ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

MODELOS PREDICTIVOS PARA LA GESTIÓN DE RIESGO DE CRÉDITO: APLICACIÓN DE RANDOM FOREST Y ÁRBOLES DE DECISIÓN

Profesor

María Salvador Gonzales Rodríguez Ph.D

Autor

Rosa Estefanía Zapata Yanez

2024

Resumen

En la actualidad, el desarrollo del sistema financiero es fundamental para impulsar el crecimiento económico, esto a su vez hace que la gestión del riesgo de crédito es esencial para asegurar la estabilidad de cualquier institución cuya actividad principal sea la colocación de créditos. Sin embargo, la complejidad y la diversidad de los factores que pueden afectar cumplimiento de los pagos por parte de los clientes requieren el desarrollo de metodologías de aprendizaje automático para discernir el comportamiento de pago de los clientes dentro del sistema financiero. Bajo este contexto el presente proyecto propone desarrollar metodologías que permitan una identificación temprana y precisa de los clientes en riesgo de morosidad. La metodología de árboles de decisión ofrece una solución interpretable y fácilmente comprensible, mientras que Random Forest proporciona un enfoque más robusto y escalable mediante la combinación de varios árboles de decisión. A través de la aplicación de estas metodologías y la utilización de variables financieras y de comportamiento del cliente se buscar proponer predicciones que permitan mejorar la capacidad de tomar decisiones informadas, se espera que los resultados obtenidos contribuyan al desarrollo de estrategias más eficaces y personalizadas para la prevención y mitigación del riesgo de crédito, promoviendo así una gestión más proactiva y eficiente de los riesgos financieros en el sector.

Abstract

Currently, the development of the financial system is essential to boost economic growth, this in turn makes credit risk management essential to ensure the stability of any institution whose main activity is the placement of credits. However, the complexity and diversity of factors that can affect customer payment compliance require the development of machine learning methodologies to discern customer payment behavior within the financial system. Under this context, this project proposes to develop methodologies that allow early and accurate identification of clients at risk of default. The decision tree methodology offers an interpretable and easily understandable solution, while Random Forest provides a more robust and scalable approach by combining several decision trees. Through applying these methodologies and using financial and behavioral variables of the client, proposals are sought that improve the ability to make informed decisions. It is expected that the results obtained will contribute to the development of more effective and personalized strategies for prevention. and credit risk mitigation, thus promoting more proactive and efficient management of financial risks in the sector.

Índice de contenido

Índice de Tablas

Índice de Figuras

Introducción

En la actual era de globalización financiera diferentes sectores de la economía se han desarrollado gracias al crecimiento exponencial que han presentado los servicios financieros, Samaniego & Rodríguez (2015) manifiestan la importancia de la contribución del sector financiero dentro del crecimiento económico, dado que lo préstamos otorgados en diferentes líneas de negocio son fundamentales para estimular de esta formar la inversión y el consumo dentro de la economía, y esto a su vez impulsa el crecimiento económico del país.

De acuerdo con el Banco Mundial los créditos colocados dentro del sector privado en el 2019 representan el 132% del (PIB) producto interno bruto a nivel mundial, para lo que corresponde a Latinoamérica y el Caribe representan el 55,6%, mientras que en Ecuador el 42,7%. (Freire Lopez, 2021)

Bajo este contexto la actividad crediticia juega importante dentro de cualquier economía, sin embargo, es crucial gestionar las operaciones bancarias de manera que las instituciones administren sus niveles de riesgo y de rentabilidad en el caso de incumplimientos en los pagos por parte de los prestamistas. De hecho, se puede encontrar gran cantidad de literatura sobre los determinantes clave dentro del riesgo de crédito e incumplimientos dentro de una organización, permitiendo capturar ciertas diferencias entre sistemas bancarios, comportamientos del usuario, entre otros. (Boudriga, Taktak, & Jellouli, 2010)

En este sentido, el presente proyecto busca analizar la relación que tiene ciertas variables dentro del comportamiento de pago y cumplimiento de una operación dentro del sistema financiero, es importante conocer cuáles son las características cruciales para determinar un buen o mal comportamiento de pago de un cliente. Los resultados buscan identificar ciertos patrones de clientes que permitan minimizar el riesgo dentro de una organización.

El proyecto se estructura de cinco capítulos en total, en el primer capítulo se presenta el desarrollo de la problemática en estudio, el contexto e importancia del

riesgo de crédito dentro del sistema financiero, y la utilización de técnicas de Árboles de Decisión y Random Forest para abordar esta problemática. De igual manera, se presenta el objetivo general y los objetivos específicos del estudio.

En segundo capítulo, presenta la revisión literaria que existe sobre la problemática planteada, se revisa investigaciones realizadas y enfoques metodológicos utilizados en estudios similares.

El tercer capítulo por su parte describe de forma detallada la metodología utilizada dentro del proyecto, una descripción de variables y los procedimientos de análisis e implementación para Random Forest y Árboles de Decisión.

En el cuarto capítulo, se exponen los resultados a partir del análisis realizado, se analizan los hallazgos y resultados con relación a la pregunta de investigación planteada, y finalmente el quinto capítulo presenta las principales conclusiones del proyecto y recomendaciones para futuras investigaciones dentro del mismo campo.

1.1 Planteamiento del Problema

El desarrollo del sistema financiero es un factor clave y determinante para el crecimiento económico, así se tiene que el papel funcional de las instituciones financieras es uno de los roles más cruciales mediante la intermediación financiera, la cual es la principal actividad de las instituciones financieras, y que de acuerdo con sus características genera la mayor parte de beneficios y a su vez los mayores riesgos, dicha actividad está relaciona a un sin número de riesgos financieros como resultado del proceso de transformación de activos y pasivos, dentro de ello el riesgo principal es el riesgo de crédito el cual es inherente dentro de la administración de carteras de crédito. (Vargas Sánchez & Mostajo Castelú, 2014)

Dado el crecimiento de las operaciones crediticias en las instituciones financieras estás se ven obligadas a implementar tecnologías dentro de sus procesos principales de negocio, pues como se mencionó con anterioridad a medida que se incrementa la colocación de operaciones crediticias el riesgo de pérdidas financieras aumenta. La falta de una adecuada evaluación de riesgo de crédito y tecnologías o modelos que ayuden a medirlos lleva a una mayor exposición a operaciones de alto riesgo lo que resulta en mayores incumplimientos, un claro ejemplo fue el caso de créditos hipotecarios que en su momento condujeron a la crisis financiera mundial de 2008. (Mian & Sufi, 2009)

Se argumenta que la crisis financiera global fue el resultado de fracturas profundas en el sistema financiero que fueron pasadas por alto o subestimadas. Uno de los problemas principales fue la ausencia de una adecuada evaluación del riesgo de crédito. Las instituciones financieras, impulsadas por incentivos a corto plazo y una búsqueda excesiva de ganancias, ignoraron los signos de advertencia y otorgaron préstamos riesgosos sin considerar adecuadamente la capacidad de pago de los prestatarios, lo que llevó a un colapso del mercado hipotecario y una crisis financiera global. (Raghuram, 2010).

De manera general, la exposición a préstamos de alto riesgo sin evaluaciones, modelos o metodologías adecuadas puede tener consecuencias

devastadoras para la estabilidad financiera de cualquier institución y de la economía en conjunto, estas dificultades pueden atribuirse a la falta de personalización en los modelos existentes, que no consideran ciertas características individuales de los clientes ni su historial de pagos de manera integral.

Adicional tampoco se realiza un análisis de los clientes posterior al otorgamiento del crédito, bajo este contexto, se propone desarrollar un modelo de Random Forest y Árbol de Decisión, los cuales permitan obtener un análisis detallado del cliente, considerando factores como su comportamiento de pago, historial crediticio y perfil financiero, con el fin de segmentarlos de manera más precisa y mitigar el riesgo crediticio.

1.2 Identificación Objeto de Estudio

El presente proyecto busca analizar características de un cliente para determinar si este puede ser clasificado como buen o mal pagador, es decir evaluar un modelo predictivo basado en Random Forest y Árboles de Decisión para la correcta gestión de riesgo de crédito utilizando técnicas modernas de aprendizaje automático y herramientas estadísticas. La implementación de modelos dentro de la gestión del riesgo de crédito es crucial para una adecuada administración y planeación estratégica dentro de una institución de manera que permita controlar y mitigar adecuadamente los riesgos que generan las carteras de créditos. (Boudriga, Taktak, & Jellouli, 2010).

