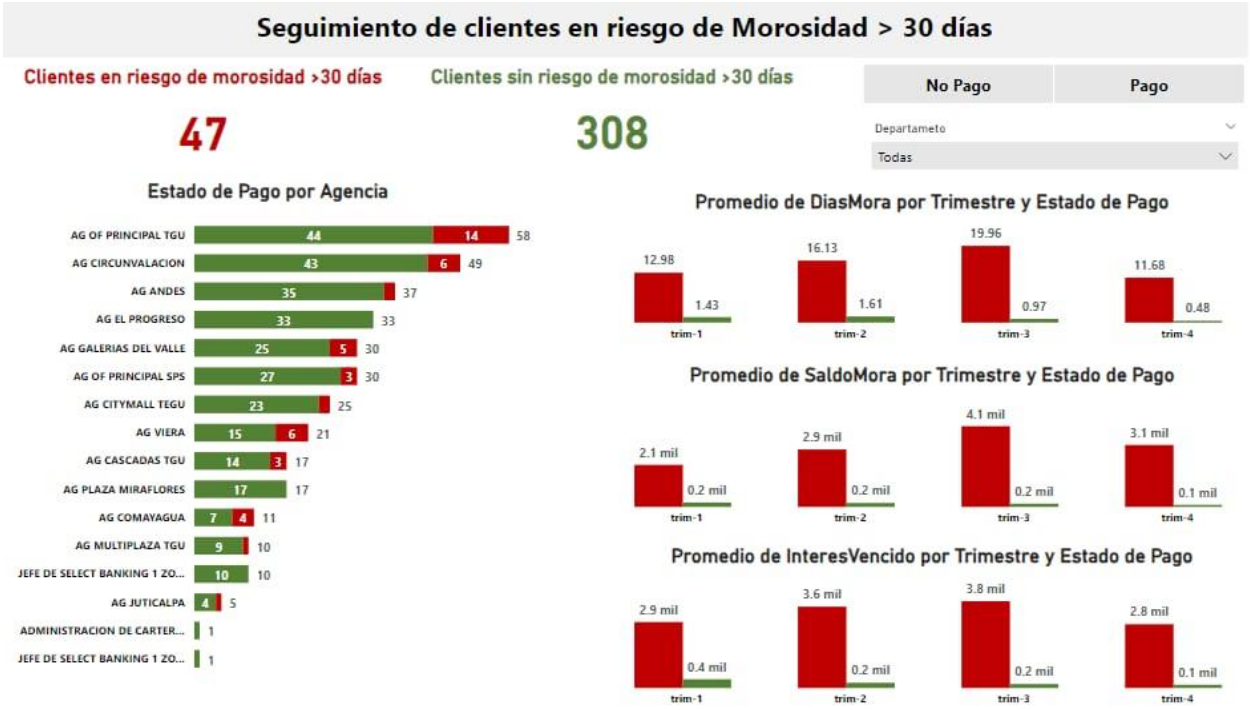


Figura 27: Seguimiento de clientes en riesgo de morosidad > 30 días

6.5.1.3 FRECUENCIA DE ACTUALIZACIÓN

La frecuencia de actualización trimestral de nuestros datos se ha seleccionado estratégicamente para brindar una visión eficaz de la evolución del riesgo crediticio. En un período de 12 meses de análisis, evaluamos el historial de pagos de los clientes e identificamos aquellos con mayor riesgo de caer en mora durante los próximos 3 meses. Es importante destacar que solo se incluyen clientes con mora menor a 15 días al final del período de observación.

6.6 CRONOGRAMA DE IMPLEMENTACIÓN Y PRESUPUESTO

Tabla 11: Cronograma de implementación y presupuesto

Actividad a desarrollar	Inversión Lps.	Semana 1	2	3	4	5	6	7
Capacitación en el uso SQL (básico)	8,000							
Capacitación en el uso de KNIME	8,000							
Capacitación en el uso de Power BI	8,000							
Adquisición de la licencia de Power BI	3,000							
Evaluación y revisión del proyecto	5,000							

Está programado que se lleven a cabo una serie de actividades clave para la implementación exitosa del proyecto. Esto incluye la capacitación en el uso de SQL nivel básico, KNIME y Power BI, así como la adquisición de la licencia necesaria para Power BI (1 año). Estas acciones nos permitirán finalizar la fase de implementación, lo único pendiente será la designación de la persona encargada y su posterior capacitación en el manejo de estas herramientas. Con estas actividades programadas y una inversión de 32,000 lempiras, se llevaría a cabo la implementación del proyecto.

