UNIVERSIDAD NACIONAL DE SAN CRISTÓBAL DE HUAMANGA

FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS:

“Predicción de fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena basado en machine learning, 2025”

Para optar el título profesional de:
INGENIERO DE SISTEMAS

PRESENTADO POR:
Bach. Pavel Wigberto GARCIA ALVARADO

ASESOR:
Dr. Manuel Avelino LAGOS BARZOLA

AYACUCHO - PERÚ

2026

Dedicatoria

A mis padres, por su palabra cumplida sin condiciones.

A mi padre, por enseñarme con el ejemplo y por su apoyo discreto pero constante.

A mi madre, por estar presente en mi vida, por cada palabra de aliento y por su fe inquebrantable en mí.

Esta tesis lleva la huella de los dos.

Agradecimiento

A Dios, por acompañarme en cada etapa de mi vida y darme la fuerza para perseverar en mis metas.

A la Universidad Nacional de San Cristóbal de Huamanga, por haberme acogido en sus aulas durante mi formación y permitirme adquirir las bases académicas que sustentan este trabajo.

Al Dr. Manuel Lagos Barzola, por su orientación constante, su tiempo dedicado y la rigurosidad con que guió el desarrollo de esta investigación. Su acompañamiento fue clave para llevarla a buen término.

A la Cooperativa de Ahorro y Crédito Santa María Magdalena (CACSMM), por brindarme las facilidades y el acceso a los datos que hicieron posible este estudio.

A las personas que acompañaron de cerca este proceso: a quienes me ayudaron a retomar y culminar mi formación, a quien me apoyó durante mis prácticas preprofesionales, y a quienes desde la cooperativa me motivaron a sostener este proyecto. Cada uno, desde su lugar, contribuyó a que esta tesis sea hoy una realidad.

Resumen

La fuga de socios representa un problema estratégico para las cooperativas de ahorro y crédito, debido a su impacto en la estabilidad financiera, la sostenibilidad institucional y la capacidad de retención de socios. En este contexto, la presente investigación tuvo como objetivo desarrollar y comparar modelos de machine learning de clasificación supervisada, basados en características bancarias y comportamientos financieros de los socios, para identificar el modelo con mejor desempeño en la predicción del riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025. La investigación fue de tipo aplicada, con enfoque cuantitativo, nivel descriptivo-comparativo con propósito predictivo, diseño no experimental y corte transversal retrospectivo. Se utilizaron 55,432 registros históricos elegibles de socios, clasificados en socios no fugados y socios con renuncia voluntaria formal. El procedimiento técnico comprendió la depuración y limpieza de datos, análisis exploratorio, ingeniería de características, balanceo de clases mediante SMOTE, entrenamiento de modelos y evaluación comparativa mediante validación cruzada. Se evaluaron los algoritmos de Regresión Logística, Random Forest, Gradient Boosting, XGBoost y LightGBM, utilizando métricas como F1-score y ROC-AUC. Los resultados mostraron que las variables con mayor contribución a las predicciones comprendieron características bancarias, comportamientos financieros recientes e indicadores derivados. Entre ellas destacaron el ratio entre retiros y aportes, la cantidad de productos activos, los préstamos otorgados, la frecuencia de uso mensual, los retiros de los últimos tres meses y el riesgo de mora.

El modelo Random Forest obtuvo el mayor F1-score según el criterio principal de selección, con un valor de 0.6941 y un ROC-AUC de 0.9076, presentando el mejor desempeño entre los modelos evaluados según el criterio establecido dentro del conjunto de datos analizado. Se concluye que el modelo propuesto constituye una herramienta útil para apoyar la toma de decisiones estratégicas orientadas a la retención de socios dentro de la cooperativa.

Palabras clave: machine learning, fuga de socios, clasificación supervisada, cooperativa de ahorro y crédito, análisis predictivo.

Abstract

Member attrition represents a strategic problem for savings and credit cooperatives because of its impact on financial stability, institutional sustainability, and member retention capacity. In this context, the objective of this study was to develop and compare supervised machine learning classification models based on members' banking characteristics and financial behaviors to identify the best-performing model for predicting member attrition risk at the Santa María Magdalena Savings and Credit Cooperative in 2025. The research was applied in nature and followed a quantitative approach, a descriptive-comparative level with a predictive purpose, a non-experimental design, and a retrospective cross-sectional scope. A total of 55,432 eligible historical member records were used, classified as non-attrited members and members with a formal voluntary resignation. The technical procedure included data screening and cleaning, exploratory data analysis, feature engineering, class balancing using SMOTE, model training, and comparative evaluation through cross-validation. Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM were evaluated using metrics such as F1-score and ROC-AUC. The results showed that the variables with the greatest contribution to the predictions comprised banking characteristics, recent financial behaviors, and derived indicators. These included the withdrawal-to-contribution ratio, number of active products, loans granted, monthly usage frequency, withdrawals during the last three months, and delinquency risk.

According to the primary selection criterion, the Random Forest model achieved the highest F1-score, with a value of 0.6941 and a ROC-AUC of 0.9076, demonstrating the best performance among the evaluated models under the established criterion within the analyzed dataset. It was concluded that the proposed model constitutes a useful tool for supporting strategic decision-making aimed at member retention within the cooperative.

Keywords: machine learning, member attrition, supervised classification, savings and credit cooperative, predictive analytics.

Índice

Dedicatoria iii

Agradecimiento iv

Resumen v

Abstract vi

Lista de tablas x

Lista de figuras xi

Capítulo I 12

Introducción 12

1.1. Planteamiento del problema 12
1.2. Formulación del problema 15
1.2.1. Problema principal 15
1.2.2. Problemas secundarios 15
1.3. Justificación de la investigación 16
1.4. Limitaciones de la investigación 17
1.5. Objetivos de la investigación 19
1.5.1. Objetivo general 19
1.5.2. Objetivos específicos 19
1.6. Hipótesis de la investigación 20

Capítulo II 21

Marco teórico 21

2.1. Antecedentes de la investigación 21
2.1.1. Antecedentes internacionales 21
2.1.2. Antecedentes nacionales 22
2.2. Marco teórico 27
2.2.1. Fuga de socios 27
2.2.2. Comportamiento financiero 30
2.2.3. Machine Learning para predicción de fuga 34

Capítulo III 51

Material y métodos 51
3.1. Tipo de investigación 51
3.2. Nivel de investigación 52
3.3. Diseño de la investigación 53
3.4. Población y muestra 53
3.4.1. Población 53
3.4.2. Muestra 54
3.4.3. Criterios de inclusión y exclusión 54
3.5. Variables e indicadores 55
3.5.1. Definición conceptual de las variables 55
3.5.2. Definición operacional de las variables 56
3.6. Técnicas e instrumentos para el tratamiento de datos e información 57
3.6.1. Técnicas de recolección de datos 57
3.6.2. Instrumentos de recolección y procesamiento 58
3.6.3. Técnica de análisis de datos 59
3.6.4. Herramientas computacionales 60
3.7. Procedimiento metodológico para el desarrollo y evaluación del modelo 60
3.7.1. Entorno de ejecución y herramientas utilizadas 60
3.7.2. Extracción de datos desde BESTERP 62
3.7.3. Aplicación de criterios de inclusión y exclusión 62
3.7.4. Construcción del dataset final 64
3.7.5. Definición de la variable objetivo 64
3.7.6. Descripción de variables predictoras 65
3.7.7. Consideraciones éticas y anonimización 68
3.7.8. Preprocesamiento y Feature Engineering 68
3.7.9. Preparación de datos para modelamiento 71
3.7.10. Definición de modelos 72
3.7.11. Entrenamiento y validación cruzada 73
3.7.12. Evaluación de desempeño 74
3.7.13. Interpretabilidad mediante SHAP 75
Capítulo IV 77

Resultados y discusión 77

4.1. Caracterización del dataset final 77
4.2. Estadísticas descriptivas de las variables numéricas..... 78
4.3. Análisis Exploratorio de Datos (EDA)..... 79
4.4. Análisis de variables categóricas 89
4.5. Evaluación Comparativa de Modelos..... 95
4.6. Matrices de confusión 97
4.7. Selección del Modelo Óptimo 98
4.8. Interpretabilidad mediante SHAP..... 98
4.9. Discusión de resultados 101

Capítulo V 109

Conclusiones y recomendaciones 109

5.1. Conclusiones 109
5.2. Recomendaciones 110

Referencias bibliográficas 112

Anexos 119

Anexo 1. Matriz de consistencia 119
Anexo 2. Matriz de operacionalización de variables 121
Anexo 3. Dataset 126
Anexo 4. Código 127

Lista de tablas

Tabla 1. Tipo de fuga de socios 28
Tabla 2. Variables asociadas al comportamiento financiero 31
Tabla 3. Algoritmos empleados en la predicción de fuga de socios 38
Tabla 4. Librerías utilizadas 61
Tabla 5. Criterios de filtrado 62
Tabla 6. Definición de variables 64
Tabla 7. Variables demográficas y socioeconómicas 65
Tabla 8. Características bancarias del socio 66
Tabla 9. Variables de comportamiento financiero 67
Tabla 10. Variables derivadas construidas durante el feature engineering . 69
Tabla 11. Codificación de variables categóricas 71
Tabla 12. Preparación de datos 71
Tabla 13. Modelos definidos y sus hiper parámetros iniciales 72
Tabla 14. Resumen del Dataset 77
Tabla 15. Correlación de Pearson de variables numéricas con fuga_socio 88
Tabla 16. Tiempo de entrenamiento y métricas iniciales por modelo 95
Tabla 17. Desempeño comparativo de los modelos sobre el conjunto de test 96
Tabla 18. Matrices de confusión por modelo 97
Tabla 19. Reporte de clasificación del modelo seleccionado Random Forest 98
Tabla 20. Top 10 features según importancia 99

Lista de figuras

Figura 1. Cantidad de socios 2008 – 2024 15
Figura 2. Marco de trabajo para el desarrollo y evaluación del modelo de machine Learning 36
Figura 3. Clasificación oficial de variables predictoras 70
Figura 4. Estadística descriptiva de variables numéricas 79
Figura 5. Distribución de Churn 80
Figura 6. Distribución de fuga de socios por edad 81
Figura 7. Distribución de fuga de socios por ingresos mensuales 81
Figura 8. Distribución de fuga de socios por antigüedad 82
Figura 9. Distribución de fuga de socios por productos activos 83
Figura 10. Distribución de fuga de socios por frecuencia de uso mensual . 83
Figura 11. Distribución de fuga de socios por préstamos otorgados 84
Figura 12. Distribución de fuga de socios por mora en días 85
Figura 13. Distribución de fuga de socios por aportes totales 85
Figura 14. Distribución de fuga de socios por retiros 86
Figura 15. Distribución de fuga de socios por movimientos 87
Figura 16. Distribución de fuga de socios por sexo 90
Figura 17. Distribución de fuga de socios por estado civil 91
Figura 18. Distribución de fuga de socios por nivel educativo 91
Figura 19. Distribución de fuga de socios por ocupación 92
Figura 20. Distribución de fuga de socios por actividad económica 93
Figura 21. Distribución de fuga de socios por historial de pagos 94

Capítulo I

Introducción

1.1. Planteamiento del problema

En el contexto actual del sistema financiero cooperativo peruano, las cooperativas de ahorro y crédito enfrentan un entorno competitivo cada vez más exigente, en el cual la fidelización de sus socios representa un factor relevante para su sostenibilidad y crecimiento (Marquez, 2020). En particular, la Cooperativa de Ahorro y Crédito Santa María Magdalena ha experimentado variaciones en su cartera de socios y registra casos de fuga, fenómeno que puede generar efectos acumulativos en la rentabilidad, la liquidez y la eficiencia operativa de la entidad.

La fuga de socios se produce cuando estos formalizan su desvinculación de la cooperativa. En la presente investigación, dicho evento se identifica mediante el registro de renuncia voluntaria consignado en el sistema institucional. La ausencia de un modelo de clasificación que estime el riesgo de pertenencia a la clase fuga limita la capacidad de la cooperativa para reconocer patrones asociados a este fenómeno y establecer criterios objetivos para priorizar posibles acciones de retención. Actualmente, las decisiones relacionadas con la fidelización y el contacto con los socios se sustentan principalmente en análisis retrospectivos o criterios generales, sin disponer de una herramienta que procese sistemáticamente los patrones históricos y las características registradas de los socios (Arotinco & Gutierrez, 2025).

En ese sentido, se identifica una brecha relacionada con el aprovechamiento de las características bancarias, el comportamiento financiero y las características demográficas y socioeconómicas disponibles en los registros institucionales. Estas variables podrían emplearse para estimar la probabilidad de que un socio pertenezca a la clase fuga, de acuerdo con los patrones identificados en la información analizada (Alvarez, 2022). La falta de esta estimación dificulta establecer criterios objetivos de clasificación y priorización para orientar acciones diferenciadas dirigidas a los socios con mayor riesgo estimado.

Asimismo, la ausencia de una herramienta analítica basada en técnicas de machine learning restringe el aprovechamiento de los datos existentes en la cooperativa. Diversas investigaciones del sector financiero han aplicado algoritmos

de aprendizaje automático para clasificar clientes según su riesgo de fuga y reconocer las variables relacionadas con dicho comportamiento, obteniendo resultados favorables en términos de desempeño predictivo (Galvez & Uribe, 2024).

Por ello, surge la necesidad de desarrollar y comparar modelos de clasificación supervisada que permitan estimar el riesgo de fuga de los socios de la Cooperativa de Ahorro y Crédito Santa María Magdalena, a partir del procesamiento de la información institucional disponible hasta el 31 de diciembre de 2024. En este estudio, la predicción se entiende como la estimación de la probabilidad de pertenencia a las clases fuga o no fuga, de acuerdo con las características observadas en el conjunto de datos. Por tanto, no comprende el pronóstico del momento futuro en el que ocurriría la desvinculación ni la evaluación longitudinal del comportamiento del socio.

La fuga de socios no solo representa una reducción de la base social de la cooperativa, sino también un problema económico y operativo, debido a que obliga a destinar nuevamente recursos para reemplazar a los socios que se desvinculan. Durante el periodo analizado se registraron aproximadamente 6 763 renuncias voluntarias, equivalentes al 12,2 % del conjunto de socios evaluados. Asimismo, se estima que el número de nuevas afiliaciones anual es alrededor de 10 % superior al número de renuncias, lo que representa aproximadamente 7 439 nuevos socios para el periodo en evaluación. Bajo esta relación, cerca del 91 % de las nuevas afiliaciones habría servido únicamente para compensar las salidas ocurridas durante el año, mientras que el crecimiento neto aproximado habría sido de solo 676 socios.

La incorporación de nuevos socios requiere la ejecución de diversas acciones de captación, entre ellas campañas en radio y televisión local, difusión mediante redes sociales, charlas informativas en instituciones, campañas promocionales en fechas conmemorativas y captación indirecta a través de los analistas de crédito. Aunque la cooperativa no dispone actualmente de un cálculo consolidado del costo de estas actividades, su ejecución demanda tiempo del personal, recursos publicitarios, infraestructura y capacidad administrativa.

Adicionalmente, cada proceso de afiliación requiere entre 30 y 60 minutos para la revisión de documentos, registro de información, procesamiento del aporte inicial y orientación al nuevo socio. Considerando aproximadamente 7 439 afiliaciones anuales, esta actividad representa entre 3 720 y 7 439 horas de trabajo administrativo, sin incluir el esfuerzo previo de promoción y captación. En condiciones

regulares, el socio realiza un pago obligatorio inicial de S/ 30, compuesto por S/ 10 por derecho de inscripción y S/ 20 como aporte mínimo inicial. Durante las aproximadamente diez campañas promocionales efectuadas al año, la cooperativa exonera definitivamente el derecho de inscripción y solicita únicamente el aporte mínimo inicial, con la finalidad de reducir la barrera de ingreso y facilitar la captación.

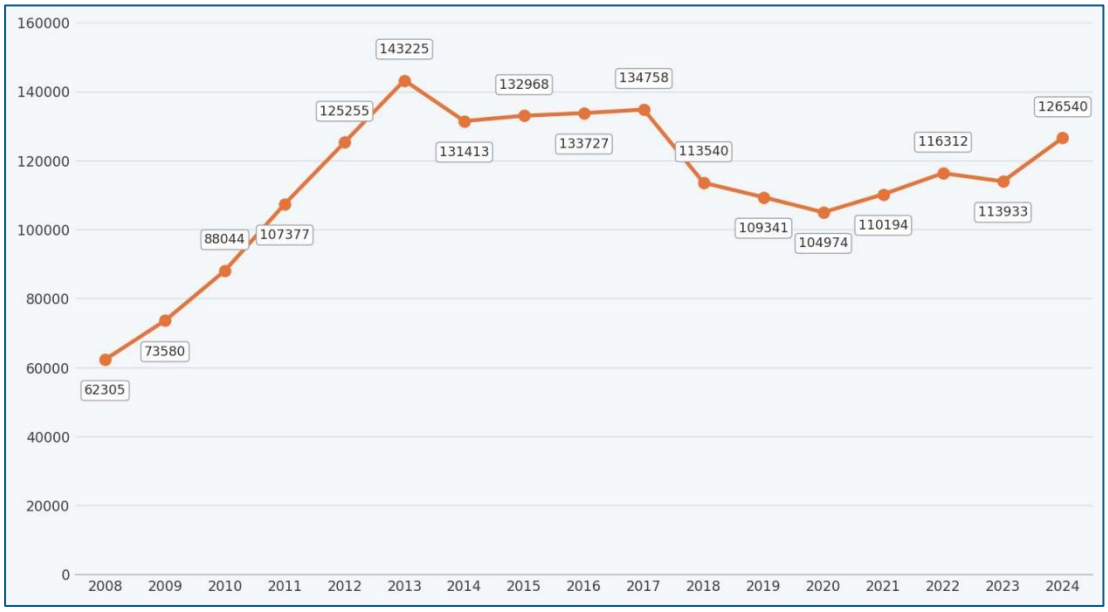
En contraste, la retención de un socio existente aprovecha una relación previamente establecida, un expediente ya registrado y un historial financiero disponible. La acción de retención considerada por la cooperativa consistiría principalmente en una llamada personalizada y en la presentación de productos o campañas comerciales acordes con las características y necesidades del socio. Este tipo de intervención requiere menos recursos operativos que ejecutar nuevamente todo el proceso de promoción, captación, documentación y afiliación necesario para reemplazar a un socio que renuncia.

En consecuencia, la identificación de socios con señales asociadas a la fuga con señales de desvinculación permitiría dirigir los esfuerzos de retención hacia grupos específicos, en lugar de aplicar campañas generales sobre toda la base social. Desde esta perspectiva, el desarrollo de un modelo de machine learning presenta una utilidad económica potencial, al contribuir a reducir la necesidad de reemplazar socios perdidos y mejorar la asignación de los recursos institucionales destinados a captación y fidelización.