Para el desarrollo del presente proyecto se utiliza un dataset o base de datos de lo que simula ser una institución financiera la cual recoge información sobre las operaciones, e incluye variables relevantes como los saldos dentro del sistema financiero, pagos, indicadores de morosidad y ciertas características sociodemográficas del cliente. Para el presente la variable de interés es binaria y corresponde a establecer al cliente como buen pagador o mal pagador considerando si cayo en default por noventa días o más.

La finalidad del presente proyecto es entonces desarrollar un modelo que permita clasificar a los clientes actuales dependiendo de sus características y comportamiento de pago para optimizar la precisión y la confiabilidad en las estimaciones y predicciones de la probabilidad de incumplimiento de los clientes.

1.3 Objetivo General

Desarrollar un modelo de clasificación, utilizando técnicas avanzadas de aprendizaje automático, que permitan controlar y mitigar el riesgo de crédito. Este modelo busca optimizar la gestión estratégica y mejorar la toma de decisiones dentro de una organización, proporcionando una evaluación adecuada y confiable del riesgo asociado para cada cliente.

1.4 Objetivos Específicos

- Analizar los datos históricos para identificar patrones y tendencias en el comportamiento crediticio, utilizando técnicas de aprendizaje automático y minería de datos.
- Evaluar la precisión y eficacia del modelo utilizando métricas estándar como la sensibilidad, especificidad, y el área bajo la curva ROC.
- Comparar el rendimiento de cada uno de los modelos de Árbol de Decisión y Random Forest para determinar el mejor modelo que se ajusta con la calidad de los datos y objetivos propuestos.

Revisión Literaria

A nivel mundial las instituciones financieras de acuerdo con sus actividades se desenvuelven en un entorno de alta competitividad para mejorar su participación dentro del mercado y su entorno, de esta forma se ven forzadas a adoptar dentro de sus procesos medidas que permitan mejorar su rendimiento y eficiencia a la vez que puedan minimizar los riesgos generados por parte de esta actividad. Bajo este contexto se han desarrollado diversos enfoques y modelos de clasificación crediticia con metodologías de aprendizaje automático cuyo objetivo principal es establecer el riesgo que genera el otorgamiento de crédito, de manera que permita realizar una gestión proactiva de los mismos. (Liebergen, 2017)

La implementación de ciertas metodologías de clasificación crediticia que utilizan algoritmos de inteligencia artificial puede generar algunos beneficios a las instituciones dentro de los cuales se menciona el aumento de la tasa de retención de clientes, mejora de eficiencia y productividad, incremento en el margen de ganancias, y la posibilidad de incluir modelos analíticos en otras áreas (Freire Lopez, 2021). Por otro lado, la implementación de dichos modelos de clasificación ha permito obtener resultados más precisos dentro conjuntos de datos grandes, facilitando de esta forma el realizar una calificación de riesgo de crédito de cada uno de los clientes. Se menciona que las instituciones financieras que utilizan inteligencia artificial dentro de sus procesos pueden reducir entre el 20% a 25% de sus costos. (Lee & Shin, 2020).

Existen varios modelos de clasificación crediticia, pero el uso de técnicas de aprendizaje automático toma popularidad en la última década. Cohen, Krishnamoorthy, & Wrigh (2019) presentan una revisión exahustiva de diferentes aplicaciones de técnicas de aprendizaje automático en la evaluación de riesgo de crédito, entre los más relevantes se consideran las redes neuronales, árboles de decisión, máquinas de vectores de soporte y los ensambles de modelos. La ventaja principal de estas técnicas es que permiten capturar relaciones no lineales y facilitan el manejo de grandes volúmes de datos.

2.1 Random Forest dentro del riesgo de crédito

Para analizar el rendimiento de los modelos de predicción dentro del riesgo de crédito se han presentado varias investigaciones como las realizadas por Subasi and Cankurt (2019), Arora and Kaur (2019), Ziemba et al. (2020) y Pradhan et al. (2020), en las cuales utilizan una variedad de técnicas y algoritmos de aprendizaje automático. Los resultados presentados en estás investigaciones muestran que los modelos que implementan Random Forest presentan los mejores resultados en la predicción del riesgo crediticio. Random Forest o Árboles Aleatorios, es una técnica de aprendizaje automático que se basa en la combinación de varios árboles de decisión sin correlación, es conocido también como un método ensamblado, el cual combina todos los resultados generados de los diferentes árboles de manera que se obtenga un único valor de predicción para el conjunto total de árboles. Este algoritmo ha demostrado ser altamente efectivo en la captura de patrones complejos y en la mejora de la precisión predictiva en comparación con otros algoritmos. (Safari, 2020)

De manera general, estos hallazgos sugieren que el Random Forest puede ser considerado como una opción prometedora y robusta para la gestión del riesgo crediticio en diversas aplicaciones financieras. A continuación, se presentan detalladamente los resultados de los estudios mencionados, destacando la superioridad de los modelos basados en Random Forest y su relevancia en el escenario de la gestión del riesgo crediticio.

Tabla 1. Resultados de investigaciones con Random Forest

Investigación	Resultados
(Arora & Kaur, 2019)	97.90%
(Pradhan, Akter, & Al Marouf, 2020)	85.00%
(Subasi & Cankurt, 2019)	89.01%
(Ziemba, Radomska - Zalas, & Becker, 2020)	98.00%

Es así como las instituciones financieras a nivel mundial han incluido un sin número de modelos estadísticos para la clasificación de clientes dentro del riesgo de crédito las cuales tienen sus inicios en los años cincuenta

Pradhan et al (2020), por su parte realiza una comparación entre algunos algoritmos en un conjunto de clientes de crédito, con 4.600 observaciones y con se considero 47 variables, dentro de las metodologías utilizadas se encuentra redes neuronales, máquina de vectores de soporte, Árboles de Decisión y Random Forest, del estudio realizado el algoritmo aplicado de Random Forest o Bosque aleatorio presentó los mejores resultados, mostrando una precisión del modelo del 85% y un porcentaje de error del 10%.

Arora & Kaur (2019) utilizaron un dataset de Kaggle, una página en línea que pertenece a una comunidad de científicos de datos y entusiastas del aprendizaje automático, la cual contaba con 42.350 registros y 143 variables con datos desde 2007 a 2011 dentro de la investigación aplicaron modelos de clasificación, en donde Random Forest presento una precisión en las predicciones de 97.9%, un indicador AUC ROC o área bajo la curva del 93.4% y una tasa de error de alrededor de 2.1%.

Otro aporte importante dentro de esta línea es el propuesto por Subasi & Cankurt (2019) en el cual comparan el comportamiento de pago de clientes dentro de varios segmentos de crédito para una base de datos de 25,000 registros con 23 variables , a su vez también presentan una comparación de la metodología utilizada versus otras metodologías como la propuesta por Arora & Kaur (2019) en la cual luego de aplicar diferentes metodologías de clasificación se presenta el resultado que al aplicar Random Forest se obtiene un valor del 89.01% en la precisión de predicciones, es decir el mayor porcentaje.

2.2 Árboles de Decisión dentro del riesgo de crédito

Como se detallo anteriormente dentro de las técnicas de clasificación se presentan varias metodologías una de ellas son los Árboles de Decisión que es una herramienta de machine learning en su mayoría utilizada dentro de modelos

enfocados en clasificación y regresión. La estructura como su nombre lo indica se asemeja a un árbol, dentro del cual cada nodo que se presenta es una decisión basada en el valor de cierto atributo, y cada rama por su parte representa el resultado de dicha decisión. (Breiman, Friedman, Olshen, & Stone, 1984).

Los primeros indicios que datan del estudio pionero de la utilización y aplicación de Árboles de Decisión se remontan a la década de 1960, Morgan (1963) exploró la viabilidad de la aplicar de Árboles de Decisión en su trabajo presentado en las Actas de la Conferencia Nacional de Inteligencia Artificial, en donde destaco la aplicación de árboles de decisión para clasificar a los clientes de acuerdo con su capacidad de pago y tomar decisiones de crédito más informadas.