Figura 28: Esquema de Implementación del proyecto Predictivo

6.7 CONCORDANCIA DE LOS SEGMENTOS DE LA TESIS CON LA PROPUESTA

La matriz concordancia de los segmentos de la tesis con la propuesta es esencial para evaluar cómo nuestra investigación se ajusta a los objetivos y enfoques establecidos en la propuesta inicial. Analizaremos cómo cada parte de la tesis se relaciona con nuestra visión original, asegurando coherencia y relevancia.

Tabla 12: Matriz de concordancia de los segmentos de la tesis con la propuesta

Capítulo I			Capítulo II	Capítulo III	Capítulo V	Capítulo VI	
Titulo Investigación	Objetivo General	Objetivos Específicos	Teorías de Sustento	Conclusiones		Nombre Propuesta	Objetivos Propuesta
Predicción de riesgo de impago en institución financiera usando modelos de machine learning	Identificar y analizar las variables más relevantes que influyen en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario, con el propósito de	Evaluar y seleccionar las variables más influyentes en la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario mediante el uso	Machine learning y su relevancia en el riesgo crediticio	Nuestra investigación demuestra que el comportamiento de pago de los clientes es fundamental para predecir el incumplimiento de pagos. Dentro de las variables analizadas, encontramos que el número promedio de cuotas vencidas por año, el historial de morosidad en términos de días de mora, el saldo en mora, los intereses vencidos entre otras variables		Implementación de un modelo de Machine Learning para predecir la probabilidad de impago de la cartera de cliente de consumo fiduciario.	• Desarrollar y aplicar un modelo de machine learning preciso y confiable que sea capaz de predecir la probabilidad de impago de

	mejorar la gestión del riesgo crediticio y reducir la morosidad del Banco	de técnicas de análisis de datos y modelos de machine learning.		son indicadores fundamentales para identificar a los clientes que podrían estar en mayor riesgo de caer en mora. Estos indicadores son esenciales para la toma de decisiones estratégicas en la gestión de riesgos crediticios y desempeñan un papel fundamental en la evaluación y mitigación de riesgos financieros.		los clientes de la cartera de consumo fiduciario. • Comparar y evaluar el rendimiento del modelo utilizando métricas como la precisión, el área bajo la curva ROC y la sensibilidad. También se verifica la capacidad predictiva utilizando datos de validación que
		Desarrollar y aplicar un modelo de machine learning precios y confiable que sea capaz de predecir la probabilidad de impago de los clientes de la	Técnicas de aprendizajes automáticos (ML) y metodologías de reprocesamiento de datos aplicadas	Basado en el análisis gráfico y la prueba chi-cuadrado se ha constatado que la mayoría de estas variables demográficas como ser agencia, zona, departamento, nivel educativo, ingresos y estado civil entre otras, no presentan una asociación estadísticamente significativa con la variable 'Estado de Pago' en nuestra		

		cartera de consumo fiduciario, con el propósito de mejorar la gestión del riesgo crediticia y reducir la morosidad del Banco.		cartera de consumo fiduciario. Sin embargo, cabe resaltar la variable género, donde hemos identificado una relación estadísticamente significativa. Es esencial la tasa de incumplimiento es más alta entre los clientes masculinos, pero también representan una proporción mayor de préstamos otorgados en comparación con las clientes femeninas.		representen un período futuro en relación con el período de entrenamiento.
		Comparar y evaluar el rendimiento de diferentes modelos de ML, utilizando métricas como la precisión, el área bajo la curva ROC y la	Técnicas de aprendizajes automáticos (ML) y metodologías de reprocesamiento de datos aplicadas	Durante esta investigación, hemos analizado exhaustivamente las variables relacionadas con la probabilidad de incumplimiento de los clientes de la cartera de consumo fiduciario en una entidad financiera. Se utilizaron diversas técnicas de análisis de datos y modelos de machine		