Los resultados permitirán identificar el modelo con mejor desempeño dentro del conjunto de datos analizado, así como las variables con mayor contribución a la clasificación. De esta manera, se busca proporcionar una herramienta analítica de soporte para la toma de decisiones y el diseño posterior de estrategias de retención. Sin embargo, la investigación no evalúa la implementación de estas estrategias ni demuestra una reducción efectiva de la fuga de socios.

Figura 1.

Cantidad de socios 2008 – 2024



Nota. Fuente: Cooperativa de ahorro y crédito Santa María Magdalena

1.2. Formulación del problema

1.2.1. Problema principal

¿Qué modelo de machine learning de clasificación supervisada, desarrollado a partir de características bancarias y comportamientos financieros de los socios, presenta el mejor desempeño para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025?

1.2.2. Problemas secundarios

a) ¿Qué características bancarias de los socios tienen mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025?

b) ¿Qué comportamientos financieros de los socios tienen mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025?

c) ¿Qué variables explican con mayor peso las predicciones del modelo de mejor desempeño, según el análisis de interpretabilidad SHAP, en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025?

1.3. Justificación de la investigación

La presente investigación se justifica desde el punto de vista teórico porque contribuye a la aplicación del conocimiento relacionado con la analítica de datos y el machine learning en el sector financiero cooperativo. Específicamente, aborda el fenómeno de la fuga de socios, el cual ha sido estudiado principalmente en contextos bancarios tradicionales, pero presenta todavía un desarrollo limitado en el ámbito de las cooperativas de ahorro y crédito, cuyas dinámicas de comportamiento financiero y vinculación institucional poseen características particulares.

Asimismo, el estudio aporta evidencia empírica sobre la contribución de las variables transaccionales, bancarias, demográficas y socioeconómicas en la clasificación del riesgo de fuga. En este sentido, permite evaluar la utilidad de la ingeniería de características basada en el conocimiento del dominio para construir indicadores derivados que representen patrones relevantes dentro del conjunto de datos institucional.

Desde el punto de vista metodológico, la investigación desarrolla un flujo estructurado de machine learning que comprende el tratamiento de los datos, la ingeniería y selección de características, el manejo del desbalance de clases, el entrenamiento de algoritmos y la evaluación de su desempeño. Para ello, se emplean técnicas como SMOTE, una partición aleatoria estratificada y validación cruzada estratificada, aplicadas de acuerdo con el procedimiento metodológico establecido. Estos recursos favorecen una evaluación consistente y comparable de los modelos dentro del conjunto de datos analizado, sin que ello implique garantizar su generalización temporal a periodos posteriores.

Además, se comparan diferentes algoritmos de clasificación, entre ellos Logistic Regression, Random Forest, Gradient Boosting, XGBoost y LightGBM. Esta comparación permite seleccionar el modelo con mejor desempeño de acuerdo con métricas previamente definidas, como F1-Score y ROC-AUC. El procedimiento fortalece la evaluación interna de los modelos y proporciona una metodología susceptible de ser replicada en investigaciones similares, siempre que se consideren las características de cada conjunto de datos y contexto institucional.

Asimismo, se incorpora un análisis de interpretabilidad mediante valores SHAP, con la finalidad de estimar la contribución de las variables a las predicciones generadas por el modelo seleccionado. Este análisis permite reconocer qué

características influyen con mayor peso en la clasificación de los socios dentro de las categorías fuga o no fuga, sin atribuir relaciones causales a las asociaciones identificadas.

En el ámbito práctico, la investigación proporciona una herramienta analítica para clasificar a los socios según su probabilidad estimada de pertenecer a la clase fuga, de acuerdo con los patrones contenidos en los registros históricos disponibles hasta el 31 de diciembre de 2024. Los resultados pueden servir como insumo para establecer criterios objetivos de priorización y diseñar posteriormente acciones diferenciadas de contacto, fidelización o retención.

El modelo permite transformar los datos históricos en información de apoyo para la toma de decisiones. La identificación de patrones relacionados con los retiros, los aportes, la frecuencia de uso, la vinculación con productos financieros o el comportamiento de pago puede orientar el análisis de los perfiles que presentan mayor riesgo estimado. Sin embargo, la investigación no evalúa la implementación de estrategias de retención ni demuestra que el uso del modelo reduzca efectivamente la fuga de socios.

Finalmente, la investigación se justifica tecnológicamente porque evalúa la viabilidad de desarrollar modelos de clasificación mediante herramientas de código abierto como scikit-learn, XGBoost y LightGBM, reduciendo la necesidad de licencias especializadas. El modelo seleccionado presenta potencial para ser incorporado posteriormente en sistemas analíticos o plataformas de apoyo a la toma de decisiones de la cooperativa. No obstante, antes de una aplicación operativa en periodos futuros será necesario realizar una validación temporal, monitorear su desempeño y establecer procedimientos de recalibración o reentrenamiento frente a posibles cambios en los datos.

1.4. Limitaciones de la investigación

Limitaciones teóricas

Desde el punto de vista teórico, la investigación se circunscribe al enfoque de machine learning supervisado aplicado a la predicción de fuga de socios (churn), considerando principalmente variables demográficas, transaccionales y de relación institucional disponibles en los sistemas internos de la cooperativa. En ese sentido, no se abordan otros enfoques teóricos complementarios como modelos

econométricos causales, análisis cualitativos del comportamiento del socio o teorías de satisfacción y lealtad del usuario.

Asimismo, el estudio se enfoca en algoritmos de aprendizaje automático tradicionales (como Random Forest, XGBoost y otros modelos de clasificación), sin considerar técnicas más avanzadas como deep learning o modelos híbridos, lo que delimita el alcance del análisis al estado del arte aplicado y accesible en contextos organizacionales similares.

Limitaciones temporales

La investigación se desarrolló a partir de registros históricos disponibles hasta el 31 de diciembre de 2024, analizados mediante un diseño transversal retrospectivo. En consecuencia, los datos representaron las características de los socios y su pertenencia a las clases fuga o no fuga dentro del periodo considerado, sin efectuar un seguimiento longitudinal posterior.

En este estudio, el término predicción se utilizó en el sentido propio del aprendizaje supervisado: la asignación de una clase o la estimación de la probabilidad de pertenencia a la clase fuga a partir de las variables observadas. Por tanto, el modelo no estimó el momento futuro exacto en el que ocurriría la desvinculación ni demostró capacidad para pronosticar temporalmente dicho evento.

Para entrenar y evaluar los modelos se aplicó una partición aleatoria estratificada del conjunto de datos. No se realizó una partición cronológica ni una validación out-of-time con información correspondiente a periodos posteriores a la fecha de corte. En consecuencia, las métricas reportadas expresan el desempeño de los modelos dentro del conjunto de datos disponible y bajo condiciones semejantes a las observadas hasta el 31 de diciembre de 2024.

La ausencia de validación temporal limita la posibilidad de afirmar que el desempeño obtenido se mantendrá sin variaciones en periodos futuros. Los cambios en el comportamiento de los socios, la estacionalidad, las campañas institucionales, las modificaciones en los productos financieros o las condiciones económicas podrían alterar las relaciones identificadas por el modelo. Por ello, una eventual aplicación operativa requerirá monitorear periódicamente su desempeño, verificar posibles cambios en la distribución de los datos y efectuar procesos de recalibración o reentrenamiento cuando resulte necesario.

Esta limitación no invalida la comparación realizada entre los algoritmos dentro del conjunto de datos estudiado, pero restringe la generalización temporal de sus resultados. La validación out-of-time y la evaluación prospectiva del modelo se plantean como líneas de investigación posteriores.

Limitaciones espaciales

Desde la perspectiva espacial, el estudio se limita a la Cooperativa de Ahorro y Crédito Santa María Magdalena, utilizando exclusivamente información de sus socios registrados en su sistema institucional. Por tanto, los resultados obtenidos responden a las características propias de esta organización, tales como su estructura operativa, perfil de socios y contexto socioeconómico.

En consecuencia, la generalización de los resultados a otras cooperativas, entidades financieras o regiones debe realizarse con cautela, considerando las diferencias en comportamiento del socio, oferta de productos y condiciones del entorno.

1.5. Objetivos de la investigación

1.5.1. Objetivo general

Desarrollar y comparar modelos de machine learning de clasificación supervisada, basados en características bancarias y comportamientos financieros de los socios, para identificar el modelo con mejor desempeño en la predicción del riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.

1.5.2. Objetivos específicos

a) Determinar las características bancarias de los socios con mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.
b) Determinar los comportamientos financieros de los socios con mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.
c) Interpretar las predicciones del modelo de mejor desempeño mediante el análisis de importancia de variables SHAP, para identificar los factores con mayor influencia en la predicción del riesgo de fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.

1.6. Hipótesis de la investigación

La presente investigación no formula hipótesis, debido a que se orienta al desarrollo, la evaluación y la comparación de modelos de clasificación supervisada, y no a la comprobación de relaciones causales entre las variables estudiadas. El estudio analiza registros históricos de los socios, entrena diferentes algoritmos de machine learning y compara su capacidad para asignar la clase y estimar la probabilidad de pertenencia a las categorías fuga o no fuga.

La selección del modelo con mejor desempeño se realiza de acuerdo con métricas de clasificación previamente establecidas, aplicadas sobre el conjunto de prueba obtenido mediante una partición aleatoria estratificada. Por tanto, los resultados expresan el desempeño comparativo alcanzado dentro del conjunto de datos analizado y no constituyen una contrastación inferencial de hipótesis ni un pronóstico temporal de futuras desvinculaciones.

Capítulo II

Marco teórico

2.1. Antecedentes de la investigación

2.1.1. Antecedentes internacionales

En el estudio realizado por Marquez (2020) la cual titula “Identificación de clientes que realizaron fuga de equipos móviles en una empresa de telecomunicaciones utilizando el algoritmo random forest” se planteó como objetivo principal diseñar e implementar un proceso de monitoreo automatizado que permita identificar, de forma anticipada al monitoreo actual, a los socios que hicieron fuga de equipos móviles en una empresa de telecomunicaciones del Perú utilizando el algoritmo Random Forest. La metodologia fue aplicada con un diseño no experimental con un nivel descriptivo. Los resultados fueron que se logró diseñar e implementar un proceso de monitoreo automatizado para identificar a los socios que hicieron fuga de equipos móviles en una empresa de telecomunicaciones del Perú utilizando el algoritmo Random Forest. Se logró reducir la cantidad de días del monitoreo actual de 45 días a 25 días sin perder información relevante sobre el tráfico telefónico de los socios, al momento de identificar a aquellos que hicieron fuga. Las principales variables que permiten identificar a los socios fugadores en una ventana de 25 días son: “prom_min_sal”, “prom_llam_ent”, “prom_sms”, “Marca_cat” y “Precio”. Se concluyó que el modelo basado en random forest permite identificar de manera precisa y anticipada a los socios que harán fuga de equipos móviles.

El estudio contribuye metodológicamente a la presente investigación al demostrar el valor del algoritmo Random Forest para la identificación precisa y anticipada de socios que realizan fuga, además de resaltar la importancia de un monitoreo automatizado y eficiente de eventos transaccionales. El abordaje de reducción en el tiempo de monitoreo sin pérdida de información relevante representa una mejora significativa en procesos de toma de decisiones preventivas, aspecto crucial tanto en telecomunicaciones como en el sector financiero cooperativo. Sin embargo, la investigación está enfocada en la fuga de equipos móviles en el mercado de telecomunicaciones, por lo que no considera variables contextuales ni comportamientos asociados a los socios de una cooperativa de ahorro y crédito. Esta limitación abre la oportunidad de adaptar el modelo a un sector diferente, integrando variables propias del ámbito cooperativo, para fortalecer la anticipación y precisión en la gestión de la retención de socios.

En la investigación llevada a cabo por Khalilbayli y Sertkaya (2025), titulada “Enhancing Customer Churn Prediction in the Finance Sector through Explainable AI and Machine Learning”, planteó como objetivo investigar cómo la integración de metodologías de inteligencia artificial explicable (XAI) puede mejorar los modelos predictivos de fuga de socios en el sector financiero, abordando el desafío de la transparencia en la toma de decisiones. La metodología del estudio evaluó el desempeño y la interpretabilidad de varios algoritmos de machine learning —entre ellos Random Forest, XGBoost, Light-GBM y CatBoost—, utilizando técnicas de explicabilidad como SHAP (Shapley Additive Explanations). El caso de estudio se enfocó en la empresa financiera Eminevim, basándose en conjuntos de datos complejos y un enfoque comparativo para identificar los métodos más efectivos. Los resultados fueron que los modelos predictivos de fuga de socios, complementados con metodologías de XAI, lograron mantener alta precisión y, a la vez, mejoraron significativamente la transparencia y comprensión de las decisiones del modelo. Esto permitió a las instituciones financieras tomar decisiones informadas y fundamentadas para reducir la deserción, optimizando la sostenibilidad del negocio a largo plazo. Se concluyó que la incorporación de inteligencia artificial explicable en los modelos de predicción de fuga de socios representa una solución efectiva a la problemática de la "caja negra" en machine learning, empoderando a las empresas financieras para mitigar la fuga de socios mediante estrategias basadas en datos y evidencias.

Este antecedente aborda directamente la predicción de la fuga de socios en el ámbito financiero mediante técnicas avanzadas de machine learning y explicabilidad. La utilización de algoritmos de última generación junto con métodos interpretables como SHAP, valida la relevancia de integrar transparencias y fundamentos técnicos en la gestión de la retención de socios. Si bien el estudio enfoca su aplicación en una entidad financiera específica, sus hallazgos y enfoque metodológico pueden transferirse y adaptarse de manera efectiva al contexto de cooperativas de ahorro y crédito en el Perú, fortaleciendo la argumentación del uso de modelos interpretables y precisos para la predicción y gestión de la fuga de socios.

2.1.2. Antecedentes nacionales

En la investigación desarrollada por Bernachea et al. (2021), titulada "Técnica de Machine Learning para el cálculo de la probabilidad de fuga de los clientes de la empresa Bitel", se propuso como objetivo principal la implementación de un algoritmo de aprendizaje automático orientado a estimar la probabilidad de abandono por parte

de los socios de la mencionada empresa de telecomunicaciones. El estudio adoptó un enfoque cuantitativo con diseño no experimental, técnica de regresión logística aplicada a clientes de telecomunicaciones. Durante el proceso metodológico, se integraron diversas bases de datos que contenían información relevante de los usuarios, como el consumo de datos móviles en megabytes, el uso de minutos de llamada, el envío de mensajes de texto (SMS), así como variables demográficas (sexo, uso de la aplicación MiBitel, entre otras). Inicialmente se obtuvo una muestra de 16,512 registros, sin embargo, al analizar la calidad de los datos se identificaron campos nulos en las variables "Sexo" y "Tipo de Dispositivo", las cuales presentaban valores faltantes en aproximadamente el 47% de los casos. Tras la depuración de registros incompletos, se conservó una base final de 8,572 entradas válidas. Posteriormente, se procedió a la carga de los datos mediante archivos en formato CSV en el entorno Jupyter Notebook, utilizando la biblioteca Pandas de Python. En esta etapa se eliminaron variables que mostraban alta correlación entre sí, se aplicó la normalización a los datos cuantitativos y se transformaron variables categóricas en variables dummy, con el propósito de optimizar el entrenamiento del modelo predictivo. Finalmente, se concluyó que el modelo construido logró integrar exitosamente variables cualitativas relacionadas con el consumo y comportamiento transaccional del socio, así como datos demográficos, generando una solución efectiva para la predicción de la fuga de socios.

Este estudio resulta relevante al demostrar la eficacia del uso de técnicas de machine learning en la predicción de fuga de socios dentro empresas del sector telecomunicaciones, lo que evidencia la aplicabilidad de estos métodos en contextos de grandes volúmenes de datos y variables transaccionales. El enfoque en la depuración y preprocesamiento de datos, así como la transformación de variables, proporciona una metodología sólida que puede ser adaptada para el sector financiero cooperativo. Sin embargo, el trabajo se limita a una empresa de telecomunicaciones y no aborda particularidades propias de cooperativas de ahorro y crédito, como la tipología de socios, servicios exclusivos y variables asociadas al comportamiento financiero. Por tanto, este antecedente fundamenta la utilidad del aprendizaje automático para el problema a resolver, pero resalta la oportunidad de explorar su desempeño en contextos cooperativos, cubriendo así una brecha metodológica y sectorial que se plantea en la presente investigación.

En la investigación desarrollada por Gallardo (2021), titulada “Modelo de predicción de fuga de clientes en empresa SaaS de inteligencia logística”, se propuso como objetivo principal construir un modelo predictivo orientado a estimar la probabilidad de fuga de socios, con el propósito de generar alertas tempranas que permitan atender oportunamente las necesidades no satisfechas de los usuarios. La metodología adoptada fue de tipo aplicada, con un diseño experimental y un enfoque cuantitativo. Entre los principales resultados, se destaca la implementación y evaluación de dos modelos de clasificación: XGBoost y regresión logística (logit), entrenados bajo diez configuraciones distintas de variables. El modelo XGBoost alcanzó un rendimiento de ROC_AUC de 0.76, una sensibilidad de 0.56 y una precisión de 0.22; mientras que el modelo logit obtuvo un ROC_AUC de 0.72, sensibilidad de 0.76 y precisión de 0.04. Estos resultados evidencian un desempeño aceptable en contextos de clasificación con datos desbalanceados, donde se comprobó la efectividad de aplicar técnicas de ponderación de errores para mejorar la capacidad predictiva. En conclusión, el estudio permitió desarrollar un modelo predictivo funcional para la detección de fuga de socios en una empresa de servicios logísticos basados en software como servicio (SaaS), haciendo uso de un conjunto de datos reducido y altamente desbalanceado, compuesto por 7 variables, 2372 registros, 235 socios y 38 casos de fuga. Además, se resaltó la importancia de comprender la influencia de las variables utilizadas en el comportamiento del modelo predictivo.

La investigación al abordar la predicción de fuga de socios en un entorno SaaS con retos de datos reducidos y desbalanceados, retos que también pueden presentarse en el sector cooperativo. El análisis comparativo de modelos (XGBoost y regresión logística), así como la aplicación de técnicas de ponderación para mejorar el rendimiento en la clasificación, proporciona enfoques útiles que pueden ser transferidos y adaptados a la predicción de fuga de socios en cooperativas de ahorro y crédito. No obstante, el estudio se limita a empresas de inteligencia logística y socios corporativos, sin explorar variables propias de aplicaciones financieras cooperativas ni relaciones de largo plazo con socios. Por tanto, este trabajo fundamenta metodologías valiosas y resalta la necesidad de desarrollar modelos que sean robustos incluso ante bases de datos con baja representatividad, abriendo la oportunidad de aplicar y validar estas estrategias en el ámbito cooperativo.