López et al (2017) y Pérez et al (2019) realizan estudios de como evaluar el riesgo dentro de instituciones financieras para lo cual se utiliza datos de préstamos de entidades bancarias para desarrollar modelos de Árboles de Decisión para clasificar a las operaciones de los clientes de la entidad. Además, evaluaron la eficiencia de dichas metodologías en la predicción del comportamiento de los prestatarios en el cumplimiento de sus pagos y en la identificación de patrones de riesgo clave, es asi que ambos muestran resultados similares en sus estudios los cuales indican que los Árboles de Decisión mejoraron significativamente la precisión en la evaluación del riesgo de crédito, por lo que se sugiere su utilidad en la toma de decisiones crediticias más informadas y efectivas.

Khandani et al (2010) por su parte realizan un estudio sobre la predicción de riesgo de crédito utilizando técnicas de machine learning utilizando variables de historial de transacciones, comportamiento de gastos, historial de pagos, saldo de cuentas, edad, ingresos, estado civil, indicadores económicos e interacciones en redes sociales, entre otras. El uso de estos datos permitió a los autores construir varias metodologías para la predicción del riesgo de crédito como redes neuronales, support vector machines, random forest y árboles de decisión, los principales hallazgos de la investigación detallan que en particular la metodología de Árboles de Decisión proporciona una mejor robustez y precisión en la predicción del riesgo de crédito comparado con los modelos tradicionales, además que capturan

relaciones no lineales entre las variables, lo cual dentro del riesgo de crédito es crucial dado que las interacciones entre factores puede ser complejas de identificarlas en la aplicabilidad.

Finalmente, Zhou et al (2020) realizan una inclusión de variables innovadoras para analizar el riesgo de crédito y clasificar a los clientes de una institución, en el mencionado estudio se incluyen variables como: historial crediticio, datos demográficos (edad, ingresos, estado civil, ocupación), interacciones en redes sociales como publicación de contenido y redes de contacto, sentimiento y comportamiento ante publicaciones y finalmente redes y conexiones, es decir solidez de las redes de contactos del prestatario, que pueden indicar su capital social y apoyo comunitario. Para dicho estudio de igual manera se aplican metodologías de árboles de decisión y NPL Procesamiento de lenguaje natural. Dentro de sus principales hallazgos muestran que la utilización de Árboles de Decisión dentro de modelos de clasificación son altamente efectivos para manejar y procesar datos no estructurados especialmente de redes sociales, otro hallazgo relevante es también la importancia de utilizar variables como historial crediticio y datos demográficos, pero sobre todo datos de redes sociales.

De forma general, se presenta varios estudios que muestran la importacia de aplicar metodologías de clasificación dentro de instituciones financieras, de manera que permite mejorar la precisión de sus modelos de riesgo y a su vez explorar nuevas fuentes de datos que proporcionen una visión más completa y matizada del prestatario.

Metodología

3.1 Origen de los Datos

El presente proyecto desarrolla un modelo de clasificación de clientes para una institución financiera, utilizando técnicas avanzadas de aprendizaje automático, con el principal objetivo de controlar y mitigar el riesgo de crédito, para ello el presente desarrollo utiliza el conjutno de datos de la base “Give Me Some Credit”, la cual es una dataset público de Kaggle, mismo que proporciona características detalladas sobre el comportamiento de un cliente. Dentro del mundo de análisis de datos esta base ha sido ya utilizada para varias metodologías de machine learning debido a su utilidad dentro la exploración y modelización del riesgo de crédito. (Kaggle, s.f.)

Esta base fue seleccionada para el desarrollo del presente proyecto por varias razones como las siguientes: la calidad de los datos, dado la naturaleza del conjunto de datos, estos proporcionan información pertinente acerca de los factores de riesgo de crédito lo que va de la mano con el objetivo planteado en este estudio; la disponibilidad de variables con las que cuenta la base permite explorar varias opciones dentro del comportamiento crediticio; y finalmente las competencias de la base dentro del análisis de datos, debido a que la base ha sido ya utilizada dentro de la evaluación de modelos predictivos y ofrece desafíos complejos para el proceso de predicción y estimación del riesgo de crédito.

Para el proceso de adquisición de datos como es una base pública el proceso de accesibilidad y descarga implican el registro correspondiente dentro de la plataforma de Kaggle. De igual forma, al ser una base pública esta se encuentra alineada con las políticas de Kaggle en lo que se refiere a las consideraciones éticas dado que permite utilizar los datos con fines educativos e investigativos, por lo que se garantiza el manejo adecuado de los datos respetando las normas de confidencialidad y privacidad, considerando que los datos no cuentan con un identificador único que permitan ligar a información personal de una persona. (Kaggle, s.f.)

3.2 Descripción de variables

El conjunto de datos antes expuesto consta de 150.000 registros y un total de once variables las cuales incluyen características sociodemográficas de un cliente, capacidad de pago e información crediticia, las cuales se detallan a continuación:

Tabla 2: Detalle de variables seleccionadas

Variable	Descripción	Tipo
Morosidad últimos 2 años	Variable binaria que indica si un cliente ha experimentado una morosidad grave en los últimos dos años	Categórica
Utilización línea no segura	Tasa de utilización de líneas de crédito no garantizadas.	Numérica %
Edad	Edad del cliente	Numérica
Retraso 30-59días	Número de veces que el cliente ha estado entre 30 y 59 días atrasado en su pago, pero no más grave, en los últimos dos años	Numérica
Ratio deuda	Relación entre la deuda y los ingresos mensuales	Numérica %
Ingreso mensual	Ingreso mensual del cliente	Numérica
Créditos abiertos	Número de líneas de crédito y préstamos abiertos	Numérica
Retraso 90días	Número de veces que el cliente ha estado 90 días o más atrasado en su pago.	Numérica
Prestámos inmobiliarios	Número de préstamos o líneas de crédito inmobiliario	Numérica
Retraso 60-89días	Número de veces que el cliente ha estado entre 60 y 89 días atrasado en su pago, pero no más grave, en los últimos dos años	Numérica
Dependientes	Número de dependientes en la familia excluyendo al cliente mismo	Numérica

Fuente: Kaggle
Elaboración propia

En este caso la variable objetivo en este caso es “Morosidad en los últimos dos años”, la cual es una binaria que divide a los clientes en dos grupos, asigna el valor

de 1 si el cliente cayó en default y presentó valores vencidos o morosidad en los últimos dos años, y asigna el valor de 0 si el cliente no cayó en default, es así como se podría denominar a la clase 1 como mal cliente y a la clase 0 como buen cliente.

3.3 Preparación e Integración de Datos

La presente sección muestra el tratamiento y preparación de los datos, esta etapa previa del procesamiento de datos es relevante debido a que ayuda a tener una visión clara del análisis de datos, y de sta forma se elimina el ruido o variables irrelevantes que no aportan valor al estudio (Kotsiantis, Kanellopoulos, & Pintelas, 2006). El objetivo es identificar la presencia de valores pérdidos o atípicos que generen ruido a las predicciones y resultados finales.

En la tabla 3 se detalla la estadistica descriptiva para cada una de las variables en análisis, de manera preliminar las estadísticos muestran que las variables “Ingreso mensual” y “Dependientes” presentan valores perdidos, adicional las variables “Ratio deuda” e “Ingreso mensual” presentan valores extremos (outliers) que afectan significativamente el promedio y su vez pueden afectar las estimaciones.

Tabla 3 Estadística Descriptiva

	Morosidad últimos 2 años	Utilización línea no segura	Edad	Retraso 30-59días	Ratio deuda	Ingreso mensual	Créditos abiertos	Retraso 90días	Prestámos inmobiliarios	Retraso 60-89días	Dependientes
count	150,000.00	150,000.00	150,000.00	150,000.00	150,000.00	120,269.00	150,000.00	150,000.00	150,000.00	150,000.00	146,076.00
mean	0.07	6.05	52.30	0.42	353.01	6,670.22	8.45	0.27	1.02	0.24	0.76
std	0.25	249.76	14.77	4.19	2,037.82	14,384.67	5.15	4.17	1.13	4.16	1.12
min	-	-	-	-	-	-	-	-	-	-	-
25%	-	0.03	41.00	-	0.18	3,400.00	5.00	-	-	-	-
50%	-	0.15	52.00	-	0.37	5,400.00	8.00	-	1.00	-	-
75%	-	0.56	63.00	-	0.87	8,249.00	11.00	-	2.00	-	1.00
max	1.00	50,708.00	109.00	98.00	329,664.00	3,008,750.00	58.00	98.00	54.00	98.00	20.00