		sensibilidad, para determinar el modelo que mejor desempeño tiene.		learning, evaluamos y seleccionamos las variables más influyentes para predecir el impago de los clientes.		
		Desarrollar un dashboard diseñado para visualizar y supervisar de manera dinámica a los clientes con alta probabilidad de caer en mora en la cartera de consumo fiduciaria.	Técnicas de aprendizajes automáticos (ML) y metodologías de reprocesamiento de datos aplicadas	Con los resultados obtenidos por los modelos ML aplicados, se considera la creación de un dashboard diseñado específicamente para visualizar y supervisar de manera dinámica a los clientes con alta probabilidad de caer en mora en la cartera de consumo fiduciaria es una medida estratégica fundamental para la institución financiera. Esta herramienta proporciona una visión más clara y actualizada de los riesgos asociados a los préstamos y permite tomar decisiones		

				informadas de manera proactiva. Al utilizar datos en tiempo real y análisis predictivos, el dashboard facilita la identificación temprana de clientes en riesgo, lo que a su vez permite implementar acciones preventivas y estratégicas para reducir la morosidad y proteger los activos financieros.		

Fuentes: Elaboración propia

6.8 BENEFICIOS Y CONCLUSIONES

6.8.1.1 BENEFICIOS

Mejora en la Gestión del Riesgo Crediticio: La implementación exitosa de nuestro modelo de predicción de riesgo crediticio se traduce en una gestión del riesgo más efectiva. Esto permite a las instituciones financieras identificar a los clientes con mayor probabilidad de impago y tomar medidas preventivas de manera oportuna.

Optimización de Recursos: Una de las ventajas clave de nuestro proyecto es la optimización de recursos. Nuestro modelo brindara la capacidad de asignar los recursos de manera más eficiente al enfocarnos en los clientes con riesgo crediticio elevado.

6.8.1.2 CONCLUSIONES

- Se desarrollo un modelo de machine learning de alta precisión que puede predecir con éxito el 60% de los clientes que enfrentarán impagos. La capacidad de anticipar y prevenir incumplimientos de pago no solo permitirá una administración más eficiente de los recursos, sino que también contribuirá a la estabilidad financiera y al crecimiento sostenible.

- Se valido el modelo con métricas como la exactitud 84%, precisión 97%, el área bajo la curva ROC 78% y la sensibilidad 61%, además de validaciones cruzada con datos de varios períodos futuros, se desprende una conclusión: el modelo de machine learning exhibe una capacidad sólida y confiable para prever el riesgo de impago entre los clientes de la cartera de consumo fiduciario. Estos resultados respaldan de manera contundente la efectividad de nuestro modelo en la gestión del riesgo crediticio y su valor para la toma de decisiones fundamentadas en el ámbito financiero.

REFERENCIAS BIBLIOGRÁFICAS

Bobadilla, J. (2020). Machine Learning y Deep Learning: Usando Python, Scikit y Keras. RA-MA Editorial. https://elibro.net/es/ereader/unitechn/222698?page=10

Campo León, E., & Alcalá Nalvaiz, J. T. (2016). Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervisado. Universidad de Zaragoza.

García, M. L. S., & García, M. J. S. (2010). Modelos para medir el riesgo de crédito de la banca. Cuadernos de Administración, 23(40), Article 40. https://doi.org/10.11144/Javeriana.cao23-40.mpmr

Gómez Fernández-Aguado, P., & Partal Ureña, A. (2010). Gestión y control del riesgo de crédito en la banca. Delta Publicaciones. https://elibro.net/es/ereader/unitechn/169698

Grau Álvarez, J. (2020). Machine Learning y riesgo de crédito. https://repositorio.comillas.edu/xmlui/handle/11531/39062

Ledesma Martinez, Z. M., & Sanchez Machado, I. R. (2007). Análisis del riesgo crediticio bancario en la economía cubana. Teoría y Praxis: turismo, negocios, recursos naturales. http://risisbi.uqroo.mx/handle/20.500.12249/782

Morales Castro, A., & Morales Castro, J. A. (2015). Crédito y cobranza. Grupo Editorial Patria. https://elibro.net/es/ereader/unitechn/39380