En la investigación llevada a cabo por Segura (2022), titulada “Desarrollo de un modelo de predicción de fuga de clientes y diseño de experimento para la aplicación de estrategias de fidelización en factoring”, se planteó como objetivo fundamental identificar de forma proactiva a los socios activos con alta probabilidad de deserción, con el fin de incrementar el volumen de socios recurrentes. Para ello, se propuso la aplicación de modelos predictivos que sirvieran como base para implementar estrategias comerciales específicas, complementadas con diseños experimentales. La metodología fue de tipo aplicada, con diseño experimental y enfoque cuantitativo. En cuanto a los resultados, se evidenció que incentivar el retorno de socios implica asumir una reducción en las utilidades a corto plazo. Para sustentar esta afirmación, se llevó a cabo una simulación de la implementación del diseño experimental, la cual mostró una disminución promedio del 10.7 % en las utilidades de la empresa, equivalente a $7,268,443. No obstante, esta estrategia resultó efectiva en términos de retención, al lograrse un incremento del 6.1 % en la cantidad de socios activos, equivalente a 28 socios adicionales en comparación con el escenario sin intervención. Dado que los beneficios en el corto plazo no compensan completamente la inversión realizada, se extendió el análisis a un escenario de mediano plazo (12 meses), mediante una nueva simulación. En este caso, se aplicó una oferta selectiva a un 20 % de los socios identificados por el modelo como los más propensos a desertar. Como resultado, se obtuvo un incremento del 3.3 % en las utilidades acumuladas ($28,545,380) y un crecimiento del 7.2 % en el número acumulado de socios activos (34 socios adicionales). En conclusión, el estudio permitió el desarrollo de un modelo predictivo robusto que no solo anticipa la fuga de socios en el sector de factoring, sino que también sustenta la formulación e implementación de estrategias de fidelización eficaces, con impacto medible en la rentabilidad y la retención de socios a mediano plazo.

El trabajo aporta una visión integral de la predicción de fuga de socios combinada con la implementación de estrategias de fidelización y análisis de impacto financiero en el sector de factoring. La principal contribución radica en el procedimiento de vincular los resultados de modelos predictivos con acciones comerciales concretas, así como en el uso de simulaciones para evaluar la rentabilidad de tales estrategias en diferentes horizontes temporales. Sin embargo, el estudio se centra en empresas de factoring y no abarca particularidades del sector cooperativo, como la segmentación de socios y la diversidad de servicios financieros ofrecidos. Esto deja una oportunidad significativa: adaptar y profundizar en el análisis

del impacto de estrategias de retención basadas en predicción, específicamente para cooperativas de ahorro y crédito, considerando sus modelos de negocio y relaciones con socios, aspecto que busca cubrir la presente investigación.

En el trabajo de investigación desarrollado por Arotinco y Gutiérrez (2025), titulado “Modelo de predicción de deserción de clientes en el sector de aseguradoras peruanas de pólizas vehiculares mediante el uso del algoritmo Random Forest”, se formuló como objetivo principal asegurar la validez y consistencia de un modelo predictivo enfocado en estimar la tasa de deserción de socios en una aseguradora vehicular, priorizando la minimización de la discrepancia relativa entre los resultados obtenidos en las fases de prueba y los registrados en el entorno real. La investigación fue de tipo aplicada y se desarrolló bajo un enfoque cuantitativo. En cuanto a los resultados, en la primera fase, correspondiente a la validación con datos históricos, el modelo alcanzó métricas significativas: una precisión del 88.89%, un recall del 94.12%, un F1-score de 91.43% y un área bajo la curva (AUC) de 92.71%. En la segunda fase, orientada a la validación en condiciones operativas reales, se evidenciaron mejoras sustanciales, logrando una exactitud del 97.08%, recall del 97.76% y un AUC del 97.34%. Estos resultados reflejaron una alta capacidad de adaptación del modelo al contexto real de negocio, reforzando su utilidad para predecir de manera efectiva la deserción de socios. Se concluyó que el modelo propuesto constituye una herramienta predictiva sólida, validada tanto cuantitativa como cualitativamente, y cuya aplicación contribuye de manera significativa a optimizar las estrategias de retención en el sector asegurador.

La investigación realizada es de gran utilidad, ya que ejemplifica cómo los modelos de machine learning en particular el algoritmo Random Forest pueden ser adaptados para la predicción de deserción de socios en sectores afines al financiero mediante validaciones robustas con datos históricos y operativos. El énfasis en la validación real y la comparación de métricas entre fases aporta una metodología aplicable directamente a la problemática tratada en cooperativas de ahorro y crédito. No obstante, el estudio se enfoca exclusivamente en aseguradoras de pólizas vehiculares, dejando abierta la oportunidad de explorar la adaptabilidad de estos modelos para diferentes sectores financieros y tipos de socios, como en el caso de socios de cooperativas. Así, este antecedente fortalece la justificación metodológica de la presente investigación y resalta la necesidad de evaluar el desempeño de

técnicas predictivas en nuevos escenarios sectoriales, lo cual constituye precisamente el aporte de esta propuesta.

En síntesis, los antecedentes revisados evidencian la aplicación de diferentes algoritmos de clasificación para estimar el riesgo de fuga en sectores como telecomunicaciones, seguros, *factoring*, servicios financieros y empresas de software. Entre los algoritmos empleados con mayor frecuencia se encuentran Random Forest, XGBoost y regresión logística, mientras que otros estudios incorporaron modelos como LightGBM y CatBoost. Para evaluar su desempeño se utilizaron principalmente métricas como precisión, sensibilidad o *recall*, F1-score y ROC-AUC. En los estudios que informaron valores numéricos de ROC-AUC, los resultados oscilaron entre 0,72 y 0,9734, aunque estas cifras deben interpretarse considerando las diferencias en las bases de datos, la definición de la fuga, el grado de desbalance y los procedimientos de validación utilizados. Asimismo, los trabajos revisados muestran la importancia del preprocesamiento, la ingeniería de características, el tratamiento del desbalance y la interpretación de las variables que contribuyen a las predicciones. Sin embargo, se observa una limitada evidencia sobre la aplicación de estos modelos en cooperativas de ahorro y crédito peruanas y no se identificaron estudios desarrollados específicamente en Ayacucho. Frente a esta brecha, la presente investigación compara algoritmos de clasificación aplicados a los registros históricos de una cooperativa de ahorro y crédito, incorpora PR-AUC como métrica complementaria ante el desbalance de clases y utiliza valores SHAP para interpretar la contribución de las variables a las estimaciones del modelo, sin atribuirles relaciones causales.

2.2. Marco teórico

2.2.1. Fuga de socios

2.2.1.1. Definición de fuga de socios

La fuga de socios, también conocida como churn o deserción de socios, se refiere al fenómeno mediante el cual un socio que previamente mantenía una relación activa con una organización, decide interrumpirla de manera voluntaria o involuntaria. Según García (2021), en el contexto de las entidades financieras, especialmente en las cooperativas de ahorro y crédito, la fuga se manifiesta cuando un socio deja de utilizar los productos y servicios financieros ofrecidos, como depósitos, créditos, tarjetas o cualquier otro instrumento financiero, ya sea

por inactividad prolongada, cancelación de cuentas o traslado a otra institución financiera.

Este comportamiento implica una pérdida directa de ingresos para la entidad y representa un riesgo considerable en términos de sostenibilidad comercial, dado que captar nuevos socios suele requerir más recursos que mantener a los actuales. Según Vicente (2023) identificar tempranamente los indicios de una posible fuga es esencial para implementar acciones correctivas que favorezcan la retención del socio.

Desde una perspectiva analítica, la fuga de socios se puede conceptualizar como una variable binaria que clasifica a los usuarios en dos categorías: aquellos que han abandonado la entidad y aquellos que continúan operando con ella. Esta distinción permite aplicar herramientas estadísticas y de aprendizaje automático para predecir con anticipación qué socios tienen mayor probabilidad de abandonar la organización.

Las estrategias tradicionales de retención de socios en el ámbito financiero se centran en evitar el abandono mediante el fortalecimiento de la relación con el socio, la mejora del servicio y la personalización de las soluciones ofrecidas (Vásquez y Ramirez, 2025). Estas estrategias han sido empleadas por años y, si bien algunas presentan limitaciones frente a enfoques más modernos basados en analítica avanzada, siguen siendo fundamentales dentro de los programas de fidelización.

2.2.1.2. Tipos de fuga de socios

En las entidades financieras, la fuga de socios puede clasificarse en diferentes tipos, dependiendo de la causa, el grado de participación del socio y el tipo de salida que se produce (Alvarez, 2022). La siguiente tabla resume las principales categorías:

Tabla 1.

Tipo de fuga de socios

Tipo de Fuga	Descripción	Ejemplo en Cooperativas Financieras
Fuga Voluntaria	Ocurre cuando el socio decide por iniciativa propia dejar de utilizar los servicios.	Un socio cancela su cuenta de ahorro por no estar satisfecho con las tasas de interés.

Fuga Involuntaria	El socio deja la entidad por factores ajenos a su voluntad o por decisiones internas de la entidad.	Suspensión de servicios por incumplimiento de pagos o fallecimiento del socio.
Fuga Total	El socio interrumpe completamente toda relación comercial con la entidad.	Cierra todas sus cuentas, cancela sus productos y no mantiene contacto con la cooperativa.
Fuga Parcial	El socio reduce su nivel de vinculación, pero mantiene algún tipo de relación residual.	Mantiene una cuenta inactiva, pero transfiere su actividad principal a otro banco.
Fuga Activa	El | comunica de forma explícita su intención de abandonar la entidad.	Envía una carta solicitando el cierre de su cuenta de ahorro y tarjeta de crédito.
Fuga Pasiva o Silenciosa	El socio deja de utilizar los servicios sin notificar su salida.	Deja de realizar operaciones y transacciones durante meses sin formalizar su desvinculación.
Fuga Preventiva (Latente)	Aún no ha ocurrido la salida, pero se identifican señales tempranas de posible abandono.	Disminución de movimientos financieros, consultas frecuentes sobre costos o beneficios.

Nota. Elaboración propia con base en (Álvarez, 2022).

2.2.1.3. Causas comunes de la fuga en entidades financieras

La fuga de socios en el sector financiero es un fenómeno complejo que responde a múltiples factores, tanto internos como externos a la organización. Identificar las causas permite no solo comprender el comportamiento de abandono, sino también diseñar estrategias de retención más efectivas (Galvez y Uribe, 2024). A continuación, se describen las principales causas identificadas en el ámbito financiero:

- **Falta de satisfacción con los productos o servicios:** Una de las causas más recurrentes de fuga es la percepción negativa del socio respecto a la calidad de los servicios brindados. Esto puede incluir lentitud en la atención, procedimientos engorrosos, falta de personal capacitado o productos financieros que no se ajustan a sus necesidades.
- **Altos costos o comisiones poco competitivas:** Cuando los costos de mantenimiento, tasas de interés, o comisiones asociadas a productos financieros son superiores a los de otras entidades del mercado, los socios tienden a migrar hacia alternativas más económicas.
- **Deficiente atención al socio:** La falta de respuesta oportuna, el trato impersonal o la resolución ineficiente de quejas generan desconfianza y descontento. En un entorno donde la experiencia del socio es determinante, este aspecto se vuelve crucial.
- **Avances tecnológicos insuficientes:** Los socios actuales demandan servicios digitales ágiles y seguros. La ausencia de canales digitales

funcionales, como aplicativos móviles o plataformas de banca en línea eficientes, limita la conveniencia y empuja a los socios hacia competidores más innovadores.

- **Cambios en la situación financiera o personal del socio:** La pérdida de empleo, cambios de residencia, o decisiones personales también pueden influir en la desvinculación de una entidad financiera, aunque no estén necesariamente relacionadas con deficiencias del servicio.
- **Oferta más atractiva de la competencia:** La captación agresiva por parte de otras entidades mediante promociones, beneficios adicionales o servicios personalizados puede influir significativamente en la decisión del socio de abandonar su institución actual.
- **Falta de personalización en la oferta financiera:** Cuando la entidad no realiza una segmentación adecuada de su cartera de socios, ni adapta sus productos o comunicaciones según sus perfiles, se produce una desconexión que puede culminar en la fuga.
- **Pérdida de confianza institucional:** Factores como escándalos financieros, problemas de transparencia o sanciones del ente regulador pueden dañar la imagen institucional y generar temor en los socios, provocando su retiro.

2.2.2. Comportamiento financiero

2.2.2.1. Definición de comportamiento financiero

Como menciona García et al. (2021), el comportamiento financiero hace referencia al conjunto de decisiones, actitudes, hábitos y patrones de conducta que adopta una persona o entidad frente a la administración de sus recursos económicos. Este comportamiento incluye aspectos como el ahorro, el gasto, la inversión, el endeudamiento y el cumplimiento de obligaciones financieras, los cuales están influenciados tanto por factores racionales como emocionales.

En el caso de los socios de entidades financieras, el comportamiento financiero se refleja en la forma en que interactúan con los productos y servicios que les son ofrecidos, la frecuencia con que realizan operaciones, su puntualidad en el pago de créditos, la disposición al ahorro, entre otros elementos. Estos comportamientos no solo permiten comprender mejor al socio, sino también anticipar posibles riesgos como la morosidad o la decisión de abandonar la institución financiera.

2.2.2.2. Variables asociadas al comportamiento financiero

Las variables asociadas al comportamiento financiero permiten cuantificar y caracterizar las acciones y decisiones de los socios respecto al uso de productos y servicios financieros. Estas variables facilitan la identificación de patrones que pueden ser utilizados para predecir comportamientos futuros, como el abandono de la entidad o la contratación de nuevos productos.

Las variables pueden clasificarse en diferentes categorías, tales como transaccionales, de uso de productos, de endeudamiento, de cumplimiento, entre otras (García et al., 2021). A continuación, se presenta una tabla que resume las principales variables asociadas al comportamiento financiero:

Tabla 2.

Variables asociadas al comportamiento financiero

Categoría	Variable	Descripción
Transaccional	Frecuencia de transacciones	Número de operaciones financieras realizadas en un periodo determinado.
	Monto promedio de transacciones	Valor promedio de los movimientos financieros efectuados.
Uso de productos	Diversidad de productos contratados	Cantidad de productos financieros utilizados por el socio (cuentas, créditos, seguros, etc.).
	Tiempo de permanencia con cada producto	Duración del uso continuo de un producto financiero específico.
Endeudamiento	Nivel de endeudamiento	Relación entre el total de deudas y la capacidad de pago del socio.
	Frecuencia de solicitudes de crédito	Cantidad de veces que el socio solicita productos crediticios.
Cumplimiento	Puntualidad en pagos	Cumplimiento de los plazos establecidos para el pago de obligaciones.
	Historial de morosidad	Número de veces que el socio ha incurrido en incumplimiento o retrasos.
Interacción con la entidad	Contacto con canales de atención	Frecuencia de contacto con canales físicos o digitales de la cooperativa.
	Respuesta a campañas de fidelización	Participación en promociones u ofertas dirigidas a su perfil.

Nota. Elaboración propia con base en (García et al., 2021).

2.2.2.3. Análisis del historial de transacciones

El análisis del historial de transacciones es una técnica clave en el estudio del comportamiento financiero, ya que proporciona información detallada y cronológica sobre las operaciones que realiza un socio con una entidad financiera. Este análisis permite identificar patrones de uso, detectar señales tempranas de abandono, evaluar el nivel de compromiso con los productos financieros y

construir modelos predictivos de comportamiento, como la fuga de socios (Barreno et al., 2024).

1. Recolección de datos transaccionales

El primer paso consiste en la recolección de datos históricos relacionados con las transacciones de los socios. Esta información generalmente incluye:

- Fecha y hora de la operación.
- Tipo de transacción (retiros, depósitos, transferencias, pagos, etc.).
- Canal utilizado (cajero automático, banca móvil, ventanilla, etc.).
- Monto de la transacción.
- Frecuencia de uso por canal o tipo.

2. Limpieza y normalización de datos

Los datos recopilados pueden contener valores nulos, inconsistencias o duplicidades. Por ello, es necesario realizar una limpieza que garantice su calidad. Además, se normalizan las variables (por ejemplo, estandarizar montos y frecuencias) para facilitar su análisis estadístico y procesamiento mediante algoritmos de machine learning.

3. Construcción de variables derivadas

A partir de los datos transaccionales básicos, se pueden derivar nuevas variables que enriquecen el análisis, tales como:

- Promedio mensual de transacciones.
- Variabilidad en los montos transaccionados.
- Estacionalidad (detectar cambios de comportamiento por fechas o ciclos).
- Tendencia de uso de canales digitales vs. tradicionales.

4. Identificación de patrones de comportamiento

Mediante técnicas de minería de datos y análisis estadístico, se detectan patrones repetitivos que indican estabilidad o cambios en el comportamiento del socio. Por ejemplo:

- Disminución progresiva en el número de transacciones puede indicar pérdida de interés o vinculación.

- Cambios bruscos en el monto de las transacciones pueden asociarse a eventos externos (problemas económicos, cambio de entidad financiera, entre otros).

5. Visualización y análisis exploratorio

El uso de herramientas de visualización permite detectar tendencias o anomalías que no son fácilmente visibles en tablas de datos. Gráficos de series temporales, histogramas, mapas de calor o diagramas de dispersión son útiles para este propósito.

6. Aplicación en modelos predictivos

Finalmente, las variables obtenidas del historial de transacciones se utilizan como insumos en modelos predictivos, como redes neuronales, árboles de decisión o regresiones logísticas, que permiten anticipar la probabilidad de fuga de un socio con base en su comportamiento histórico.

Segmentación de socios según comportamiento

Como menciona Goyo (2024), la segmentación de socios según comportamiento consiste en clasificar a los usuarios de una entidad financiera en grupos homogéneos basados en sus hábitos, patrones de consumo, y nivel de interacción con los productos o servicios. Esta segmentación permite comprender mejor las necesidades y riesgos asociados a cada grupo, lo cual es clave para anticipar la fuga de socios y diseñar estrategias de retención personalizadas.

Los principales criterios utilizados en la segmentación comportamental incluyen:

- Frecuencia de uso de productos financieros (ahorro, créditos, tarjetas).
- Canales preferidos de interacción (banca digital, presencial, telefónica).
- Nivel de actividad transaccional (alta, media, baja).
- Estabilidad del comportamiento financiero (socios consistentes vs. volátiles).
- Respuestas ante campañas o beneficios ofrecidos.