Fuente: Kaggle
Elaboración propia

De acuerdo con lo antes expuesto se realiza el análisis de datos pérdidos para cada variable del modelo. En efecto en la Tabla 4 se puede observar la cantidad y porcentaje de datos perdidos para cada una de las variables, la variable ingreso mensual presenta el 20% de datos perdidos en relación con las observaciones

totales, de igual manera la variable dependientes presenta un 3% de datos perdidos del total de las observaciones. Ante la presencia de datos perdidos es importante considerar aplicar métodos de imputación en ambos casos. Sin embargo, se debe analizar primero la presencia de datos atípicos, pues existe la posibilidad de que los valores imputados estén influenciados de manera negativa por los valores extremos y sesgar asi los resultados. (Tukey, 1977)

Tabla 4 Datos perdidos

Variable	Datos Perdidos		Porcentaje
Morosidad últimos 2 años	0		0%
Utilización línea no segura	0		0%
Edad	0		0%
Retraso 30-59días	0		0%
Ratio deuda	0		0%
Ingreso mensual		29,731	20%
Créditos abiertos	0		0%
Retraso 90días	0		0%
Prestámos inmobiliarios	0		0%
Retraso 60-89días	0		0%
Dependientes	3,924		3%

Fuente: Kaggle
Elaboración propia

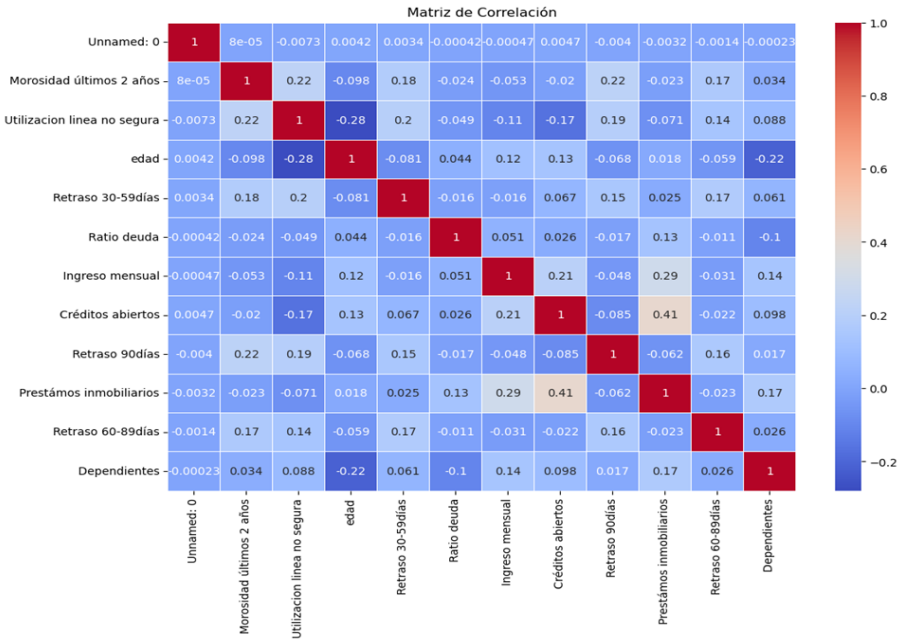
Como se detalla en la Tabla 3 el valor de los estadísticos descriptivos señala la presencia de datos atípicos para las variables, lo cual se corrobora en el diagrama de caja y bigotes, el cual se presenta en la sección de anexos, Anexo 1.

Para la exclusión de datos atípicos se aplico la técnica de rango intercuartílico el cual mide la variabilidad de los datos, aplicando el rango entre el primer cuartil (Q1) y el tercer cuartil (Q3), es decir IQR = Q3-Q1, el IQR permite excluir los datos atípicos dado que cualquier dato que se encuentre por debajo del valor de Q1 - 1.5*IQR o sobre el valor de Q3 + 1.5*IQR suele considerarse un valor atípico. (Tukey, 1977).

Una vez realizado el ajuste de los outliers o datos atípicos se continua con la corrección de los datos faltantes de las variables ingreso mensual y dependientes, para imputar estos datos se aplicó la técnica de imputación por la media, la cual se basa en reemplazar los valores ausentes o faltantes con el valor promedio de esa variable calculado a partir de los datos disponibles, la ventaja que tiene esta técnica es que mantiene la media de los datos evitando que los valores imputados alteren el promedio de la variable. (Little & Rubin, 2002)

Por otro lado, para analizar las relaciones que se pueden presentar entre las variables en estudio se muestra la matriz de correlación, en la cual cada celda contiene el coeficiente de correlación entre las variables. La matriz de correlación no solo nos permite identificar relaciones directas entre pares de variables, sino que también puede ayudarnos a detectar patrones más complejos y a identificar variables redundantes. Por ejemplo, si dos variables están altamente correlacionadas, podría ser no necesario incluir ambas en un modelo predictivo, ya que aportan información similar. Se puede observar en este caso en la Figura 1 que las variables de créditos inmobiliarios con créditos abiertos tienen una correlación positiva alta, mientras que la variable edad tiene una correlacion negativa con el número de dependientes que puede tener una persona, es decir a medida que la edad aumenta el número dependientes para una persona o dentro de una familia disminuye. (Field, 2013)

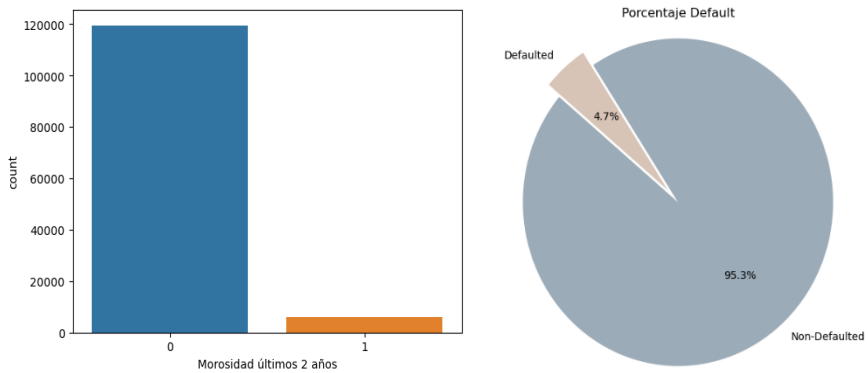
Figura 1: Matriz de correlación



Fuente: Kaggle
Elaboración propia

En lo que corresponde a la variable objetivo la cual divide en dos grupos a las observaciones, personas con morosidad en los últimos dos años toman el valor de uno y caso contrario el valor de cero, se puede observar en la Figura 2 que el 4.7% del total de la base cayó en default, mientras que el 95.3% no presento saldos vencidos.

Figura 2 : Frecuencia variable objetivo



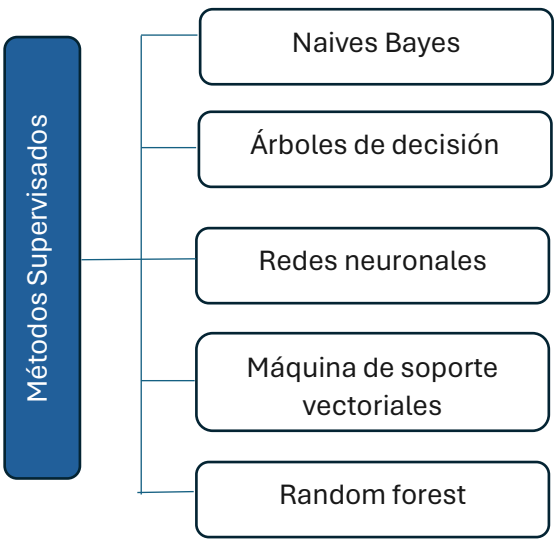
Fuente: Kaggle
Elaboración propia

3.4 Justificación y aplicación de la metodología

Dentro el acápite de revisión literaria del presente proyecto se expuso varios autores con proyectos similares en donde aplican metodologías de clasificación, estudios como el de Khandani et al (2010), Zhou et al (2020) utilizan Árboles de Decisión para realizar predicciones dentro del riesgo de crédito y exponen que se obtiene mejores resultados con su aplicación comparado con otro tipo de metodologías. Por otro lado, Arora & Kaur (2019), Ziemba et al (2020) , Pradhan et al (2020) exponen mejores resultados y eficiencia en el modelo con la aplicación de Random Forest. Es asi que la metodología propuesta para el presente proyecto de investigación es la aplicación de Árboles de Decisión y Random Forest, con la aplicación de ambas metodologías se busca determinar aquella que presente el mejor nivel de ajuste y predicción de los resultados.