Ossa Giraldo, W., & Jaramillo Marin, V. (2021). Machine Learning para la estimación del riesgo de crédito en una cartera de consumo [masterThesis, Universidad EAFIT]. http://repository.eafit.edu.co/handle/10784/29589

Pineda Pertuz, C. (2022). Aprendizaje automático y profundo en Python: Una mirada hacia la inteligencia artificial. https://elibro.net/es/ereader/unitechn/230579

Riesgo, C. de R. | S. (s. f.). Calificadora de Riesgo | SC Riesgo. Recuperado 20 de agosto de

2023, de https://scriesgo.com/Publication/detail/373/informe-sectorial-de-seguros-en-centroamerica-diciembre-2021

Saunders, A., & Allen, L. (2010). Credit Risk Management In and Out of the Financial Crisis: New Approaches to Value at Risk and Other Paradigms.

Soules, L. M. (2020). Modelos predictivos competitivos de morosidad crediticia para entidades argentinas Análisis descriptivo y predictivo con datos públicos. https://repositorio.utdt.edu/handle/20.500.13098/11288

Valle Carrascal, J. M. (2017). Modelos de medición del riesgo de crédito. https://hdl.handle.net/20.500.14352/21643

ANEXOS

ANEXO 1: CARTA DE AUTORIZACIÓN DE LA EMPRESA O INSTITUCIÓN

CARTA DE AUTORIZACIÓN DE LA EMPRESA O INSTITUCIÓN

Tegucigalpa, Francisco Morazán, 21/09/2023
(Cuidad), (Departamento) (Día, mes y año)

Kenia Mariela Velasquez Menendez
(Nombre y apellidos del director o Gerente)

Gerente de Recursos Humanos
(Puesto Laboral)

Banco Ficensa
(Empresa o Institución)

Gerencia de Recursos Humanos
(Dirección principal de la empresa o institución)

Estimado Señor(a): Licencia Kenia Velasquez
Reciba un cordial y atento saludo. Por medio de la presente deseamos solicitar su apoyo, dado que somos alumnos de UNITEC y nos encontramos desarrollando el Trabajo Final de Graduación previo a obtener nuestro título de maestría en Analítica de Negocios.
Hemos seleccionado como tema: Predicción del riesgo de Impago en Institución Financiero usando Machine Leraning, por lo que estaríamos muy agradecidos de contar con el apoyo de la empresa que usted representa para poder desarrollar nuestra investigación. En particular, dicha solicitud se circunscribe a peticionar que se nos autorice a realizar: a hacer uso de la información proporcionados por su institución financiera con el fin de llevar a cabo nuestro estudio de predicción de riesgo de impago.
A la espera de su aprobación, me suscribo de Usted.
Atentamente,

José Manuel García
Firma, nombre y apellidos Firma, nombre y apellidos

Walter Torres
Firma, nombre y apellidos Firma, nombre y apellidos

No. de cuenta: 11623063

No. de cuenta: 12213071

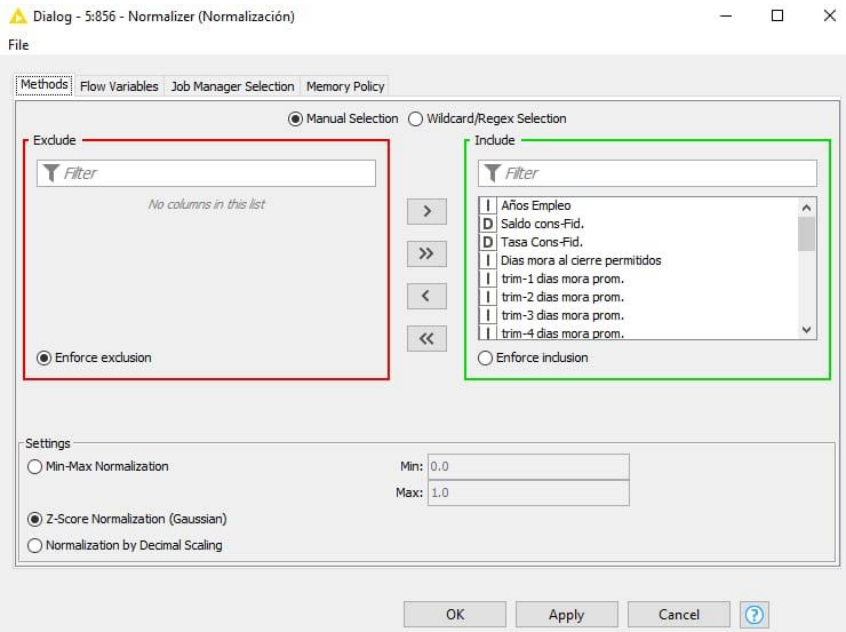
Por este medio, Banco Ficensa
(empresa / institución),
Autoriza la realización dentro de sus instalaciones el proyecto de investigación de Postgrado antes mencionado.