2.2.2.4. Relación entre comportamiento financiero y fuga

La decisión de un socio de continuar o abandonar una entidad financiera depende de múltiples factores, tanto internos como externos, que afectan su

experiencia, confianza y percepción de valor (Moreno, 2025). A continuación, se detallan los factores más relevantes:

- **Calidad del servicio**: La atención rápida, personalizada y eficaz impacta directamente en la satisfacción del socio.
- **Tasas y comisiones**: Costos elevados o poco competitivos pueden incentivar la migración hacia otras entidades.
- **Accesibilidad a productos financieros**: La facilidad de acceso a créditos, ahorros u otros servicios influye en la permanencia.
- **Tecnología y canales digitales**: La disponibilidad y funcionalidad de aplicaciones móviles, plataformas web y otros canales digitales son determinantes en la experiencia del socio.
- **Experiencia previa**: Problemas anteriores no resueltos o malas experiencias pueden acelerar la decisión de abandono.
- **Falta de personalización**: La ausencia de ofertas o servicios adaptados a las necesidades individuales disminuye el sentido de fidelización.
- **Cambios en la situación económica del socio**: Pérdida de ingresos, desempleo o sobreendeudamiento pueden llevar a cerrar cuentas o reducir actividad.
- **Reputación institucional**: Escándalos financieros, mala prensa o pérdida de confianza institucional pueden motivar la fuga.
- **Falta de programas de fidelización**: La ausencia de incentivos o beneficios por permanencia puede restar atractivo a la entidad.
- **Competencia agresiva**: Promociones, tasas preferenciales o beneficios ofrecidos por otras entidades influyen en la decisión de cambio

2.2.3. Machine Learning para predicción de fuga

2.2.3.1. Marco de trabajo para el desarrollo y evaluación de los modelos

La investigación adoptó un marco de trabajo secuencial para el desarrollo y la evaluación de modelos de clasificación supervisada. Este comprende la obtención de los registros institucionales, la aplicación de criterios de inclusión y exclusión, la definición de la variable objetivo, la validación de la calidad de los datos, la construcción y transformación de las variables predictoras, la división estratificada del conjunto de datos, el tratamiento del desbalance, el

entrenamiento de los algoritmos, la evaluación comparativa y la interpretación del modelo seleccionado.

En la primera etapa se obtuvieron los datos procedentes del sistema BESTERP de la Cooperativa de Ahorro y Crédito Santa María Magdalena, con fecha de corte al 31 de diciembre de 2024. Posteriormente, se aplicaron los criterios de inclusión y exclusión establecidos para conformar la población analizada y se definió la variable objetivo binaria fuga_socio, en la que el valor 1 representó la renuncia voluntaria y el valor 0 la ausencia de dicho registro de renuncia.

En la etapa de preparación se verificó la integridad de los registros, se excluyó el identificador anonimizado del conjunto de variables predictoras, se codificaron las variables categóricas y se construyeron características derivadas a partir de las variables institucionales. Estas transformaciones tuvieron como finalidad representar relaciones financieras y transaccionales potencialmente relevantes para la clasificación, sin atribuirles efectos causales sobre la fuga.

A continuación, el conjunto de datos fue dividido aleatoriamente en entrenamiento y prueba mediante una partición estratificada de 80 % y 20 %, respectivamente. El conjunto de prueba no intervino en el entrenamiento, el balanceo de clases ni el ajuste de los modelos. Sin embargo, sus resultados se emplearon para la evaluación comparativa y la selección del modelo final, utilizando el F1-score como criterio principal. Debido al desbalance entre las clases, SMOTE se integró únicamente en los pliegues de entrenamiento de la validación cruzada, evitando que las observaciones sintéticas alcanzaran los pliegues de validación o el conjunto de prueba.

Los modelos candidatos fueron regresión logística, Random Forest, Gradient Boosting, XGBoost y LightGBM. Estos algoritmos fueron entrenados bajo una estructura comparable y evaluados mediante validación cruzada estratificada de cinco pliegues en el conjunto de entrenamiento. La evaluación final se efectuó sobre el conjunto de prueba empleando precisión, sensibilidad, F1-score, ROC-AUC y PR-AUC. El F1-score se utilizó como criterio principal de comparación debido a la necesidad de equilibrar la precisión y la detección de la clase minoritaria, mientras que ROC-AUC y PR-AUC permitieron complementar el análisis de la capacidad discriminativa.

Finalmente, el modelo con mejor desempeño comparativo fue interpretado mediante valores SHAP. Este procedimiento permitió estimar la contribución de las variables a las clasificaciones generadas por el modelo, tanto de manera global como individual, sin que dichos valores deban interpretarse como evidencia de relaciones causales.

Figura 2.

Marco de trabajo para el desarrollo y evaluación del modelo de machine Learning



Nota. El conjunto de prueba permaneció separado durante el entrenamiento. SMOTE se aplicó únicamente dentro de cada pliegue de entrenamiento para evitar la transferencia de información hacia los datos de validación y prueba.

2.2.3.2. Definición de aprendizaje automático

Según Forero y Negre (2024), el aprendizaje automático es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos matemáticos que permiten a los sistemas informáticos identificar patrones y aprender de los datos, sin ser programados explícitamente para cada tarea específica. A través de este enfoque, las máquinas pueden realizar predicciones o tomar decisiones basadas en datos históricos y actuales.

El aprendizaje automático es una tecnología que permite a los sistemas adaptarse y mejorar su desempeño en tareas específicas conforme se expone a nuevos datos. Este tipo de aprendizaje posibilita el desarrollo de modelos predictivos que se ajustan dinámicamente, siendo especialmente útil en contextos donde el comportamiento del entorno o de los usuarios puede cambiar con el tiempo, como en el caso del análisis de fuga de socios en instituciones financieras (Sharifani y Amini, 2023).

Aplicaciones de Machine Learning en el sector financiero

Como menciona Hermitaño (2022), el Machine Learning se ha consolidado como una herramienta clave en la transformación digital del sector financiero. Su capacidad para analizar grandes volúmenes de datos, identificar patrones y realizar predicciones precisas lo convierte en un recurso valioso para optimizar operaciones, mejorar la experiencia del socio y gestionar riesgos.

En el sector financiero, el machine learning se utiliza para evaluación crediticia, detección de fraude y predicción de fuga de socios. Dentro de estas aplicaciones, la predicción de fuga resulta relevante porque permite identificar oportunamente usuarios con riesgo de abandono.

2.2.3.3. Aplicación en entidades financieras

En las entidades financieras, la predicción de fuga puede abordarse como un problema de clasificación supervisada, en el cual se emplean datos históricos para identificar patrones asociados con la permanencia o desvinculación de los clientes o socios. Entre los datos utilizados suelen encontrarse las características financieras, crediticias, transaccionales y de vinculación con los productos ofrecidos por la entidad.

La revisión de los antecedentes desarrollados en el apartado 2.1 evidencia que los estudios sobre fuga de clientes han empleado diferentes algoritmos de clasificación. Por esta razón, en la presente investigación no se establece anticipadamente un algoritmo como el más adecu

ado, sino que se considera un conjunto de modelos candidatos cuya selección responde a sus características metodológicas y a su aplicabilidad en datos tabulares.

Tabla 3.

Algoritmos empleados en la predicción de fuga de socios

Algoritmo	Descripción	Ventajas	Limitaciones	Aplicabilidad en la fuga de socios
Regresión logística	Modelo de clasificación que estima la probabilidad de pertenencia a una clase binaria mediante una función logística aplicada a una combinación de variables predictoras.	Es sencilla, eficiente e interpretable; permite establecer una referencia clara para comparar modelos más complejos.	Puede presentar menor desempeño cuando existen relaciones no lineales o interacciones complejas entre las variables.	Se emplea como línea base interpretable para estimar la probabilidad de fuga y contrastar el desempeño de los modelos de ensamblado.
Random Forest	Método de ensamblado basado en múltiples árboles de decisión entrenados con diferentes muestras de datos y subconjuntos aleatorios de variables; la clasificación final se obtiene	Captura relaciones no lineales, reduce el sobreajuste respecto de un árbol individual y presenta robustez ante datos con numerosas variables.	Su interpretación directa es más compleja y puede requerir mayor capacidad computacional que un modelo lineal.	Permite reconocer patrones complejos presentes en las características financieras, crediticias y transaccionales asociadas con la fuga.

	mediante votación.			
Gradient Boosting	Método de ensamblado que construye árboles de manera secuencial, de modo que cada nuevo estimador busca corregir los errores cometidos por los anteriores.	Presenta elevada capacidad predictiva y modela relaciones no lineales e interacciones entre variables.	Es sensible a la selección de hiperparámetros y puede sobreajustarse si no se configura adecuadamente.	Facilita la identificación progresiva de patrones vinculados con socios que permanecen o se desvinculan de la cooperativa.
XGBoost	Implementación optimizada de *gradient boosting* que incorpora regularización, procesamiento paralelo y mecanismos para mejorar la eficiencia del entrenamiento.	Ofrece alto desempeño predictivo, control del sobreajuste, escalabilidad y eficiencia computacional.	Requiere ajustar varios hiperparámetros y posee menor interpretabilidad directa que la regresión logística.	Permite evaluar patrones complejos de fuga en datos tabulares y manejar interacciones entre las características de los socios.
LightGBM	Algoritmo de *gradient boosting* basado en árboles que utiliza estrategias eficientes de crecimiento y agrupamiento para reducir el tiempo y el uso de memoria.	Proporciona rapidez de entrenamiento, eficiencia computacional y buen desempeño con grandes conjuntos de datos tabulares.	Puede sobreajustarse en determinadas configuraciones y requiere controlar cuidadosamente sus hiperparámetros.	Resulta adecuado para analizar eficientemente grandes volúmenes de registros financieros y clasificar el riesgo de fuga.

Nota. Elaboración propia con base en Breiman (2001), Friedman (2001), Chen y Guestrin (2016) y Ke et al. (2017)

La regresión logística se incorporó como línea base interpretable, mientras que Random Forest, Gradient Boosting, XGBoost y LightGBM fueron

considerados métodos de ensamblado candidatos debido a su capacidad para modelar relaciones no lineales e interacciones complejas en datos tabulares. La inclusión de estas técnicas no supone la selección anticipada de alguna de ellas, puesto que la identificación del modelo con mejor desempeño se realizó mediante la comparación empírica presentada en el Capítulo IV, utilizando las métricas de evaluación definidas en la metodología.

2.2.3.4. Machine Learning como rama de la inteligencia artificial

El Machine Learning (aprendizaje automático) constituye una de las principales ramas de la Inteligencia Artificial (IA), orientada al desarrollo de sistemas capaces de aprender a partir de datos y mejorar su desempeño sin ser programados explícitamente para cada tarea. En este sentido, la IA se define como el campo general que busca simular procesos de inteligencia humana mediante sistemas computacionales, mientras que el Machine Learning se enfoca específicamente en el desarrollo de modelos que identifican patrones y generan predicciones a partir de datos históricos (Sarker, 2021).

En la actualidad, el Machine Learning ha adquirido gran relevancia debido a su capacidad para procesar grandes volúmenes de información y extraer conocimiento útil para la toma de decisiones. En el ámbito financiero, su aplicación resulta especialmente valiosa en problemas como la predicción de fuga de socios (churn), donde permite anticipar comportamientos futuros y diseñar estrategias de retención más efectivas (Prabadevi et al., 2023).

Desde una perspectiva teórica, el Machine Learning se clasifica en tres tipos principales de aprendizaje, los cuales se diferencian en función de la disponibilidad de datos etiquetados y del objetivo del modelo:

- Aprendizaje supervisado: Se basa en el uso de datos etiquetados, donde cada registro contiene una variable de entrada y una salida conocida. El modelo aprende una función que permite predecir la variable objetivo a partir de nuevas observaciones. Este tipo de aprendizaje es ampliamente utilizado en problemas de clasificación y regresión, siendo especialmente relevante en estudios de predicción de fuga, donde la variable objetivo suele ser binaria (por ejemplo, fuga o no fuga) (Osman, 2021).
- Aprendizaje no supervisado: Se caracteriza por trabajar con datos sin etiquetas, es decir, sin una variable objetivo definida. El propósito del modelo

es identificar patrones ocultos en los datos, como agrupaciones o segmentaciones. Este enfoque es útil para el análisis exploratorio, permitiendo descubrir estructuras internas en los datos, como perfiles de socios con comportamientos similares (Sarker, 2021).

- Aprendizaje por refuerzo: Se basa en la interacción de un agente con un entorno, donde el aprendizaje ocurre mediante un sistema de recompensas y penalizaciones. El agente busca maximizar una recompensa acumulada a través de la toma de decisiones secuenciales. Este enfoque es común en problemas dinámicos y de optimización, aunque su aplicación en el ámbito financiero es aún más limitada en comparación con los enfoques supervisados (IJCRT, 2021).

El presente estudio utiliza aprendizaje supervisado debido a que la variable objetivo se encuentra etiquetada como fuga y no fuga.

2.2.3.5. Aprendizaje supervisado

El aprendizaje supervisado es una de las técnicas más utilizadas dentro del Machine Learning, caracterizada por el uso de datos etiquetados para entrenar modelos capaces de realizar predicciones sobre nuevos datos. En este enfoque, cada instancia del conjunto de entrenamiento contiene variables de entrada (features) y una variable de salida conocida (target), lo que permite al algoritmo aprender una función que relacione ambas (Kotsiantis, 2007).

Este tipo de aprendizaje es ampliamente empleado en problemas donde se requiere predecir resultados específicos, como la detección de fraude, diagnóstico médico y predicción de fuga de socios. Su principal ventaja radica en la capacidad de generalizar patrones a partir de datos históricos, permitiendo obtener predicciones confiables cuando el modelo es correctamente entrenado y validado (Shalev-Shwartz & Ben-David, 2021).

En el aprendizaje supervisado, los algoritmos se dividen principalmente en dos categorías: clasificación y regresión. La clasificación se utiliza cuando la variable objetivo es categórica, mientras que la regresión se emplea cuando la variable es continua. En el contexto de la presente investigación, el problema de predicción de fuga de socios corresponde a una tarea de clasificación, dado que se busca determinar si un socio pertenece a una categoría específica (fuga o no fuga) (Hastie et al., 2021).

Además, el desempeño de los modelos supervisados depende en gran medida de la calidad de los datos utilizados, así como del proceso de entrenamiento, validación y selección del modelo. Por ello, es fundamental aplicar técnicas adecuadas de preprocesamiento, balanceo de clases y evaluación para garantizar resultados confiables y aplicables en entornos reales (Géron, 2022).

Clasificación binaria

La clasificación binaria es un tipo específico de aprendizaje supervisado en el que la variable objetivo solo puede tomar dos valores posibles, representando dos clases mutuamente excluyentes. Este tipo de problema es común en múltiples aplicaciones, tales como la detección de spam (correo no deseado o válido), diagnóstico médico (presencia o ausencia de enfermedad) y, en el caso de la presente investigación, la predicción de fuga de socios (fuga o no fuga) (Brownlee, 2020).

El objetivo principal de la clasificación binaria es construir un modelo que permita asignar correctamente nuevas observaciones a una de las dos clases posibles, en función de sus características. Para ello, los algoritmos generan una probabilidad asociada a cada clase, la cual posteriormente es transformada en una decisión mediante un umbral de clasificación (threshold), comúnmente fijado en 0.5, aunque este puede ajustarse según los requerimientos del problema (Raschka & Mirjalili, 2020).

Una de las características más importantes de la clasificación binaria es la necesidad de evaluar el modelo mediante métricas específicas que permitan medir su capacidad de discriminación entre clases. En particular, en problemas desbalanceados como la predicción de fuga, es fundamental considerar métricas como precision, recall y F1-Score, ya que el uso exclusivo de accuracy puede generar interpretaciones erróneas del desempeño del modelo. Asimismo, la clasificación binaria permite identificar patrones de comportamiento que diferencian claramente a los individuos de cada clase, lo que resulta especialmente útil en la toma de decisiones estratégicas. En el caso de la fuga de socios, este enfoque permite detectar socios con alta probabilidad de abandono, facilitando la implementación de acciones preventivas por parte de la organización.

2.2.3.6. Algoritmos de clasificación aplicados en el estudio

En el contexto del aprendizaje supervisado, los algoritmos de clasificación constituyen herramientas fundamentales para la predicción de variables categóricas a partir de datos históricos. En estudios orientados a la predicción de fuga de socios, estos algoritmos permiten identificar patrones complejos en el comportamiento de los usuarios, facilitando la estimación de la probabilidad de abandono y la toma de decisiones estratégicas en las organizaciones.

La selección de algoritmos de clasificación en una investigación depende de diversos factores, tales como la naturaleza de los datos, el tamaño del dataset, la presencia de desbalance de clases y la necesidad de interpretabilidad del modelo. En el presente estudio se emplearon diferentes algoritmos representativos del aprendizaje supervisado, con el objetivo de comparar su desempeño y seleccionar el modelo óptimo para la predicción de la fuga de socios.

2.2.3.6.1. Regresión Logística

La regresión logística es uno de los algoritmos más utilizados en problemas de clasificación binaria debido a su simplicidad, interpretabilidad y eficiencia computacional. Este modelo estima la probabilidad de pertenencia a una clase mediante una función logística (sigmoide), la cual transforma una combinación lineal de variables independientes en un valor comprendido entre 0 y 1.

Una de las principales ventajas de la regresión logística es su capacidad para interpretar la relación entre las variables predictoras y la variable objetivo, lo que permite identificar la influencia de cada variable en la probabilidad de ocurrencia del evento. Esta característica la convierte en un modelo base o de referencia en estudios comparativos de machine learning (Biecek & Burzykowski, 2021).

Sin embargo, la regresión logística presenta limitaciones cuando los datos contienen relaciones no lineales o interacciones complejas entre variables, lo que puede reducir su capacidad predictiva en comparación con modelos más avanzados basados en árboles de decisión o técnicas de ensamblado.

2.2.3.6.2. Random Forest

Random Forest es un algoritmo de aprendizaje basado en el ensamblado de múltiples árboles de decisión, donde cada árbol se entrena sobre una muestra

aleatoria del dataset y un subconjunto aleatorio de variables. Este enfoque permite reducir el sobreajuste (overfitting) y mejorar la capacidad de generalización del modelo (Breiman, 2001).

El principio fundamental de Random Forest consiste en la combinación de múltiples predicciones individuales (árboles) mediante votación mayoritaria en problemas de clasificación. Esta estrategia permite capturar relaciones no lineales y complejas en los datos, lo que lo convierte en un algoritmo robusto y altamente efectivo en problemas de predicción de churn (Probst et al., 2020).

Además, Random Forest es capaz de manejar datos con ruido, variables irrelevantes y distribuciones no balanceadas, lo que lo hace especialmente adecuado para entornos reales donde los datos no siguen supuestos estadísticos estrictos. Asimismo, proporciona medidas de importancia de variables, lo que facilita la interpretación del modelo y la identificación de factores relevantes en la predicción (Lundberg et al., 2020).

2.2.3.6.3. Gradient Boosting

Gradient Boosting es una técnica de ensamblado que construye modelos de forma secuencial, donde cada nuevo modelo se entrena para corregir los errores cometidos por los modelos anteriores. Este enfoque permite optimizar progresivamente el desempeño del modelo mediante la minimización de una función de pérdida (Chen & Guestrin, 2016).

Entre las implementaciones más destacadas de Gradient Boosting se encuentran XGBoost y LightGBM, ambos ampliamente utilizados en aplicaciones de machine learning por su alta eficiencia y precisión.