Dentro del campo de la inteligencia artificial, el aprendizaje automático es un método que implementa métodos computacionales capaces de detectar patrones y predicciones, dentro del aprendizaje automático existen diferentes tipos de algoritmos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. El aprendizaje supervisado es una técnica que utiliza algoritmos que trabajan con conjuntos de entrenamiento conformados etiquetas o clases, en esta técnica los algoritmos aprenden en función de los datos históricos y de esta manera realizan las predicciones de repuesta dependiendo de la clase *(Freire Lopez, 2021)*. A continuación se presenta algunas metodologías del aprendizaje supervisado. (Jiang, Gradus, & Rosellini , 2020)

Figura 3: Métodos de aprendizaje supervisado



3.4.1 Árboles de decisión

Un árbol de decisión se basa en una estructura jerárquica donde cada nodo que lo conforma representa una condicion sobre las carácterísticas del conjunto de datos, y cada hoja representa la predicción de la variable objetivo. El desarrollo y la construcción del árbol implica dividir los datos en regiones iguales con relación a la variable objetivo. Para la construcción del árbol se aplican los siguientes pasos:

1. Selección variable objetivo

Una vez identificada la variable objetivo este es el punto de corte que divide el conjunto de datos en términos de dicha variable, para ello se utiliza diferentes criterios que dependen si el problema es de clasificación o regresión.

- *Clasificación*

El índice de Gini mide la hetorogeneidad de un nodo, es decir ayuda a evaluar a calidad de división de datos y se define como:

$$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$$

Donde:

C es el número de clases

p_i es la proporción de observaciones pertenecientes a la clase i en el nodo t.

- *Regresión*

El error cuadrático mide la variabilidad dentro de las regiones de un árbol de decisión, evalua la calidad de una división cuantificando la discrepancia entre los valores predichos versus los observados y se define como:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y)^2$$

Donde y_i es el valor observado y es el valor predicho y N el número de observaciones

2. División del conjunto de datos

Los datos se dividen en dos subconjuntos utilizando la variable objetivo como punto de corte, la divisiòn de los datos se realiza de manera que se minimiza el error en los subconjuntos relevantes de la siguiente manera:

$$\min_{j,s}\left[\sum_{i\in R_1(j,s)} L(y_i, R_1) + \sum_{i\in R_2(j,s)} L(y_i, R_2)\right]$$

Donde L es la función de pérdida

3. Iteracción

Este procedimiento de selección y partición se repite de manera recursiva para cada subconjunto hasta que se alcance un criterio de finalización, como la profundidad máxima del árbol o el número mínimo de muestras en un nodo, o la ganancia mínima de información. (Breiman, Friedman, Olshen, & Stone, 1984).

3.4.2 Random Forest

Random Forest es una técnica de aprendizaje automático mayormente utilizada en tareas de clasificación y regresión al igual que los árboles de decisión. La idea principal de Random Forest es combinar varios árboles de decisión de manera que permita obtener un modelo más estable y preciso. (Breiman, 2001).Para la construcción del modelo se aplican los siguientes pasos:

1. Muestreo

En el muestreo se generan B subconjuntos de datos del conjunto de datos inicial. Cada subconjunto en este caso tiene el mismo tamaño del conjunto original.

$$\{D_{1,}D_{2,,,,,,,,,}D_{B,}\}$$

2. Entrenamiento de Árboles

Para cada subconjunto de datos generado $D_b(b = 1{,}2{,,,,}B)$ se entrena un árbol de decisión T_b. Dentro de cada uno de los nodos del árbol se selecciona de manera

aleatoria un subconjunto de características m de las características p disponibles, para determinar la mejor división.

3. Agregación de Resultados

-*Clasificación*

Se toma un voto mayoritario de las predicciones de todos los árboles para determinar la clase final. La predicción $\hat{y}$ para cada observación x se define como:

$$\hat{y} = mode\ (\{T_1(x), T_2(x), , , , , , , , , T_B(x)\})$$

-*Regresión*

En este caso se calcula el promedio de las predicciones de todos los árboles, de esta forma se obtiene el valor final. En este caso la predicción $\hat{y}$ para cada observación x se define como:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

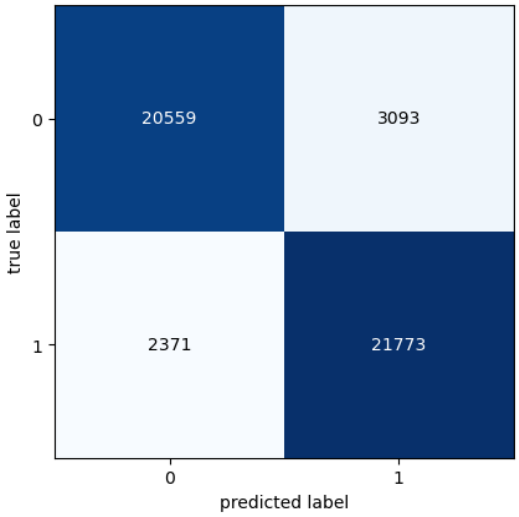
Resultados

En el presente capítulo se presenta los principales resultados para cada una de las metodologías propuestas y una evaluación correspondiente sobre las mismas.

4.1 Resultados Árbol de Decisión

Para modelo de clasificación de Árbol de Decisión se muestra en la Figura 4 la matriz de confusión generada, la cual refleja la evaluación del modelo y las clasificaciones generadas, de esta manera el modelo identificó correctamente a 20,559 lo que corresponde a la tasa de verdaderos positivos. De la misma manera de los clientes que eran malos, el modelo identifico correctamente a 21,773 de ellos, también conocidos como verdaderos negativos. Sin embargo, hubo casos en los que el modelo se equivocó, identificó incorrectamente a 3,093 clientes buenos como malos es decir como falsos negativos y 2,371 clientes malos como buenos es decir falsos positivos. Se puede determinar que el modelo tiende a presentar un cierto grado de dificultad para clasificar a los clientes como malos.

Figura 4 : Matriz de confusión



Fuente: Kaggle

Elaboración propia

Por otro lado en la Tabla 5 se presenta las métricas de desempeño del modelo, la precisión mide la exactitud del modelo, en este caso el valor de 88% indica que el 88% de las clasificaciones que fueron realizadas por el modelo fueron correctas, ya sea en la predicción de clientes buenos o malos.

La recuperación o Recall es conocida como la sensibilidad o capacidad que tiene modelo para determinar todos los casos positivos. En donde en este caso el modelo logro una recuperación del 87% de los clientes buenos y el 90% de clientes malos del conjunto de datos.

Por su parte el puntaje F1 es la media armónica de precisión y recuperación. Es útil cuando las clases están desequilibradas, como parece ser el caso aquí. El puntaje de 0.88 para F1 muestra un equilibrio entre precisión y recuperación lo que se evidencia con la clasificación entre los clientes de clase 0 y clase 1.

Tabla 5: Desempeño del Modelo Árbol de Decisión

	Precision	Recall	F1-score	Support
Clase 0	0.89	0.87	0.88	35,698
Clase 1	0.87	0.90	0.88	35,995
Accuracy			0.88	71,693
Macro avg	0.88	0.88	0.88	71,693
Weighted avg	0.88	0.88	0.88	71,693

Fuente: Kaggle

Elaboración propia

Una vez analizado el desempeño del modelo es importante validar el AUC-ROC [1], la cual es una métrica que permite evaluar el rendimiento del modelo en una clasificación binaria, teniendo en cuenta la capacidad de distinguir entre las clases, la independencia del umbral de clasificación y la resistencia ante clases desbalanceadas. En este caso un valor de 0.88 sugiere que el modelo tiene una buena capacidad discrimitaria y no hay evidencia de un sobreajuste en el modelo. (Ver Anexo 2).

[1] Área Bajo la Curva de Característica Operativa del Receptor

Finalmente, se analiza la validación cruzada con el obtjetivo de evaluar la capacidad de generalización del modelo a datos no vistos. Los resultados en este caso indican que, en promedio, el modelo mantuvo un rendimiento del 85.5% en los diferentes subconjuntos de datos. La pequeña desviación estándar (0.0723) sugiere que el modelo es estable y consistente en su rendimiento.

En resumen, estos resultados sugieren que el modelo de clasificación de Árbol de decisión es robusto y efectivo para identificar clientes buenos y malos, con un buen equilibrio entre precisión y recuperación, respaldado por su capacidad de generalización demostrada en la validación cruzada.