(Nombre y sello del Director / Gerente)

Vo.Bo.

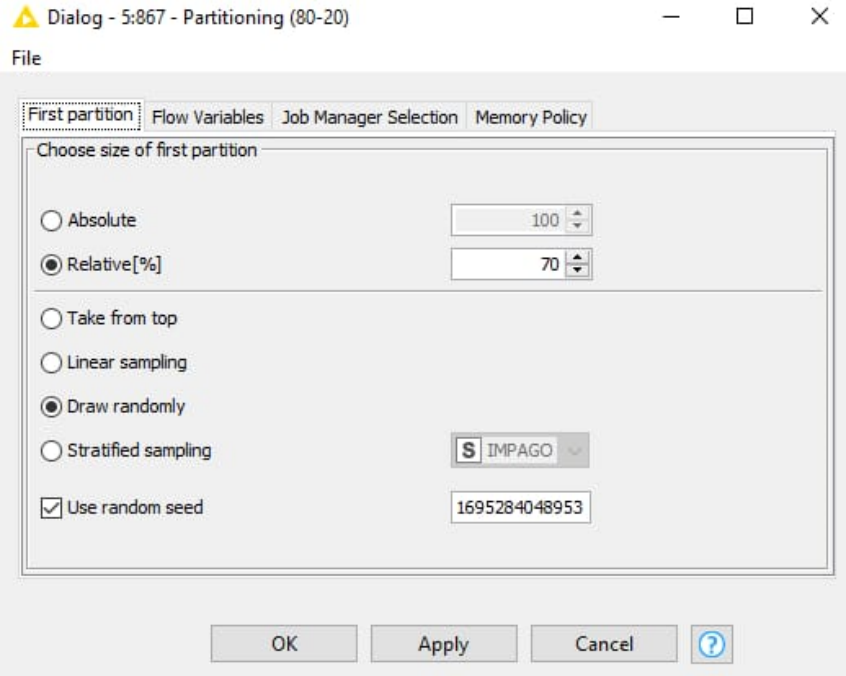
Fuente: Elaboración propia

ANEXO 2: CONFIGURACIÓN NODO NORMALIZACIÓN



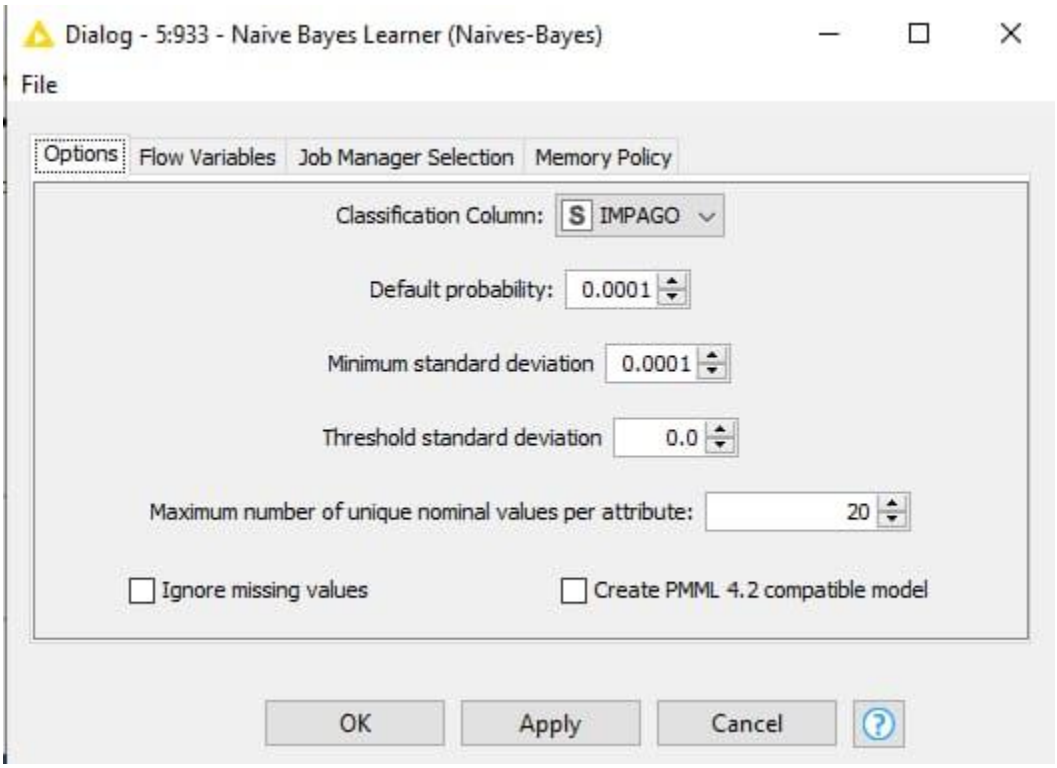
Fuente: Elaboración propia

ANEXO 3: CONFIGURACIÓN NODO PARTICIPACIÓN