2.2.3.6.4. XGBoost

XGBoost (Extreme Gradient Boosting) es un algoritmo optimizado que incorpora regularización, manejo eficiente de valores faltantes y paralelización del entrenamiento, lo que permite obtener modelos con alto rendimiento y menor riesgo de sobreajuste. Su capacidad para manejar grandes volúmenes de datos y relaciones complejas lo convierte en una de las herramientas más utilizadas en competencias de ciencia de datos y aplicaciones empresariales (Chen & Guestrin, 2016).

2.2.3.6.5. LightGBM

Por su parte, **LightGBM (Light Gradient Boosting Machine)** introduce un enfoque basado en el crecimiento de árboles tipo leaf-wise, lo que permite una mayor eficiencia en el uso de memoria y tiempos de entrenamiento significativamente menores en comparación con otros algoritmos de boosting. Además, es especialmente adecuado para datasets de gran tamaño, manteniendo altos niveles de precisión en la predicción (Ke et al., 2017).

2.2.3.6.6. Criterios de selección de los algoritmos

La selección de los modelos respondió a un criterio de diversidad metodológica y comparabilidad. Se incluyó la regresión logística como línea base lineal e interpretable, con la finalidad de determinar si las relaciones presentes en los datos podían ser representadas adecuadamente mediante una función lineal. Random Forest fue incorporado como método de ensamblado basado en árboles, debido a su capacidad para modelar relaciones no lineales e interacciones entre variables, así como por su robustez frente a distribuciones asimétricas.

Gradient Boosting se consideró como representante del aprendizaje secuencial por ensamblado, en el cual cada estimador busca reducir los errores residuales del modelo anterior. XGBoost y LightGBM se incluyeron como implementaciones optimizadas del *gradient boosting*, debido a sus mecanismos de regularización, eficiencia computacional y capacidad para procesar conjuntos de datos tabulares de tamaño considerable. La comparación de estos algoritmos permitió evaluar modelos con distintos niveles de complejidad sin asumir anticipadamente cuál presentaría el mejor desempeño.

La inclusión de los modelos no se sustentó únicamente en su utilización en investigaciones anteriores. También se consideraron la naturaleza binaria de la variable objetivo, el predominio de datos tabulares, la posible presencia de relaciones no lineales, el tamaño del conjunto de datos, el desbalance entre las clases y la necesidad de interpretar posteriormente las estimaciones del modelo seleccionado.

2.2.3.6.7. Justificación de los algoritmos excluidos

No se incorporaron algoritmos como k vecinos más cercanos, máquinas de vectores de soporte, redes neuronales o métodos de agrupamiento. K vecinos

más cercanos puede presentar un costo elevado de predicción y sensibilidad a la escala y dimensionalidad de las variables. Las máquinas de vectores de soporte pueden requerir mayor costo computacional en conjuntos de datos con decenas de miles de observaciones y ofrecen una interpretación directa limitada. Las redes neuronales no fueron priorizadas debido a que el conjunto de datos utilizado es tabular y el estudio buscó mantener un equilibrio entre desempeño, complejidad e interpretabilidad. Los métodos de agrupamiento fueron excluidos porque corresponden al aprendizaje no supervisado y no responden directamente al propósito de clasificar una variable objetivo previamente etiquetada.

Asimismo, no se incluyó CatBoost debido a que la comparación ya incorporó tres alternativas de *boosting* basadas en árboles —Gradient Boosting, XGBoost y LightGBM—, consideradas suficientes para representar esta familia algorítmica dentro del alcance establecido. La exclusión de estos modelos no implica que sean inadecuados, sino que responde a la delimitación metodológica y computacional de la investigación.

2.2.3.7. Tratamiento de datos desbalanceados

En problemas de clasificación supervisada, es común encontrar datasets con distribución desigual entre las clases, fenómeno conocido como desbalance de clases. Esta situación ocurre cuando una de las clases (generalmente la de interés, como la fuga de socios) presenta una menor proporción de registros en comparación con la clase mayoritaria. Este desbalance puede afectar significativamente el desempeño de los modelos de machine learning, ya que muchos algoritmos tienden a favorecer la clase predominante, generando resultados sesgados y reduciendo la capacidad de detección de la clase minoritaria (Elreedy & Atiya, 2019).

En el contexto de la predicción de fuga, este problema es especialmente crítico, dado que la clase minoritaria representa precisamente a los socios que abandonan la cooperativa, los cuales son de mayor interés para la toma de decisiones. Por ello, es necesario aplicar técnicas que permitan equilibrar la distribución de clases y mejorar la capacidad predictiva del modelo.

Una de las técnicas más utilizadas para abordar este problema es SMOTE (Synthetic Minority Over-sampling Technique), la cual consiste en generar nuevas muestras sintéticas de la clase minoritaria a partir de sus vecinos más cercanos.

A diferencia de los métodos tradicionales de sobremuestreo, que replican instancias existentes, SMOTE crea nuevos registros interpolando entre ejemplos reales, lo que permite enriquecer la representación de la clase minoritaria sin introducir duplicados (Fernández et al., 2018).

El funcionamiento de SMOTE se basa en seleccionar una instancia de la clase minoritaria y generar nuevos puntos en el espacio de características mediante la combinación de dicha instancia con sus vecinos más cercanos. Este proceso contribuye a mejorar la capacidad del modelo para identificar patrones asociados a la clase minoritaria, reduciendo el sesgo hacia la clase mayoritaria y aumentando métricas clave como el recall y el F1-Score.

No obstante, es importante considerar que el uso de SMOTE debe aplicarse cuidadosamente dentro del proceso de entrenamiento, evitando su implementación sobre el conjunto completo de datos, ya que esto podría generar filtración de información (data leakage). Por ello, se recomienda aplicar esta técnica únicamente en el conjunto de entrenamiento o dentro de procesos de validación cruzada, garantizando así una evaluación más realista del modelo.

2.2.3.8. Validación y métricas de evaluación

La evaluación de modelos de clasificación constituye una etapa fundamental en el proceso de machine learning, ya que permite medir el desempeño del modelo y determinar su capacidad para generalizar correctamente sobre nuevos datos. En problemas de clasificación binaria, como la predicción de fuga de socios, es necesario emplear métricas específicas que permitan analizar el comportamiento del modelo desde diferentes perspectivas, especialmente en contextos donde existe desbalance de clases (Tharwat, 2021).

El uso de múltiples métricas resulta indispensable, dado que una única medida, como la exactitud (accuracy), puede no reflejar adecuadamente el desempeño del modelo en escenarios donde la clase minoritaria es la más relevante. Por ello, se utilizan métricas complementarias que permiten evaluar la capacidad de discriminación, detección y precisión del modelo (Sofaer et al., 2020).

2.2.3.8.1. Matriz de confusión

La matriz de confusión es una herramienta fundamental para evaluar el desempeño de un modelo de clasificación, ya que permite visualizar de manera detallada los aciertos y errores del modelo en términos de predicciones correctas e incorrectas. Esta matriz se compone de cuatro elementos: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN), los cuales representan las diferentes combinaciones entre valores reales y predichos (Tharwat, 2021).

El análisis de la matriz de confusión permite comprender el tipo de errores que comete el modelo, siendo especialmente relevante en problemas donde ciertos errores tienen mayor impacto que otros. En el caso de la predicción de fuga, los falsos negativos representan socios que abandonan la cooperativa sin ser detectados, lo que implica una pérdida directa para la organización, mientras que los falsos positivos representan alertas innecesarias, pero menos costosas (Sofaer et al., 2020).

2.2.3.8.2. Accuracy, Precision y Recall

La exactitud (accuracy) es una de las métricas más utilizadas en la evaluación de modelos de clasificación, y se define como la proporción de predicciones correctas respecto al total de observaciones. Sin embargo, en problemas desbalanceados, esta métrica puede resultar engañosa, ya que un modelo puede alcanzar altos niveles de accuracy simplemente prediciendo la clase mayoritaria (Hossin & Sulaiman, 2020).

Por ello, se emplean métricas complementarias como la **precision** y el **recall**. La precision mide la proporción de verdaderos positivos sobre el total de predicciones positivas, indicando qué tan confiables son las predicciones del modelo. Por su parte, el recall mide la proporción de verdaderos positivos sobre el total de casos reales positivos, reflejando la capacidad del modelo para detectar correctamente la clase de interés (Chicco & Jurman, 2020).

Estas métricas son especialmente relevantes en problemas de churn, donde el objetivo principal es identificar correctamente a los socios que abandonarán la organización, priorizando la detección sobre la exactitud global del modelo.

2.2.3.8.3. F1-Score

El F1-Score es una métrica que combina la precision y el recall en una sola medida, mediante su media armónica. Esta métrica resulta particularmente útil en problemas de clasificación desbalanceada, ya que proporciona un equilibrio entre la capacidad de detección del modelo y la precisión de sus predicciones (Chicco & Jurman, 2020).

El F1-Score penaliza aquellos modelos que presentan valores extremos en precision o recall, favoreciendo aquellos que mantienen un balance adecuado entre ambas métricas. Por esta razón, es ampliamente utilizado como criterio principal de evaluación en problemas de predicción de fuga, donde se busca maximizar la identificación de casos reales sin generar un exceso de falsas alarmas (Powers, 2020).

2.2.3.8.4. ROC-AUC y curva ROC

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica que permite evaluar el desempeño de un modelo de clasificación a diferentes umbrales de decisión, representando la relación entre la tasa de verdaderos positivos (recall) y la tasa de falsos positivos (FPR). Esta curva permite analizar la capacidad del modelo para discriminar entre clases independientemente del umbral seleccionado (Fawcett, 2006).

El área bajo la curva ROC (ROC-AUC) es una métrica que resume el desempeño del modelo en todos los posibles umbrales, tomando valores entre 0 y 1. Un valor cercano a 1 indica un modelo con alta capacidad de discriminación, mientras que un valor cercano a 0.5 sugiere un comportamiento similar al azar (Saito & Rehmsmeier, 2020).

En problemas desbalanceados, la curva ROC debe complementarse con otras métricas como el F1-Score o curvas Precision-Recall, ya que puede presentar una visión optimista del modelo cuando la clase mayoritaria domina el dataset.

2.2.3.8.5. PR-AUC

El área bajo la curva precisión-recall resume la relación entre la precisión y la sensibilidad alcanzadas por el modelo a través de diferentes umbrales. Esta métrica resulta especialmente pertinente cuando existe desbalance de clases,

porque concentra la evaluación en el desempeño correspondiente a la clase positiva. En la investigación se utilizó como métrica complementaria para analizar la capacidad del modelo de identificar a los socios clasificados con riesgo de fuga.

2.2.3.9. Interpretabilidad del modelo mediante valores SHAP

La interpretabilidad de los modelos de machine learning se ha convertido en un aspecto fundamental, especialmente en aplicaciones donde las decisiones tienen impacto directo en personas u organizaciones. En este contexto, técnicas como SHAP (SHapley Additive exPlanations) permiten explicar las predicciones de modelos complejos, proporcionando información sobre la contribución de cada variable en el resultado final (Molnar, 2022).

SHAP se basa en la teoría de valores de Shapley, proveniente de la teoría de juegos, la cual asigna a cada variable una contribución proporcional a su impacto en la predicción del modelo. Esta metodología permite descomponer una predicción individual en términos de la influencia de cada característica, facilitando la comprensión del comportamiento del modelo (Covert et al., 2020).

Una de las principales ventajas de SHAP es su capacidad para ofrecer tanto interpretaciones globales como locales. A nivel global, permite identificar las variables más importantes en el modelo, mientras que a nivel local explica por qué una observación específica fue clasificada de determinada manera. Esto resulta especialmente útil en problemas de predicción de fuga, donde es necesario comprender las razones detrás del riesgo de abandono de cada socio (Slack et al., 2020).

Asimismo, la interpretabilidad mediante SHAP contribuye a comprender la influencia de las variables en los resultados obtenidos por los modelos predictivos (Vowels et al., 2023). En entornos empresariales, esta capacidad es clave para la adopción de soluciones basadas en inteligencia artificial, ya que facilita la transparencia y la trazabilidad de las decisiones automatizadas (Bhatt et al., 2020).

Capítulo III

Material y métodos

3.1. Tipo de investigación

Tipo

Según Vizcaíno, Cedeño y Maldonado (2023), el tipo de investigación hace referencia a la finalidad del estudio en relación con la generación y aplicación del conocimiento. En este caso, la investigación es de tipo aplicada, ya que su propósito principal no es solamente ampliar el conocimiento teórico, sino utilizar dicho conocimiento para resolver un problema concreto dentro de un contexto real y específico.

La presente investigación se clasifica como aplicada porque busca desarrollar un modelo predictivo basado en técnicas de Machine Learning para identificar la fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena. El objetivo no es únicamente describir el fenómeno, sino también diseñar y proponer una solución tecnológica que contribuya directamente a mejorar la gestión institucional y la toma de decisiones en el área de retención de socios. De este modo, el estudio se orienta a la utilización práctica de herramientas científicas y tecnológicas para intervenir en una problemática del entorno organizacional.

Enfoque

Como menciona Vizcaíno, Cedeño y Maldonado (2023), el enfoque cuantitativo se caracteriza por el uso de datos numéricos, procedimientos estadísticos y métodos sistemáticos orientados a medir, analizar y explicar fenómenos de manera objetiva. Este enfoque busca identificar patrones, relaciones y comportamientos mediante la recolección y procesamiento estructurado de información, permitiendo obtener resultados verificables y generalizables. Asimismo, el enfoque cuantitativo se sustenta en el uso de técnicas matemáticas y estadísticas que facilitan la validación de hipótesis y la evaluación precisa de variables dentro de un contexto de investigación determinado.

La presente investigación se desarrolla bajo un enfoque cuantitativo debido a que utiliza datos históricos estructurados correspondientes a los registros financieros y transaccionales de los socios de la Cooperativa de Ahorro y Crédito Santa María Magdalena. A partir de dichos datos se aplicaron técnicas de analítica de datos y

algoritmos de machine learning para medir y predecir el riesgo de fuga de socios mediante métricas cuantificables como F1-Score, ROC-AUC y recall. Además, el estudio emplea procedimientos estadísticos y computacionales para el procesamiento, entrenamiento y evaluación de modelos predictivos, permitiendo analizar objetivamente el comportamiento de los socios y determinar patrones asociados a la probabilidad de abandono.

3.2. Nivel de investigación

Como indica Vizcaíno, Cedeño y Maldonado (2023), el nivel de investigación se refiere a la profundidad con la que se aborda el objeto de estudio, y está vinculado con los objetivos y el alcance que se pretende lograr. En el presente estudio, el nivel de investigación es descriptivo-comparativo con propósito predictivo, dado que se busca, por un lado, caracterizar detalladamente las variables involucradas, y estimar la probabilidad de pertenencia de los socios a las clases fuga o no fuga mediante modelos de clasificación supervisada.

En su componente descriptivo, la investigación busca identificar y analizar las características sociodemográficas, financieras, crediticias y transaccionales de los socios de la Cooperativa de Ahorro y Crédito Santa María Magdalena. Este análisis permitió reconocer los patrones presentes en los datos y describir el perfil de los socios que presentan mayor riesgo de fuga.

En su componente comparativo, se evaluó el desempeño de los modelos de machine learning implementados mediante métricas de clasificación previamente establecidas, como exactitud, precisión, sensibilidad, F1-score y área bajo la curva ROC. Esta comparación permitió determinar cuál de los modelos presenta el mejor desempeño para identificar a los socios con riesgo de fuga.

Finalmente, el propósito predictivo estuvo orientado a utilizar los patrones identificados en los registros históricos para estimar la probabilidad de fuga de los socios. Por tanto, la investigación no pretende establecer relaciones de causalidad ni contrastar diferencias mediante una hipótesis estadística, sino desarrollar y evaluar modelos predictivos para seleccionar aquel que ofrezca los resultados más adecuados de acuerdo con las métricas definidas.

3.3. Diseño de la investigación

La investigación adoptó un diseño no experimental, de corte transversal retrospectivo, debido a que se analizaron registros históricos de los socios de la Cooperativa de Ahorro y Crédito Santa María Magdalena disponibles hasta el 31 de diciembre de 2024, sin manipular deliberadamente las variables ni intervenir sobre los sujetos de estudio. El diseño no experimental permitió analizar las características registradas y su asociación con la condición de fuga observada en el sistema institucional (Vizcaíno et al., 2023).

A partir del conjunto de datos obtenido, se desarrollaron modelos de clasificación supervisada para estimar la probabilidad de pertenencia a las clases fuga o no fuga. El conjunto de datos se dividió mediante una partición aleatoria estratificada en subconjuntos de entrenamiento y prueba, preservando la proporción de ambas clases. Este procedimiento permitió comparar el desempeño de los algoritmos sobre observaciones no utilizadas durante su entrenamiento.

Debido al carácter transversal retrospectivo del diseño y al uso de una partición aleatoria, la capacidad predictiva evaluada corresponde a la clasificación de observaciones pertenecientes al periodo analizado. No se efectuó una partición cronológica ni una validación out-of-time con registros de periodos posteriores. Por consiguiente, los resultados no representan un pronóstico del momento futuro de la fuga ni permiten asegurar que el desempeño se mantendrá sin variaciones en periodos posteriores.

3.4. Población y muestra

3.4.1. Población

Según Reyes (2022), la población en una investigación se refiere al conjunto total de elementos, personas, eventos u objetos que poseen ciertas características comunes y que son objeto de estudio para el investigador. Constituye el universo al cual se pretende generalizar los resultados obtenidos.

En el presente estudio, la población inicial es de 126,540 registros de socios registrados en BESTERP al 31 de diciembre de 2024; y población de análisis de 55,432 registros elegibles de socios de la CACSMM, clasificados en socios no fugados y socios con renuncia voluntaria formal de la CACSMM. Esta población incluye a todos los socios que han mantenido algún tipo de relación financiera con la

entidad, tales como cuentas de ahorro, créditos o transacciones financieras durante el periodo analizado.

3.4.2. Muestra

Como menciona Reyes (2022), la muestra es un subconjunto representativo de la población, seleccionado mediante procedimientos específicos, y que permite inferir conclusiones aplicables al total del universo en estudio. No obstante, en ciertos casos, cuando el tamaño y la accesibilidad lo permiten, se puede optar por el estudio de toda la población, lo cual se denomina censo.

Dado que se cuenta con acceso a la totalidad de los registros financieros de los socios de la cooperativa y se dispone de herramientas tecnológicas que facilitan el procesamiento masivo de datos, no se aplicó muestreo probabilístico y se ha determinado realizar un censo de la población de análisis, conformada por 55,432 registros elegibles. En consecuencia, la muestra estará compuesta por los 55,432 registros elegibles de socios de la CACSMM, clasificados en socios no fugados y socios con renuncia voluntaria formal que conforman la población total, permitiendo así un análisis integral y detallado para la predicción de la fuga de socios utilizando técnicas de aprendizaje automático.