Tabla 6: Validación cruzada Árbol de Decisión

Iteracción	Entrenamiento	Prueba
1	0.855	0.854
2	0.855	0.855
3	0.853	0.850
4	0.856	0.843
5	0.854	0.851
Promedio	0.855	0.849
Desviación estándar	0.007	0.030

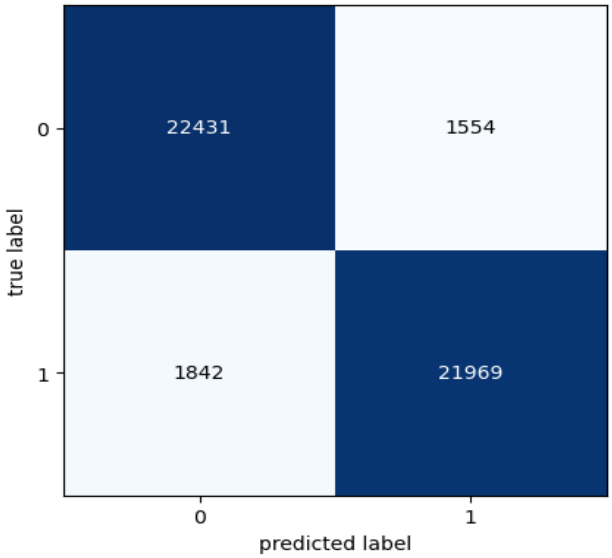
Fuente: Kaggle
Elaboración propia

El desarrollo y contrucción del arbol de decisión en el presente proyecto permite evidenciar que las variables "Utilización línea no segura" y "Dependientes" emergen como los principales predictores en la clasificación de nuestros datos. Estas variables, al ser seleccionadas en los primeros nodos del árbol, indican su alta importancia para discernir entre las clases objetivo, su presencia en los nodos iniciales sugiere que poseen una fuerte influencia en la determinación de la clasificación de las instancias. La revisión inicial de nuestro modelo revela que estas variables son fundamentales para entender las relaciones subyacentes en nuestros datos. Sin embargo, para una comprensión más detallada y exhaustiva de las interacciones y patrones presentes, se invita a consultar el Anexo 3, que proporciona el Árbol de Decisión completo y ofrece una exploración detallada de las relaciones entre las variables.

4.1 Resultados Random Forest

Con base en el modelo de clasificación Random Forest o bosque aleatorio se observa la Figura 5, la cual permite analizar la calidad de predicción de los clientes, misma que reflejo un alto de grado de precisión. En lo que corresponde a los clientes buenos, 22,431 fueron correctamente identificados y 21,969 clientes malos fueron correctamente identificados. Por otro lado, la cantidad de falsos positivos y negativos es mínima, pues se presentó 1,554 clientes identificados como malos y en realizada fueron malo y a su vez 1,842 clientes identificados como buenos los cuales en realidad fueron buenos.

Figura 5: Matriz de confusión Random Forest



Fuente: Kaggle
Elaboración propia

Por otro lado, en la Tabla 7 se presentan las métricas del modelo, donde se puede observar que el modelo logró una precisión y un recall del 92% para los clientes buenos y del 93% para los clientes malos, lo que indica una alta exactitud tanto en la identificación de clientes valiosos como en la detección de aquellos con riesgo potencial, y se obtuvo un puntaje F1 de 0.93 para ambas clases, lo que muestra un equilibrio entre precisión y recall en la clasificación.

Tabla 7: Desempeño del Modelo Random Forest

	Precision	Recall	F1-score	Support
Clase 0	0.93	0.94	0.93	23,985
Clase 1	0.93	0.92	0.93	23,811
Accuracy			0.93	47,796
Macro avg	0.93	0.93	0.93	47,796
Weighted avg	0.93	0.93	0.93	47,796

Fuente: Kaggle
Elaboración propia

El área bajo la curva por su parte fue excepcionalmente alta, alcanzando 0.9804. Este valor destaca la capacidad que presenta el modelo para diferenciar entre las distintas clases con una precisión notable. (Ver Anexo 4)

Finalmente, para analizar la capacidad del modelo se presenta en la Tabla 8 la validación cruzada, la cual arrojó una puntuación media de 0.9216, con una desviación estándar de 0.0005. Estos resultados muestran una consistencia y estabilidad en el rendimiento del modelo en los diferentes conjuntos de datos, lo que respalda su capacidad de generalización.

En conjunto, estos hallazgos muestran la eficacia y robustez del modelo de clasificación aplicado en la evaluación de clientes, con una precisión notable, un rendimiento sólido en términos de AUC-ROC y una buena capacidad de generalización demostrada mediante la validación cruzada.

Tabla 8: Validación Cruzada Random Forest

Iteracción	Entrenamiento	Prueba
1	0.921	0.880
2	0.922	0.877
3	0.921	0.886
4	0.922	0.880
5	0.920	0.880
Promedio	0.921	0.881
Desviación estándar	0.0005	0.002

Fuente: Kaggle
Elaboración propia

El modelo de Random Forest está compuesto por varios árboles de decisión, por lo cual visualizar estos árboles generados de manera individual puede ser un poco complicado o innecesario, especialmente porque el bosque aleatorio está formado por varios árboles. Sin embargo, dentro del análisis se puede generar gráficos que nos permita determinar las variables relevantes dentro de la construcción del modelo, estos gráficos pueden proporcionar una comprensión más detallada del funcionamiento del modelo en conjunto. En este caso para el modelo generado se establece que las variables que toman relevancia son "Utilización línea no segura" y "Dependientes", cuyo gráfico correspondiente se puede visualizar en el Anexo xx

Discusión de Resultados y Propuesta de Solución

De acuerdo con la descripción de resultados detallada en el acápite anterior, se puede indicar que ambas metodologías de Árbol de Decisión y Random Forest muestran un considerable desempeño dentro de la clasificación de clientes dentro del riesgo de crédito. Por su parte ambos modelos lograron una precisión y un F1-score de alrededor del 0.88, indicando así una capacidad adecuada para distinguir entre las clases clientes entre 0 y 1.

Adicional, al realizar el análisis de la validación cruzada, se observa una consistencia en el desempeño del Árbol de Decisión, con un promedio de precisión del 0.855 en el conjunto de entrenamiento y del 0.849 en el conjunto de prueba, con una desviación estándar de 0.007 y 0.030 respectivamente. Por otro lado, el modelo de bosque aleatorio exhibió un desempeño ligeramente superior, con un promedio de precisión del 0.921 en el conjunto de entrenamiento y del 0.881 en el conjunto de prueba, con una desviación estándar muy baja de 0.0005 y 0.002 respectivamente.

Bajo este contexto es importante mencionar que, a pesar que el modelo de bosque aleatorio mostró una mayor precisión promedio en la validación cruzada, la diferencia en el desempeño entre ambos modelos no fue tan significativa como se esperaba. Esto permite sugerir que la aplicación de la metodología de Árbol de Decisión es robusta y estable en diferentes particiones de datos, lo que lo hace una opción confiable para la clasificación de clientes.

En lo que corresponde a la propuesta de solución y basándonos en los resultados de la validación cruzada y en la comparación de desempeño entre los modelos de Árbol de Decisión y Random Forest, se propone la implementación del modelo de el Árbol de Decisión como la solución preferida para la clasificación de clientes dentro del riesgo de crédito. Además, la estabilidad y consistencia del Árbol de Decisión en diferentes particiones de datos, junto con su fácil interpretabilidad, hacen de esta metodología las más adecuada para su implementación y práctica dentro de entornos financieros. Pues como se detalló anteriormente el Árbol de Decisión generado nos permite analizar cada una de sus ramificaciones, es decir

las relaciones que tienen las variables entre sí dentro de la decisión de clasificación de los clientes.

Por otro lado, se observa que el modelo de Random Forest no muestra una mejora significativa en el desempeño en comparación con el Árbol de Decisión, pero la complejidad del Random Forest no puede ser justificada en este caso dado que el Árbol de Decisión ya proporciona un sólido rendimiento y consistente en la clasificación de clientes como se evidencia en la validación cruzada y métricas de desempeño. Dado que la interpretabilidad de un modelo es importante dentro de los entornos financieros para la comprensión de las decisiones propuestas, el Árbol de Decisión ofrece una estructura clara e interpretable para entender la clasificación.