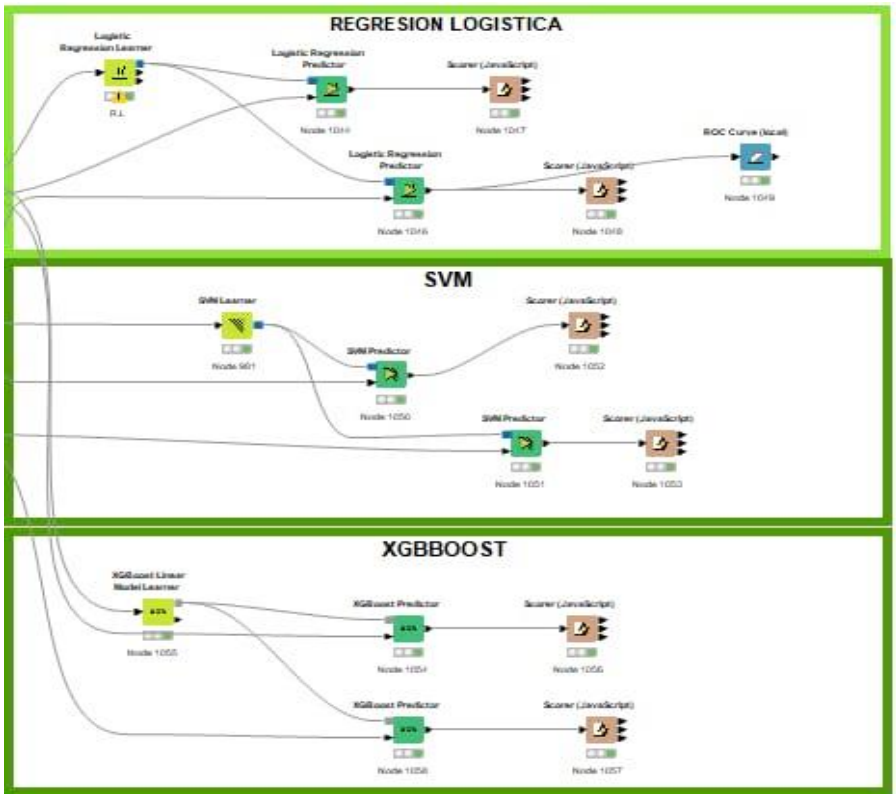
Fuente: Elaboración propia

ANEXO 4: CONFIGURACIÓN NODO NAIVES BAYES



Fuente: Elaboración propia

ANEXO 5: CONFIGURACIÓN NODO NAIVES BAYES



Fuente: Elaboración propia

ANEXO 6: DASHBOARD DETALLE CLIENTES EN RIESGO DE MOROSIDAD

Fuente: Elaboración propia