3.4.3. Criterios de inclusión y exclusión

Para la construcción del dataset utilizado en el entrenamiento y evaluación de los modelos de machine learning, se aplicaron criterios de selección sobre la base de datos histórica de socios de la CACSMM con fecha de corte al 31 de diciembre de 2024. Estos criterios permitieron garantizar la calidad, consistencia y pertinencia de los registros incluidos en el estudio.

Criterios de inclusión

- Socios registrados como personas naturales.
- Socios con edad igual o superior a 18 años al momento del corte de información.
- Socios activos que mantenían vínculo vigente con la cooperativa o que hubieran presentado una renuncia voluntaria formal registrada en el sistema institucional.
- Registros que contaban con información válida y completa para las variables consideradas en el modelamiento predictivo.

Criterios de exclusión

- Personas jurídicas, tales como empresas, asociaciones u otras organizaciones, debido a que presentan patrones financieros distintos a los de los socios individuales.
- Socios menores de 18 años, por tratarse de cuentas administradas por representantes legales y no reflejar decisiones financieras autónomas.
- Socios candidatos a depuración administrativa según la política institucional de la cooperativa, definida para aquellos que mantienen saldos inferiores a S/ 20.00 en todas sus cuentas y no registran movimientos durante los últimos 24 meses.
- Registros correspondientes a procesos de desafiliación por depuración institucional, debido a que representan casos de abandono pasivo y no renuncias voluntarias formales, que constituyen la definición operativa de fuga utilizada en la presente investigación.

3.5. Variables e indicadores

3.5.1. Definición conceptual de las variables

Variable objetivo (Y): Fuga de socios

La fuga de socios se entiende como la desvinculación voluntaria de un socio respecto de la cooperativa, formalizada mediante su renuncia y la finalización de su relación institucional. Este fenómeno puede afectar la sostenibilidad de la entidad debido a la reducción del capital social, la disminución del uso de productos financieros y la pérdida de socios.

Variable predictora 1 (X1): Características bancarias del socio

Son los atributos que describen la vinculación institucional y financiera del socio con la cooperativa. Comprenden aspectos como su antigüedad, la cantidad de productos financieros activos y el número de préstamos otorgados durante su relación con la entidad. Estas características permiten identificar el grado de vinculación del socio con la cooperativa.

Variable predictora 2 (X2): Comportamiento financiero del socio

Es el conjunto de patrones relacionados con el uso y manejo de los productos financieros del socio. Comprende la frecuencia de sus operaciones, los retiros realizados, el movimiento transaccional, los aportes acumulados, el historial de pagos

y los días de mora. Su análisis permite reconocer conductas financieras relacionadas con la permanencia o fuga del socio.

Variable predictora 3 (X3): Características demográficas y socioeconómicas del socio

Son los atributos personales, sociales y económicos que describen el perfil del socio. Comprenden la edad, el sexo, el estado civil, el nivel educativo, la ocupación, la actividad económica y los ingresos mensuales. Estas características permiten analizar las diferencias existentes entre los perfiles de los socios y su posible relación con la fuga.

3.5.2. Definición operacional de las variables

Variable objetivo (Y): Fuga de socios

La fuga de socios se midió mediante el registro de renuncia voluntaria formal consignado en el sistema BESTERP de la CACSMM. Se representó como una variable nominal dicotómica: 0 = no fuga y 1 = fuga. Esta variable constituyó la clase objetivo de los modelos de clasificación supervisada.

En esta investigación, el término predicción se empleó en el sentido de clasificación supervisada y se entendió como la estimación de una clase o de la probabilidad de pertenencia a una clase a partir de las variables observadas. El modelo asignó a cada socio una probabilidad estimada de pertenecer a las categorías no fuga o fuga, de acuerdo con los patrones identificados en los registros disponibles hasta el 31 de diciembre de 2024. Por tanto, el modelo no estimó el momento futuro exacto en que ocurriría la desvinculación ni fue validado como un instrumento de pronóstico temporal. Su desempeño representa la capacidad de clasificación alcanzada dentro del conjunto de datos analizado y bajo el procedimiento de partición aleatoria utilizado.

Variable predictora 1 (X1): Características bancarias del socio

Las características bancarias se midieron mediante la antigüedad en meses, la cantidad de productos activos y el número de préstamos otorgados al socio. Los datos se obtuvieron de los registros históricos del sistema BESTERP y se incorporaron como variables de entrada en los modelos de clasificación.

Variable predictora 2 (X2): Comportamiento financiero del socio

El comportamiento financiero se midió mediante el total de aportes acumulados, los retiros efectuados durante los últimos tres meses, el movimiento total de los últimos seis meses, la frecuencia mensual de uso, los días de mora y el historial de pagos. Los datos se obtuvieron de los registros históricos y transaccionales del sistema BESTERP y se emplearon como variables de entrada en los modelos de clasificación.

Variable predictora 3 (X3): Características demográficas y socioeconómicas del socio

Las características demográficas y socioeconómicas se midieron mediante la edad, el sexo, el estado civil, el nivel educativo, la ocupación, la actividad económica y los ingresos mensuales registrados en el sistema BESTERP. Las variables categóricas se codificaron para su incorporación en los modelos, mientras que las variables numéricas conservaron su naturaleza cuantitativa durante el procesamiento.

Variables derivadas mediante ingeniería de características

Adicionalmente, durante la etapa de ingeniería de características se generaron cinco indicadores derivados a partir de las variables predictoras originales: ratio retiro/aportes, intensidad de movimiento, uso por producto, riesgo de mora y ratio ingreso/movimiento. Estos indicadores se calcularon durante el procesamiento del conjunto de datos. Posteriormente, se incorporaron a la matriz de entrenamiento únicamente aquellos que fueron conservados después del proceso de selección de características.

3.6. Técnicas e instrumentos para el tratamiento de datos e información

3.6.1. Técnicas de recolección de datos

Las técnicas de recolección de datos constituyen los procedimientos utilizados para obtener información relevante en función de los objetivos de investigación (Reyes, 2022). En la presente investigación, la información fue obtenida a partir de registros históricos almacenados en el sistema BESTERP de la Cooperativa de Ahorro y Crédito Santa María Magdalena.

En la presente investigación se emplearon dos técnicas de recolección de datos: el análisis documental y la extracción de registros automatizados. El análisis

documental permitió revisar documentación institucional relacionada con la gestión de socios, políticas de retención, procedimientos internos y normativa aplicable a la cooperativa, con la finalidad de comprender el contexto del problema de investigación y apoyar la selección de variables relevantes. Por su parte, la extracción de registros automatizados permitió obtener información histórica de los socios almacenada en el sistema BESTERP de la Cooperativa de Ahorro y Crédito Santa María Magdalena. Estos registros incluyeron variables demográficas, financieras y transaccionales utilizadas para la construcción y evaluación del modelo predictivo de fuga de socios.

3.6.2. Instrumentos de recolección y procesamiento

Para la presente investigación se emplearon dos instrumentos complementarios para la obtención y preparación de la información utilizada en el desarrollo del modelo predictivo.

Ficha de registro digital de datos históricos. Este instrumento fue utilizado para recopilar y organizar la información extraída del sistema BESTERP de la Cooperativa de Ahorro y Crédito Santa María Magdalena. La ficha permitió registrar las variables originales de los socios que forman parte del dataset, incluyendo edad, sexo, estado civil, nivel educativo, ocupación, actividad económica, ingresos mensuales, aportes totales, retiros de los últimos tres meses, movimiento total de los últimos seis meses, productos activos, frecuencia de uso mensual, préstamos otorgados, días de mora, antigüedad como socio e historial de pagos. Estas variables constituyeron la base inicial para el análisis exploratorio y el entrenamiento de los modelos de machine learning.

Ficha de procesamiento de variables derivadas. Este instrumento fue utilizado durante la etapa de preprocesamiento y feature engineering para documentar y organizar las nuevas variables construidas a partir de las variables originales. La ficha permitió registrar indicadores derivados orientados a capturar patrones de comportamiento financiero relevantes para la predicción de fuga de socios. Entre las variables derivadas generadas se encuentran el ratio retiro/aportes, la intensidad de movimiento, el uso por producto, el riesgo de mora y el ratio ingreso/movimiento. Estas variables fueron incorporadas al dataset final utilizado para el entrenamiento y evaluación de los algoritmos de clasificación supervisada.

Finalmente, la información recopilada mediante ambos instrumentos fue integrada en un dataset estructurado, el cual fue sometido a procesos de limpieza,

transformación, codificación y balanceo de clases para su utilización en el desarrollo y evaluación de los modelos predictivos.

3.6.3. Técnica de análisis de datos

El análisis de datos se realizó bajo un enfoque cuantitativo, empleando técnicas estadísticas y de machine learning orientadas a la construcción y evaluación de un modelo predictivo de fuga de socios.

En una primera etapa se aplicó la estadística descriptiva, mediante el uso de medidas de tendencia central, medidas de dispersión, tablas de frecuencia y representaciones gráficas, con el propósito de caracterizar las variables incluidas en el estudio. Posteriormente, se desarrolló un análisis exploratorio de datos (Exploratory Data Analysis - EDA) para identificar patrones de comportamiento, distribuciones, relaciones entre variables, valores atípicos y posibles problemas de calidad de los datos.

La fase de preparación de datos comprendió técnicas de limpieza y depuración de registros, tratamiento de inconsistencias, codificación de variables categóricas, así como la aplicación de feature engineering, orientada a la construcción de variables derivadas que permitieran representar de manera más eficiente los patrones de comportamiento financiero de los socios. Asimismo, debido al desbalance existente entre las clases de la variable objetivo, se empleó la técnica de balanceo de clases mediante SMOTE (Synthetic Minority Oversampling Technique).

Para la etapa de modelamiento se utilizaron técnicas de clasificación supervisada, evaluando los algoritmos Regresión Logística, Random Forest, Gradient Boosting, XGBoost y LightGBM. El proceso de entrenamiento y validación se realizó mediante validación cruzada estratificada, con la finalidad de estimar la capacidad de generalización de los modelos y reducir el riesgo de sobreajuste.

Finalmente, el desempeño de los modelos fue evaluado mediante las métricas Accuracy, Precision, Recall, F1-Score y Área Bajo la Curva ROC (ROC-AUC), ampliamente utilizadas en problemas de clasificación binaria desbalanceada. Adicionalmente, se incorporó una aproximación de interpretabilidad basada en SHAP (SHapley Additive Explanations), con el propósito de identificar las variables que presentan mayor contribución en las predicciones generadas por el modelo seleccionado.

3.6.4. Herramientas computacionales

Para el desarrollo de la presente investigación se emplearon herramientas computacionales especializadas en análisis de datos y machine learning. Como entorno de programación se utilizó el lenguaje Python, debido a su amplia aplicación en proyectos de ciencia de datos, aprendizaje automático y analítica predictiva. Asimismo, la ejecución de los procesos de análisis, entrenamiento y evaluación de modelos se realizó en Google Colab, plataforma que proporciona recursos computacionales y facilita la integración de librerías especializadas para el procesamiento de datos.

Entre las principales librerías utilizadas se encuentran Pandas y NumPy para la manipulación, transformación y análisis de datos; Scikit-learn para el desarrollo, entrenamiento y evaluación de modelos de clasificación supervisada; Imbalanced-learn para la aplicación de la técnica de balanceo de clases SMOTE; XGBoost y LightGBM para la implementación de algoritmos de gradient boosting; y SHAP (SHapley Additive Explanations) para el análisis de interpretabilidad del modelo seleccionado.

Asimismo, se utilizaron las librerías Matplotlib y Seaborn para la generación de gráficos estadísticos, visualizaciones exploratorias y representación de resultados. El uso conjunto de estas herramientas permitió ejecutar de manera eficiente las etapas de preparación de datos, construcción de variables derivadas, entrenamiento de modelos, evaluación de desempeño e interpretación de resultados.

3.7. Procedimiento metodológico para el desarrollo y evaluación del modelo

3.7.1. Entorno de ejecución y herramientas utilizadas

El pipeline experimental se ejecutó en **Google Colab**, utilizando un entorno de ejecución **Python 3** con procesamiento mediante **CPU**, sin acelerador GPU o TPU. Durante la ejecución, el entorno dispuso de aproximadamente **12,7 GB de memoria RAM** y **107,7 GB de almacenamiento en disco**. Debido a que Google Colab asigna dinámicamente los recursos computacionales, el modelo específico de procesador puede variar entre sesiones; por esta razón, los tiempos de entrenamiento reportados corresponden a las condiciones de ejecución disponibles durante el experimento y deben interpretarse como una referencia del costo computacional relativo entre los algoritmos, no como tiempos absolutos reproducibles en cualquier infraestructura.

Para el procesamiento, el entrenamiento y la evaluación de los modelos se emplearon librerías de código abierto ampliamente utilizadas en el campo del aprendizaje automático. Las herramientas utilizadas se presentan en la Tabla 4.

Tabla 4.

Librerías utilizadas

Librería	Versión	Propósito
scikit-learn	≥ 1.3	Preprocesamiento, modelos, métricas y validación cruzada
imbalanced-learn	≥ 0.11	SMOTE para balanceo de clases dentro de cada fold
XGBoost	≥ 1.7	Modelo de gradient boosting optimizado
LightGBM	≥ 4.0	Modelo de gradient boosting de alta eficiencia
SHAP	≥ 0.44	Interpretabilidad
pandas	≥ 2.0	Manipulación y análisis de datos tabulares
numpy	≥ 1.24	Operaciones numéricas y matrices
matplotlib / seaborn	≥ 3.7	Visualización estadística y gráficos
joblib	≥ 1.3	Serialización y carga de modelos entrenados
scipy.stats	≥ 1.11	Distribuciones para RandomizedSearchCV

Nota. Se presentan las librerías utilizadas durante la ejecución del pipeline, junto con sus versiones y propósitos principales.

3.7.2. Extracción de datos desde BESTERP

Origen y Proceso de Extracción

Los datos se extrajeron del sistema BESTERP de la CACSMM, en el cual se registraron las operaciones financieras, las características demográficas y socioeconómicas, así como el historial de vinculación de los socios. La extracción se realizó a partir de la base de datos operativa, con fecha de corte al 31 de diciembre de 2024. Asimismo, se aplicaron criterios de inclusión orientados a garantizar la calidad y pertinencia de los registros utilizados en el modelamiento predictivo.

3.7.3. Aplicación de criterios de inclusión y exclusión

Criterios de Inclusión y Filtrado

La construcción del dataset siguió un proceso de filtrado secuencial basado en tres criterios de inclusión. Cada criterio responde a una justificación operativa o metodológica específica:

Tabla 5.

Criterios de filtrado

Nº	Criterio	Justificación	Registros excluidos	Registros restantes
0	Base original	Total de registros en el sistema de CACSMM al 31/12/2024	—	126,540
1	Solo personas naturales	Se excluyen personas jurídicas (empresas, asociaciones, instituciones) cuyo comportamiento financiero difiere estructuralmente del de socios individuales y requiere modelamiento independiente.	7393 (5.84%)	119147
2	Edad ≥ 18 años	Se excluyen cuentas de menores de edad, que son administradas por sus tutores legales y no reflejan decisiones financieras autónomas del titular.	27265 (21.55%)	91882
3	Socios (activos)	Se excluyen socios que son candidatos a depuracion por la cooperativa. La política institucional de CACSMM establece una depuración	36450 (28.80%)	55432

	de socios con saldo menor a S/ 20 en todas sus cuentas y sin movimientos en los últimos 2 años. Estos socios son desafiliados administrativamente y en un ~95% de los casos nunca presentaron renuncia voluntaria.

Nota. Listado de criterios para el filtro de datos.

Justificación de la Exclusión de Socios Depurados

La exclusión de socios candidatos a depuración merece una explicación particular por su impacto en la definición del problema. CACSMM realiza un proceso de depuración institucional en el que se desafilian a socios que cumplen simultáneamente dos condiciones: (a) mantener un saldo inferior a S/ 20.00 en la totalidad de sus cuentas, y (b) no haber registrado ninguna transacción en los últimos 24 meses.

Estos socios representan un fenómeno cualitativamente distinto a la fuga voluntaria. La evidencia operativa de CACSMM indica que aproximadamente el 95% de los socios depurados nunca presentan una solicitud formal de renuncia, lo que sugiere que su desvinculación es pasiva — muchos parecen haber olvidado su condición de socios. Incluir estos registros como casos de fuga_socio = 1 contaminaría el modelo con un patrón de abandono pasivo que no es susceptible de intervención mediante estrategias de retención, que es el objetivo último del sistema predictivo.

En contraste, la fuga voluntaria capturada en el dataset corresponde a socios que toman la decisión deliberada de acudir a la cooperativa, solicitar su renuncia formal y retirar sus fondos. Este tipo de deserción sí presenta patrones transaccionales identificables (incremento de retiros, reducción de uso de productos) que un modelo predictivo puede detectar con antelación suficiente para permitir una intervención del área de retención.

3.7.4. Construcción del dataset final

Descripción General del Dataset

El presente estudio utiliza el dataset SOCIOS_CACSMM_CHURN.csv, construido a partir de los registros operativos de la Cooperativa de Ahorro y Crédito Santa María Magdalena (CACSMM) con fecha de corte al 31 de diciembre de 2024. El dataset contiene 55,432 registros correspondientes a socios individuales y 18 campos: 16 variables predictoras originales, una variable objetivo denominada fuga_socio y un identificador anonimizado denominado socio_id. De las 16 variables predictoras, 10 son numéricas y seis son categóricas. El campo socio_id fue utilizado únicamente para identificar los registros y se excluyó del modelamiento.

3.7.5. Definición de la variable objetivo

Definición de la Variable Objetivo

La variable objetivo fuga_socio es una variable binaria (0/1) que indica si el socio presentó una renuncia voluntaria explícita a la cooperativa. Siempre teniendo en cuenta que la predicción de la fuga de socios está relacionado a la estimación de una probabilidad a partir de las variables observadas. La definición operacional es la siguiente:

Tabla 6.

Definición de variables

Valor	Definición operacional
0(No fuga)	Socio activo que mantiene su vinculación con la cooperativa al 31/12/2024. No ha presentado solicitud de renuncia y continúa realizando operaciones financieras de forma regular.
1(Fuga)	Socio que presentó renuncia voluntaria formal ante la cooperativa, solicitando explícitamente su desafiliación y el retiro de sus fondos. Esta renuncia queda registrada documentalmente en el sistema BESTERP de CACSMM.

Nota. Definición de las variables utilizadas.

La distribución de la variable objetivo presenta un desbalance característico de problemas de churn: **48,669** socios activos (87.8%) frente a 6,763 socios que renunciaron (12.2%). Este ratio de desbalance (0.139) justifica el uso de técnicas de sobremuestreo sintético (SMOTE) durante el entrenamiento y la priorización de métricas como F1-Score y ROC-AUC sobre Accuracy.

La variable objetivo se definió a partir del registro institucional de renuncia voluntaria. Para los socios clasificados como fugados se verificó que las características predictoras correspondieran a información disponible antes de la renuncia. Los movimientos derivados de la liquidación o cierre posterior a la renuncia no fueron considerados como predictores.