Se reitera entonces la importancia de integrar el modelo de Árbol de Decisión en los sistemas de gestión de riesgos financieros dentro de una institución, aprovechando su capacidad para identificar tempranamente a los clientes en riesgo de morosidad y facilitar la implementación de medidas preventivas. En conclusión, la combinación de los resultados de la validación cruzada y la comparación de desempeño respalda la propuesta de elección del Árbol de Decisión como la solución óptima para la gestión de la morosidad crediticia en este contexto específico.

Es importante mencionar que el uso de estos modelos no se limita únicamente a la predicción de comportamiento y clasificación de clientes, pueden al comprender de mejor manera las características y el comportamiento de un cliente se puede proponer estrategias financieras específicas para cada uno de los segmentos adaptando los términos de los préstamos, ofreciendo servicios adicionales de asesoramiento financiero o implementando programas de educación sobre la gestión del crédito.

La aplicación e integración de este tipo de modelos en los procesos de gestión estratégica y de toma de decisiones de una institución representa una avance dentro del campo de la innovación, pues la capacidad del manejo y utilización de un gran volumen de datos históricos para predecir el comportamiento de un cliente genera nuevas oportunidades para la personalización de los servicios

financieros y optimización de la administración y gestión de riesgo, especialmente dentro del riesgo de crédito, de esta forma no solo permite mejorar la eficiencia de la institución, si no a su vez aumenta su reputación al ofrecer soluciones adaptadas con necesidades individuales.

En conclusión, la aplicación de metodologías como Random Forest y Árbol de Decisión no solo contribuye a una gestión más efectiva de la morosidad crediticia, sino que también impulsa la innovación dentro de la institución financiera. Al aprovechar el poder predictivo de estos modelos y diseñar estrategias personalizadas basadas en sus resultados, la institución puede mejorar su competitividad en el mercado y fortalecer su relación con los clientes, promoviendo así un crecimiento sostenible y una mayor estabilidad financiera.

Conclusiones y Recomendaciones

En el presente proyecto se realizó una evaluación sobre dos metodologías de aprendizaje automático, Árbol de Decisión y Random Forest, para la clasificación de clientes dentro del riesgo de crédito utilizando un conjunto de variables relevantes dentro de estudios similares desarrollados dentro de la misma rama de investigación. De manera conjunta ambos modelos muestran que los resultados sugieren su utilidad dentro del riesgo de crédito dada su capacidad de clasificación.

El análisis de la validación cruzada reveló una consistencia en el desempeño del árbol de decisión, con una precisión promedio del 0.855 en el conjunto de entrenamiento y del 0.849 en el conjunto de prueba. Por otro lado, el modelo de Random Forest mostró un desempeño ligeramente superior, con una precisión promedio del 0.921 en el conjunto de entrenamiento y del 0.881 en el conjunto de prueba. Aunque dicha diferencia en la precisión promedio es relevante, es importante tener en cuenta que la interpretación del bosque aleatorio puede ser más compleja y menos intuitiva en comparación con el árbol de decisión.

La comparación realizada entre ambos modelos resalta la utilidad de considerar no solo la precisión, sino también la interpretabilidad y la eficiencia computacional al seleccionar una metodología para la clasificación de clientes dentro del riesgo de crédito. Si bien por un lado el modelo de Random Forest puede ofrecer precisiones en las predicciones ligeramente superiores, el Árbol de Decisión brinda una ventaja significativa en términos de interpretabilidad y facilidad de implementación en entornos financieros prácticos.

Finalmente, este estudio proporciona una base sólida para la adopción del modelo de Árbol de Decisión como una herramienta efectiva en la gestión de riesgo de crédito. Sin embargo, se recomienda realizar investigaciones adicionales para explorar nuevas variables y enfoques que puedan mejorar la robustez y precisión del modelo en diferentes contextos y escenarios económicos. Con un enfoque continuo en la mejora e innovación, las instituciones financieras pueden fortalecer su capacidad para identificar y mitigar los riesgos asociados con la morosidad crediticia, promoviendo así la estabilidad y la salud financiera a largo plazo.

Es importante también realizar un monitoreo continuo sobre las metodologías aplicadas, sobre todo del desempeño del modelo de Árbol de Decisión para asegurar de esta forma su efectividad y precisión en las predicciones realizadas a través del tiempo, esto incluye también realizar una revisión periódica de los datos y la actualización del modelo de acuerdo con las necesidades de la organización.

Por otro lado, es importante que las organizaciones que apliquen este tipo de metodologías cuenten con personal calificado al cual a su vez se encuentre en constante capacitación, esto garantizará una correcta comprensión de las predicciones del modelo y una adecuada toma de decisiones, adicional se recomienda la inclusión de otras variables relevantes como en el estudio de Zhou et al (2020) en el cual se consideras variables sobre el comportamiento del cliente en redes sociales, reacciones y su capital social, dichas variables pueden mejorar la precisión del modelo de clasificación de morosidad crediticia.

Bibliografía

Arora, N., & Kaur, P. (2019). Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assesment . *Applied Soft Computing*, 86.

Boudriga, A., Taktak, N., & Jellouli, S. (2010). Bank Specific, Business and Institutional Environment Determinants of Banks Nonperforming Loans. *The Economic Research Forum (ERF)*, 3-28.

Breiman, L. (2001). Random Forests. *Machine Learning*, 4-57.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. . *Wadsworth & Brooks/Cole Advanced Books & Software.*

Cohen, R., Krishnamoorthy, G., & Wright, A. (2019). Credit Risk Assessment: A Review of Machine Learning Applications. *International Review of Financial Analysis*, 1-15.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics. *SAGE Publications*.

Freire Lopez, J. (2021). “Modelo de Clasificación de Riesgo Crediticio Utilizando Random. *Universidad UISEK*.

Jiang, T., Gradus, J., & Rosellini , A. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, 675-687.

Kaggle. (s.f.). *Give Me Some Credit*. Obtenido de https://www.kaggle.com/searchQuery=give+me+some+credit+&page=2

Khandani, A., Kim, A., & Lo, A. (2010). Consumer Credit-Risk Models via Machine-Learning Algorithms. *Journal of Banking & Finance*.

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*.

Lee, I., & Shin, Y. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, 155-158.

Liebergen, B. (2017). Machine Learning: A Revolucion in Risk Management and Compliance. *Journal of Financial Transformation*, 60-67.

Little, R., & Rubin, D. (2002). Statistical Analysis with Missing Data . *Wiley*.

López, J., García, M., & Martinez, M. (2017). Evaluación del Riesgo de Crédito en Instituciones Financieras mediante Árboles de Decisión: Un Enfoque Práctico. *Journal of Banking and Finance*.

Mian, A., & Sufi, A. (2009). Las consecuencias de la expansión del crédito hipotecario: evidencia de la crisis de impagos hipotecarios en Estados Unidos . *The quarterly journal of economics*.

Morgan , J. (1963). Decision Trees for Learning. In Proceedings of the National Conference on Artificial Intelligence.

mundial, B. (2021). *Banco mundial*. Obtenido de Banco mundial: https://www.bancomundial.org/es/home

Pérez, L., Rodríguez, C., & Gómez, F. (2019). Mejora de la Evaluación del Riesgo de Crédito en Instituciones Financieras mediante Árboles de Decisión y Análisis de Sensibilidad. *International Journal of Finance & Economics*.

Pradhan, M., Akter, S., & Al Marouf, A. (2020). Performance Evaluation of Traditional Classifiers on Prediction of Credit Recovery . *Advances in Electrical and Computer Springer Singapore*, 541-555.

Raghuram, R. (2010). Fault Lines: How Hidden Fractures Still Threaten the World Economy. *Princeton University Press*.

Safari, M. (2020). Hybridization of multivariate adaptive regression splines and random forest models with an empirical equation for sediment deposition prediction in open channel flow. *Journal of Hydrology*.

Samaniego, R., & Rodriguez, L. (2015). El sector financiero y su contribución al crecimiento económico en Ecuador. *Revista Ciencia UNEMI*, 59-68.

Subasi, A., & Cankurt, S. (2019). Prediction of default payment of credit card clients using Data Mining Techniques . *International Engineering Conference (IEC)*, 115-117.

Tukey, J. (1977). Exploratory Data Analysis. *Addison-Wesley*.

Vargas Sánchez, A., & Mostajo Castelú, S. (2014). MEDICIÓN DEL RIESGO CREDITICIO MEDIANTE LA APLICACIÓN DE MÉTODOS BASADOS EN CALIFICACIONES INTERNAS. *Scielo Investigación & Desarrollo*, 5-12.

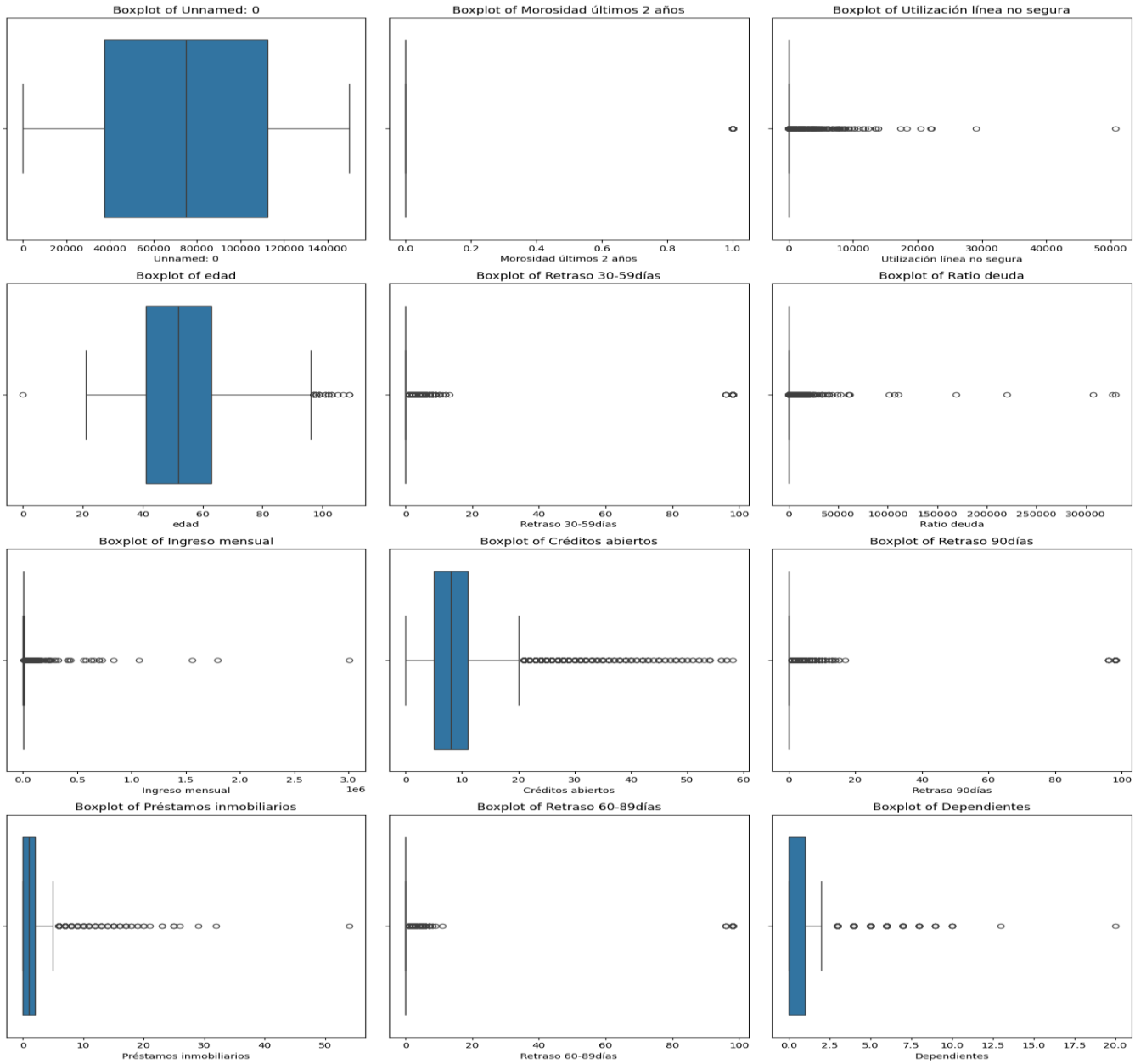
Zhou, W., Luo, C., Xia, M., & Zhang, X. (2020). Credit Scoring Using Machine Learning by Combining Social Media Information: Evidence from China. *Journal of Financial Stability*, 46-53.

Ziemba, P., Radomska - Zalas, A., & Becker, J. (2020). Client evaluation decision models in the credit scoring task . *Procedio Computer Science*, 3301-3312.

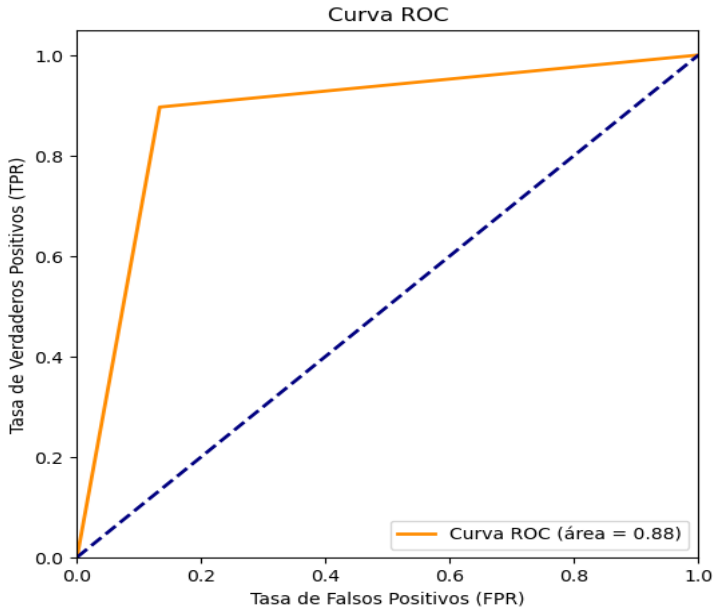
Anexos

Anexo 1. Diagrama de cajas y bigotes

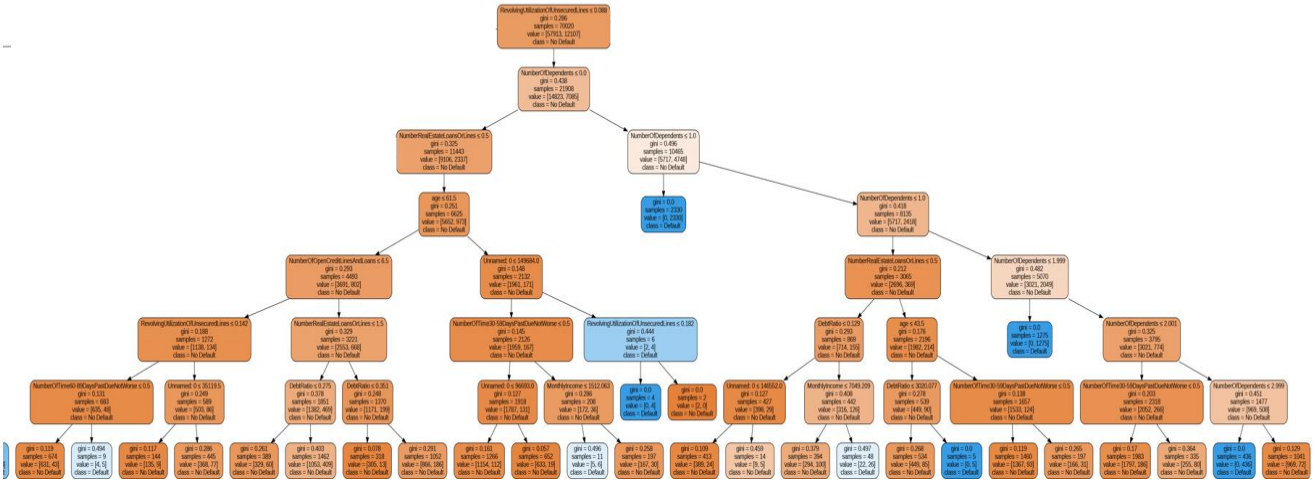
Se realizo el diagrama de caja y bigotes para cada una de las variables, de esta forma se puede evidenciar la presencia de datos atípicos, los cuales fueron corregidos mediante el proceso de rango intercuartil.



Anexo 2. AUC-ROC Árbol de Decisión



Anexo 3. Árbol de Decisión



Anexo 4. AUC-ROC Random Forest