3.7.6. Descripción de variables predictoras

Para facilitar el procesamiento y análisis de los datos, las variables predictoras se clasificaron en tres categorías: demográficas/socioeconómicas, características bancarias y comportamientos financieros.

Las variables demográficas/socioeconómicas representan las condiciones personales y económicas de los socios; las características bancarias describen su vinculación y los productos financieros registrados; y los comportamientos financieros comprenden los patrones de uso y las operaciones realizadas. Esta clasificación coincide con la matriz de operacionalización y con el diccionario de variables presentado en el Anexo 2.

En total, se utilizaron 16 variables predictoras originales: 10 numéricas y seis categóricas. Adicionalmente, el dataset incluyó el campo socio_id, empleado únicamente como identificador anonimizado y excluido del modelamiento.

Tabla 7.

Variables demográficas y socioeconómicas

Variable	Tipo	Rango / Categorías	Descripción
edad	int	18 – 85	Edad del socio en años cumplidos al 31/12/2024.
sexo	cat	M, F	Sexo biológico registrado.
estado_civil	cat	Casado, Conviviente, Soltero, Divorciado, Viudo	Estado civil declarado por el socio.

nivel_educativo	cat	Posgrado, Primaria, Secundaria, Técnico, Universitario	Máximo nivel educativo alcanzado.
ocupación	cat	Dependiente, Desempleado, Independiente, Jubilado	Tipo de ocupación laboral del socio.
actividad_economica	cat	Formal, Informal	Clasificación del sector económico del socio.
ingresos_mensuales	float	S/ 900 – S/ 4,734.55	Ingreso mensual declarado por el socio (en soles).

Nota. Variables demográficas y socioeconómicas consideradas en el modelamiento, con su tipo, rango o categorías.

Tabla 8.

Características bancarias del socio

Variable	Tipo	Rango observado	Descripción
productos_activos	int	1 – 5	Número de productos financieros que el socio mantiene activos (ahorros, préstamos, plazo fijo, etc.).
prestamos_otorgados	int	0 – 5	Cantidad de préstamos otorgados al socio en su historial.
antiguedad_meses	int	1 – 120	Meses transcurridos desde la afiliación del socio.

Nota. Variables relacionadas con la vinculación institucional y los productos financieros del socio.

Tabla 9.

Variables de comportamiento financiero

Variable	Tipo	Rango / Categorías	Descripción
historial_pagos	cat	Puntual, Tardío	Clasificación del comportamiento de pago del socio.
aportes_total	float	S/ 100 – S/ 22,207.39	Saldo acumulado de aportes del socio al cierre. Media: S/ 5,195.38.
retiros_ultimos_3m	float	S/ 0.25 – S/ 15,091.90	Monto total retirado en los últimos 3 meses.
movimiento_total_6m	float	S/ 100 – S/ 36,766.33	Suma de transacciones (depósitos + retiros) en los últimos 6 meses. Media: S/ 5,971.99.
frecuencia_uso_mensual	int	0 – 15	Número promedio de transacciones mensuales del socio.
mora_dias	int	0 – 120	Días de mora acumulados en préstamos vigentes.

Nota. Variables relacionadas con la actividad transaccional, los aportes y el comportamiento de pago del socio.

Construcción de la variable historial_pagos

La variable historial_pagos fue construida a partir de la clasificación crediticia de las obligaciones vigentes del socio registrada en el cierre mensual correspondiente a su fecha de referencia. Para los socios que permanecían activos, se utilizó la clasificación registrada al 31 de diciembre de 2024. Para los socios que renunciaron durante el año 2024, se utilizó la clasificación del cierre mensual inmediatamente anterior a la fecha de renuncia. En los casos en que el socio mantenía más de un crédito vigente, se consideró la clasificación de mayor riesgo entre sus obligaciones. Finalmente, se asignó la categoría Puntual cuando la clasificación resultante fue Normal y la categoría Tardío cuando correspondió a cualquier categoría distinta de Normal.

Además de las variables predictoras descritas, el dataset contiene el campo socio_id, de tipo cadena, cuyos valores se encuentran comprendidos entre S0000001 y S0055432. Este campo se utilizó únicamente como identificador anonimizado de cada registro y fue excluido antes del entrenamiento de los modelos, debido a que no representa una característica predictiva del socio.

3.7.7. Consideraciones éticas y anonimización

El dataset fue anonimizado previo a su uso en la investigación. Los identificadores de socios fueron reemplazados por códigos secuenciales (S0000001 a S0055432) que no permiten la identificación directa ni indirecta de los individuos. Se omitieron campos como nombre completo, número de documento de identidad, dirección y números de cuenta.

El acceso a los datos fue autorizado exclusivamente para fines de investigación académica, bajo compromiso de confidencialidad y uso ético de la información. Ningún dato individual será publicado ni compartido fuera del ámbito de la presente tesis.

3.7.8. Preprocesamiento y Feature Engineering

El preprocesamiento transformó el dataset crudo en una matriz de features lista para el modelamiento. El diagnóstico de valores nulos confirmó un dataset completo sin datos faltantes. La validación de data leakage temporal no detectó features con información del futuro. Se construyeron cinco features adicionales que capturan relaciones entre variables existentes. Estos ratios condensan comportamiento relativo del socio.

Tratamiento de valores atípicos

Durante la consulta utilizada para generar el archivo CSV se aplicaron reglas de consistencia relacionadas con el dominio de los datos. Se incluyeron únicamente socios con edad mayor o igual a 18 años y se excluyeron registros con valores negativos en las variables ingresos mensuales, aportes totales, retiros de los últimos tres meses y movimiento total de los últimos seis meses. Asimismo, se utilizó la cláusula DISTINCT para evitar la incorporación de filas completamente duplicadas en el resultado de la consulta.

Estas reglas permitieron diferenciar los registros inconsistentes de las observaciones estadísticamente extremas. Los valores que incumplían las

condiciones definidas para el dominio fueron excluidos durante la construcción del conjunto de datos. En cambio, un monto elevado no fue considerado automáticamente como un error, debido a que podía representar una operación financiera real y contener información relevante para la predicción de fuga.

No se aplicaron procedimientos automáticos de eliminación o recorte mediante rango intercuartílico, percentiles, winsorización ni límites superiores arbitrarios. Los valores elevados que cumplían las reglas de consistencia fueron conservados sin imponer topes artificiales, debido a que la magnitud de una observación, por sí sola, no constituía evidencia suficiente de ruido o error de calidad.

Tabla 10.

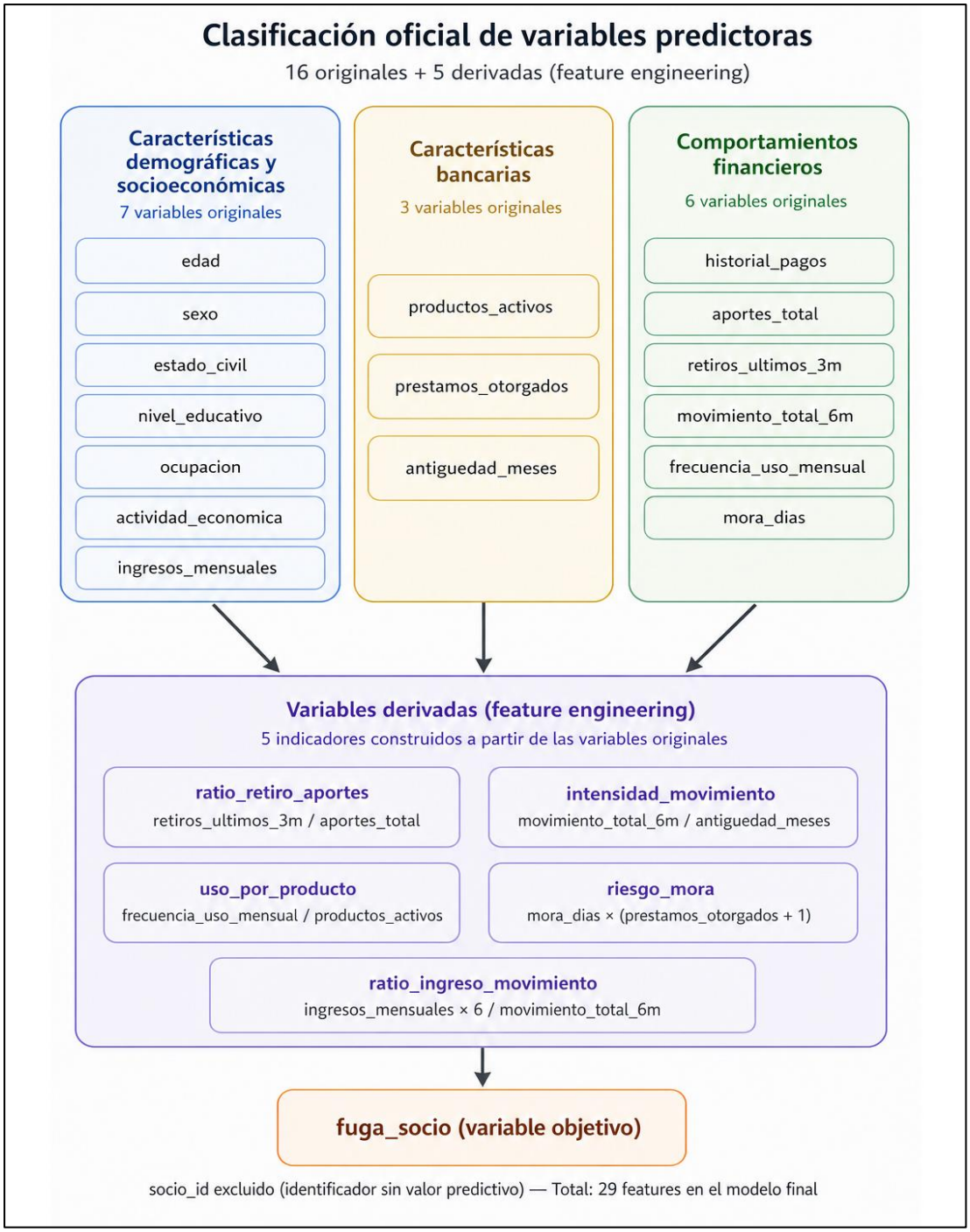
Variables derivadas construidas durante el feature engineering

Variable derivada	Fórmula	Significado económico
ratio_retiro_aportes	retiros_3m / aportes_total	Proporción de retiros recientes respecto de los aportes acumulados
intensidad_movimiento	movimiento_6m / antigüedad	Actividad transaccional relativa
uso_por_producto	frecuencia / productos_activos	Engagement por producto
riesgo_mora	mora_dias × (préstamos + 1)	Riesgo crediticio ponderado
ratio_ingreso_movimiento	ingreso×6 / movimiento_6m	Capacidad de ahorro relativa

Nota. Se detalla variables y su fórmula, Cuando el denominador presentó un valor igual a cero, la característica derivada fue asignada a cero.

Figura 3.

Clasificación oficial de variables predictoras



Las variables categóricas (sexo, estado_civil, nivel_educativo, ocupación, actividad_económica e historial_pagos) se codificaron mediante one-hot encoding con cardinalidad ≤ 10. La matriz final quedó con 30 columnas (29 features + 1 target), de las cuales todas son numéricas.

Tabla 11.

Codificación de variables categóricas

Variable	Cardinalidad	Método	Columnas Generadas
sexo	2 categorías	One-Hot Encoding (drop_first=True)	1 columna
estado_civil	5 categorías	One-Hot Encoding (drop_first=True)	4 columnas
nivel_educativo	5 categorías	One-Hot Encoding (drop_first=True)	4 columnas
ocupacion	4 categorías	One-Hot Encoding (drop_first=True)	3 columnas
actividad_economica	2 categorías	One-Hot Encoding (drop_first=True)	1 columna
historial_pagos	2 categorías	One-Hot Encoding (drop_first=True)	1 columna

Nota. Se detalla el método de codificación de las variables categóricas.

3.7.9. Preparación de datos para modelamiento

La preparación de datos para el modelamiento siguió las mejores prácticas de machine learning para problemas de clasificación binaria con desbalance de clases. Se aplicó una partición estratificada 80/20 del dataset, asegurando que ambos conjuntos (entrenamiento y prueba) mantuvieran la misma proporción de clases que el dataset original.

La división se realizó mediante train_test_split, utilizando test_size=0.20, stratify=y y random_state=42. Como resultado, el 80 % de los registros se asignó al conjunto de entrenamiento y el 20 % al conjunto de prueba, conservándose la proporción de ambas clases.

El identificador anonimizado del socio se mantuvo únicamente para fines de trazabilidad y no fue incorporado como variable predictora.

Tabla 12.

Preparación de datos

Aspecto	Detalle
Split	80% train / 20% test, estratificado
Train set	44,345 muestras (38,935 clase 0 + 5,410 clase 1)
Test set	11,087 muestras (9,734 clase 0 + 1,353 clase 1)
Validación cruzada	StratifiedKFold, 5 folds, shuffle=True

Balanceo de clases	SMOTE dentro de cada fold

Nota. Se prepararon los datos siguiendo las practicas del Machine Learning.

Es importante destacar que la técnica SMOTE (Synthetic Minority Over-sampling Technique) se aplicó exclusivamente dentro de cada fold de validación cruzada, nunca sobre el dataset completo, para evitar la filtración de información sintética al conjunto de prueba. Esta decisión metodológica es crucial para obtener estimaciones no sesgadas del rendimiento real del modelo.

3.7.10. Definición de modelos

Se seleccionaron 5 algoritmos que representan las principales familias de clasificación supervisada utilizadas en la literatura de predicción de churn. La Tabla presenta cada modelo con sus Hiperparámetros utilizados en la evaluación comparativa y la justificación de su inclusión.

Tabla 13.

Modelos definidos y sus hiper parámetros iniciales

Modelo	Hiperparámetros iniciales	Justificación
Logistic Regression	max_iter=1000, StandardScaler, SMOTE	Línea base interpretable; permite evaluar si un modelo lineal captura la señal de fuga
Random Forest	n_estimators=300, max_depth=15, min_samples_split=5, min_samples_leaf=2	Ensemble de árboles robusto a distribuciones asimétricas; manejo natural de variables mixtas
Gradient Boosting	n_estimators=200, max_depth=5, learning_rate=0.1, subsample=0.8	Boosting secuencial que optimiza residuos; alta capacidad de captura de patrones complejos
XGBoost	n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8	Implementación optimizada de gradient boosting con regularización L1/L2 integrada
LightGBM	n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8	Crecimiento leaf-wise más eficiente; óptimo para datasets grandes (55,432 registros)

Nota. Todos los modelos utilizan SMOTE(random_state=42) para balanceo de clases dentro del pipeline.

Los hiperparámetros fueron definidos previamente y se mantuvieron constantes durante la evaluación comparativa. En los resultados oficiales no se utilizó una búsqueda automatizada de hiperparámetros. Esta decisión permitió comparar los algoritmos bajo un procedimiento homogéneo y conservar la correspondencia entre el modelo entrenado, el modelo evaluado y el modelo interpretado mediante SHAP.

3.7.11. Entrenamiento y validación cruzada

Para entrenar los modelos se realizó una división estratificada del conjunto de datos en conjuntos de entrenamiento (80 %) y prueba (20 %), manteniendo la proporción original de la variable objetivo (fuga/no fuga). Esta estrategia permite preservar la distribución de clases al momento de ajustar los modelos.

En el conjunto de entrenamiento se aplicó validación cruzada estratificada de k-folds, con $k = 5$, con el fin de evaluar de manera consistente la capacidad de generalización de los modelos. La validación cruzada estratificada asegura que cada pliegue tenga una proporción similar de la clase minoritaria (fuga), lo cual es crítico en problemas de clasificación con desbalance de clases.

El balanceo de clases se integró dentro del esquema de validación cruzada mediante la técnica SMOTE, aplicada sobre cada pliegue durante el entrenamiento para evitar filtraciones o sesgos. De esta forma, los modelos fueron entrenados y evaluados internamente sobre grupos balanceados, reduciendo el impacto negativo que el desbalance original pudiera tener en el aprendizaje supervisado.

Los algoritmos considerados —Regresión Logística, Random Forest, Gradient Boosting, XGBoost y LightGBM— fueron entrenados utilizando sus implementaciones estándar en Python con bibliotecas como Scikit-Learn y XGBoost/LightGBM, ajustando únicamente parámetros básicos de entrenamiento sin búsqueda automatizada de hiperparámetros, dado que la optimización hiperparamétrica no forma parte del alcance de esta investigación.

Cabe precisar que el cuaderno de trabajo adjunto (Anexo 4) incluye etapas exploratorias adicionales, tales como una búsqueda aleatoria de hiperparámetros (RandomizedSearchCV, con validación interna de 3 pliegues por costo computacional) y un barrido del umbral de decisión, desarrolladas con fines de exploración técnica complementaria. Dichas etapas no forman parte del alcance metodológico de la presente investigación y sus salidas no se utilizan en los resultados reportados. Por consistencia y trazabilidad, la totalidad de las métricas,

las matrices de confusión, el análisis de interpretabilidad SHAP y el modelo final descritos en el Capítulo IV corresponden al mismo pipeline base —SMOTE integrado en los pliegues de entrenamiento y clasificador con los hiperparámetros iniciales declarados— evaluado con el umbral de decisión por defecto de 0.50 sobre un conjunto de prueba que no fue utilizado en el entrenamiento ni en ninguna decisión de ajuste del modelo. De esta manera, aunque el conjunto de prueba permaneció separado del entrenamiento y del ajuste interno de los modelos, sus métricas sí se utilizaron para comparar los cinco algoritmos y seleccionar el modelo final. Las cifras reportadas no incorporan los resultados de la optimización exploratoria del umbral sobre esos mismos datos.

3.7.12. Evaluación de desempeño

Para evaluar el desempeño de los modelos se emplearon métricas estándar de clasificación supervisada, enfocadas en medir tanto la calidad de las predicciones como la capacidad de distinguir entre socios con y sin riesgo de fuga. Las métricas consideradas fueron:

- Exactitud (accuracy): proporción de predicciones correctas respecto del total de casos evaluados. Debido al desbalance entre las clases, esta métrica se interpretó juntamente con las demás medidas de desempeño.
- Precisión (Precision): proporción de verdaderos positivos entre todas las predicciones positivas, útil para medir la exactitud de las predicciones de fuga.
- Exhaustividad (Recall): proporción de verdaderos positivos recuperados entre todos los casos positivos reales, importante en contextos donde no se desea omitir casos de fuga reales.
- F1-Score: medida armónica entre precisión y recall que equilibra ambos criterios.
- Área bajo la curva ROC (ROC-AUC): métrica agregada que evalúa la capacidad discriminativa de un modelo en distintos puntos del threshold de decisión.
- Área bajo la curva precisión-recall (PR-AUC): es una métrica en machine learning que evalúa qué tan bueno es un modelo de clasificación, resume la relación entre la exactitud de las predicciones positivas (Precisión) y la cantidad de casos positivos reales encontrados (Exhaustividad/Recall).

El F1-score se estableció como criterio principal de comparación porque permite evaluar conjuntamente la precisión y la sensibilidad respecto de la clase minoritaria. ROC-AUC y PR-AUC se utilizaron como métricas complementarias de discriminación. La selección no se sustentó en una sola métrica, sino en el análisis conjunto del desempeño obtenido sobre el conjunto de prueba.

Para la clasificación final se utilizó un umbral de decisión de 0,50. No se aplicó optimización del umbral en los resultados presentados en el Capítulo IV.

Las métricas se calcularon sobre el conjunto de prueba, el cual no fue utilizado durante el entrenamiento ni durante la validación cruzada. Para calcular la exactitud, la precisión, la exhaustividad, la puntuación F1 y la matriz de confusión, se empleó el umbral de decisión predeterminado de 0.5, mediante el cual los puntajes generados por los modelos se transformaron en predicciones binarias. Este umbral no fue optimizado dentro del alcance de la investigación; sin embargo, su uso se documentó para mantener la transparencia y uniformidad de la comparación.

Por otro lado, las métricas ROC-AUC y PR-AUC se calcularon directamente a partir de los puntajes o probabilidades estimadas por cada modelo, considerando diferentes umbrales de clasificación. En consecuencia, sus valores no dependen únicamente del umbral de 0.5 utilizado para obtener las predicciones binarias.

No se aplicaron procedimientos de calibración de probabilidades. Por ello, las salidas de los modelos deben interpretarse principalmente como puntajes que permiten ordenar a los socios según su propensión a pertenecer a la clase fuga y no como probabilidades calibradas de que el evento ocurra. Su utilización operativa como estimaciones directas de riesgo individual requeriría una etapa adicional de calibración, mediante procedimientos como curvas de calibración, puntuación de Brier o CalibratedClassifierCV, aspecto que se plantea como una mejora futura.

Finalmente, la matriz de confusión se utilizó como herramienta complementaria para analizar la distribución de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Esto permitió interpretar con mayor detalle los aciertos y errores de clasificación de cada modelo.

3.7.13. Interpretabilidad mediante SHAP

Esta sección responde al tercer objetivo específico de la investigación, orientado a interpretar las predicciones del modelo de mejor desempeño mediante

SHAP, con el propósito de comprender la contribución de cada variable en las predicciones realizadas por los modelos de aprendizaje automático, se empleó la técnica SHAP (SHapley Additive exPlanations). Este método se basa en la teoría de juegos cooperativos y permite cuantificar el aporte individual de cada característica sobre la predicción generada por un modelo.

La aplicación de SHAP permitió identificar cuáles variables ejercen mayor influencia en la probabilidad de fuga de los socios, así como determinar si cada característica incrementa o reduce dicha probabilidad. A diferencia de las medidas tradicionales de importancia de variables, SHAP proporciona explicaciones consistentes tanto a nivel global como individual.

Para los modelos seleccionados se calcularon los valores SHAP utilizando las implementaciones disponibles en Python. Posteriormente, se generaron gráficos de importancia global para visualizar el impacto promedio de cada variable sobre las predicciones del modelo. Estos resultados permitieron identificar los factores más relevantes asociados a la fuga de socios y complementar la evaluación del desempeño predictivo.

La interpretación de los valores SHAP se realizó considerando la magnitud absoluta de la contribución de cada variable y la dirección de su efecto sobre la predicción, facilitando una comprensión más transparente del comportamiento de los modelos desarrollados.

Capítulo IV

Resultados y discusión

El objetivo general de la presente investigación es desarrollar y comparar modelos de machine Learning de clasificación supervisada, basados en características bancarias y comportamientos financieros de los socios, para identificar el modelo con mejor desempeño en la predicción del riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.

En correspondencia con este objetivo, el capítulo presenta los resultados organizados de acuerdo con los objetivos específicos. En primer lugar, se exponen los resultados del análisis y procesamiento de las características bancarias y los comportamientos financieros de los socios. Seguidamente, se presentan el desarrollo y la evaluación de los cinco modelos considerados en el estudio: regresión logística, Random Forest, Gradient Boosting, XGBoost y LightGBM. Luego, se comparan sus resultados mediante las métricas de exactitud, precisión, recall, F1-score y ROC-AUC, con la finalidad de identificar el modelo con mejor desempeño. Finalmente, se desarrolla la interpretación del modelo seleccionado mediante SHAP, a fin de reconocer la contribución de las características analizadas en sus predicciones.

4.1. Caracterización del dataset final

Resumen Estadístico del Dataset Final

Tabla 14.

Resumen del Dataset

Característica	Valor
Registros totales	55,432 socios
Variables predictoras	16 (10 numéricas + 6 categóricas)
Identificador	socio_id (anonimizado y excluido del modelamiento)
Variable objetivo	fuga_socio (binaria: 0 = activo, 1 = renuncia voluntaria)
Total de campos del dataset	18 (16 predictoras, 1 objetivo y 1 identificador)
Tasa de churn	12.2% (6,763 renuncias / 55,432 total)
Ratio de desbalance	0.139 (clase minoritaria / clase mayoritaria)
Valores nulos	0

Fecha de corte	31 de diciembre de 2024
Fuente	Sistema BESTERP de la CACSMM — Ayacucho, Perú

Nota. El campo socio_id se utilizó únicamente como identificador anonimizado y fue excluido del modelamiento. La variable fuga_socio constituyó la variable objetivo.

Limitaciones del Dataset

Se reconocen las siguientes limitaciones inherentes al dataset y al proceso de construcción:

1. **Sesgo de selección por exclusión de socios inactivos.** Al incluir únicamente socios con actividad en los últimos 24 meses, el modelo no captura patrones de desvinculación gradual previos a la inactividad total. Los socios que transitan lentamente hacia la inactividad antes de abandonar están subrepresentados. Sin embargo, este sesgo es mitigado por la política de depuración institucional, que absorbe la mayoría de estos casos como abandonos pasivos no accionables.
2. **Ausencia de validación temporal (out-of-time).** El split de entrenamiento/test se realizó de forma aleatoria estratificada, no temporal. Si existen patrones estacionales en las renuncias (fin de año, inicio escolar, campañas de competidores), la evaluación del modelo podría sobreestimar su rendimiento en un despliegue real con datos futuros.
3. **Variables limitadas al sistema transaccional.** El dataset no incluye variables externas potencialmente relevantes como distancia a la agencia, presencia de competidores en la zona, satisfacción del socio, interacciones con servicio al socio o eventos macroeconómicos regionales. Estas variables podrían mejorar la capacidad predictiva del modelo.

4.2. Estadísticas descriptivas de las variables numéricas

El análisis descriptivo de las variables numéricas revela que la edad promedio de los socios es de 44.75 años, con un rango de 18 a 85 años aproximadamente. Los ingresos mensuales presentan una media de S/ 1,680.46, la antigüedad promedio como socio es de 56.03 meses indicando que la base de datos captura tanto socios recientes como aquellos con más de 8 años de relación con la cooperativa.

En cuanto a las características bancarias y los comportamientos financieros, los socios mantienen en promedio 2.76 productos activos y realizan 4.84 operaciones mensuales. El monto promedio de retiros en los últimos 3 meses es de S/ 1,530.84, con una dispersión elevada que sugiere heterogeneidad en los patrones de retiro. Los días promedio de mora son 6.28, lo que indica que la mayoría de los socios no presenta morosidad significativa, aunque existe una cola derecha con valores extremos.

Figura 4.

Estadística descriptiva de variables numéricas

	count	mean	std	min	25%	50%	75%	max
edad	55432.0	44.751479	14.464869	18.00	34.0000	45.000	55.000	85.00
ingresos_mensuales	55432.0	1680.462530	670.424244	900.00	1056.4750	1594.480	2132.125	4734.55
antiguedad_meses	55432.0	56.029045	30.330860	1.00	35.0000	57.000	78.000	120.00
productos_activos	55432.0	2.758930	1.201134	1.00	2.0000	3.000	4.000	5.00
frecuencia_uso_mensual	55432.0	4.835312	3.275841	0.00	2.0000	5.000	7.000	15.00
prestamos_otorgados	55432.0	1.200714	1.225734	0.00	0.0000	1.000	2.000	5.00
mora_dias	55432.0	6.283645	17.404845	0.00	0.0000	0.000	0.000	120.00
aportes_total	55432.0	5195.377309	3631.330595	100.00	2266.3550	4963.390	7689.625	22207.39
retiros_ultimos_3m	55432.0	1530.842491	1617.881331	0.25	367.6775	1045.955	2146.000	15091.90
movimiento_total_6m	55432.0	5971.988109	4969.369342	100.00	1984.4300	5081.100	8816.645	36766.33
fuga_socio	55432.0	0.122005	0.327295	0.00	0.0000	0.000	0.000	1.00

Nota. Resultados estadísticos

4.3. Análisis Exploratorio de Datos (EDA)

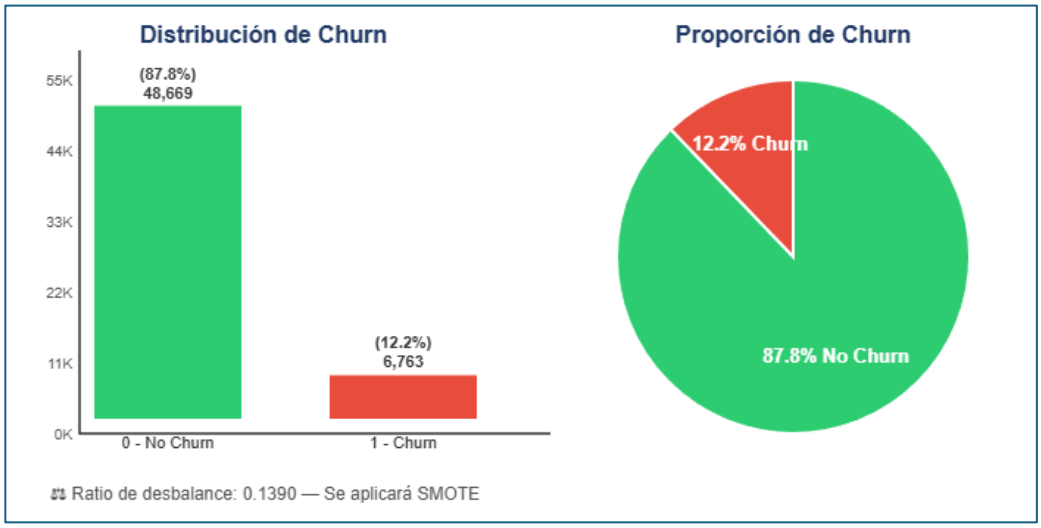
Se realizó un análisis exploratorio en cuatro dimensiones: distribución de la variable objetivo, distribución de variables numéricas, matriz de correlación con la variable objetivo, y análisis de variables categóricas segmentadas por churn.

Distribución de la variable objetivo y correlación de pearson

Se analizó la distribución de la variable fuga_socio para cuantificar el desbalance de clases. Este paso es fundamental porque un desbalance pronunciado invalida el uso de Accuracy como métrica primaria y exige estrategias especiales durante el entrenamiento.

Figura 5.

Distribución de Churn



Nota. Graficos de las distribuciones y proporciones de Churn.

Distribución de fuga_socio: No Churn=87.8% (48,669 socios), Churn=12.2% (6,763 socios). Ratio desbalance=0.139 → se aplica SMOTE.

El ratio de desbalance es 0.139, confirmando que el dataset es altamente desbalanceado (< 0.2). Este resultado justifica: (a) el uso de SMOTE para oversampling de la clase minoritaria, (b) la evaluación mediante F1-Score y ROC-AUC en lugar de Accuracy.

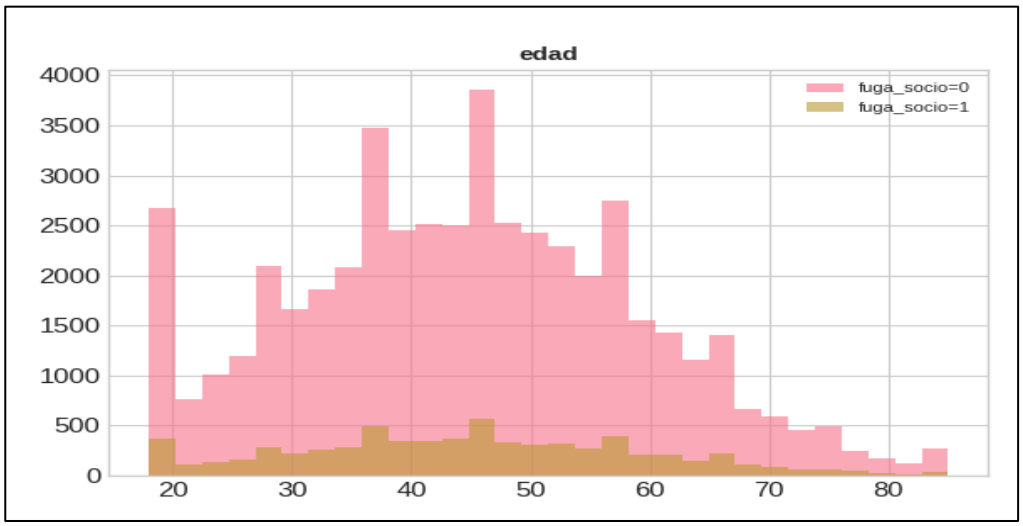
Distribución de variables numéricas y correlacion de pearson

Se analizaron las distribuciones de las 10 variables numéricas del dataset (55,432 registros), segmentadas por la variable objetivo fuga_socio. El análisis combinó histogramas superpuestos y correlaciones de Pearson para clasificar las variables según su capacidad de discriminar entre socios activos y fugados.

La Figura 6 presenta la distribución de la edad de los socios según su condición de fuga. Se observa que ambas clases comprenden principalmente edades adultas, con una mayor concentración de registros aproximadamente entre los 30 y 60 años. Aunque los socios no fugados muestran frecuencias superiores debido a su mayor presencia en el conjunto de datos, existe un considerable solapamiento entre ambas clases. Por consiguiente, la edad puede aportar información al modelo, pero su distribución indica que no permitiría diferenciar por sí sola a los socios fugados de los no fugados.

Figura 6.

Distribución de fuga de socios por edad

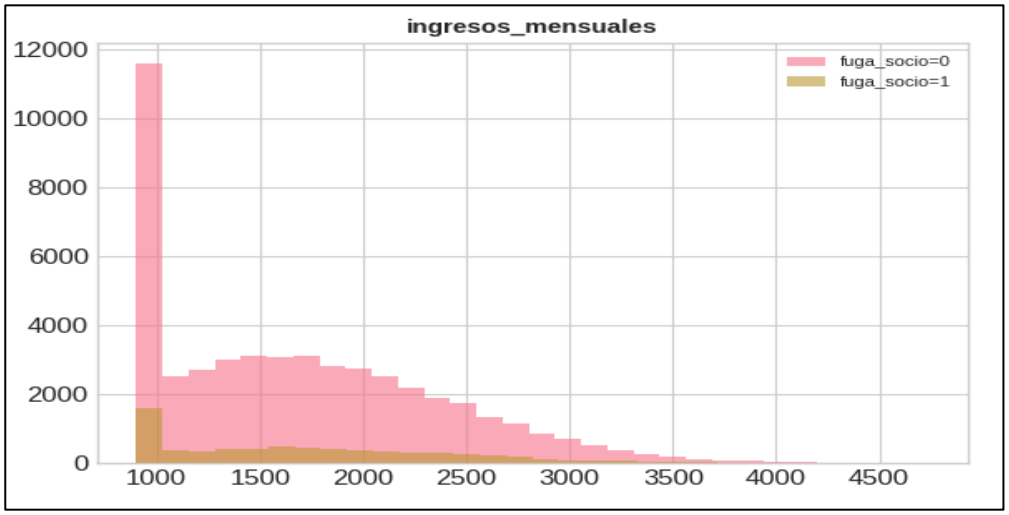


Nota. Distribución de la variable edad.

La Figura 7 muestra la distribución de los ingresos mensuales de los socios en función de su condición de fuga. En ambas clases, los registros se concentran en los niveles inferiores de ingreso y disminuyen conforme este aumenta, formando una distribución asimétrica hacia la derecha. Asimismo, se observa una cola prolongada correspondiente a socios con ingresos considerablemente superiores. Esta característica evidencia la necesidad de considerar la asimetría y los valores extremos durante el modelamiento, especialmente en los algoritmos sensibles a la escala de las variables.

Figura 7.

Distribución de fuga de socios por ingresos mensuales

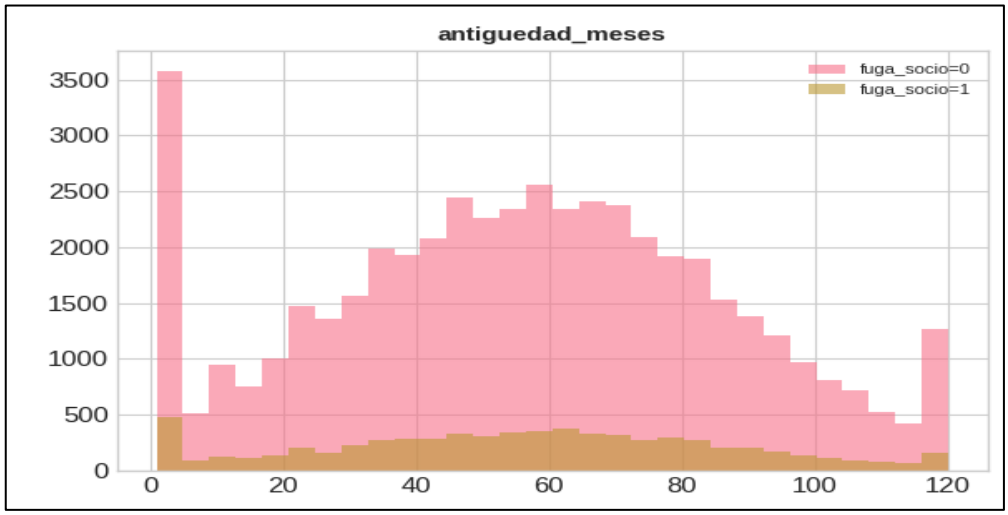


Nota. Distribución de la variable ingresos mensuales.

La Figura 8 presenta la antigüedad de los socios, expresada en meses, según su condición de fuga. La distribución muestra una elevada frecuencia de registros con poca antigüedad, seguida de una concentración gradual en valores intermedios y una disminución hacia los periodos más extensos. También se observa la presencia de socios cercanos al límite superior registrado. La forma de la distribución sugiere que la antigüedad podría aportar información relacionada con la permanencia del socio; sin embargo, el solapamiento entre las clases requiere que esta variable sea evaluada juntamente con las demás características.

Figura 8.

Distribución de fuga de socios por antigüedad

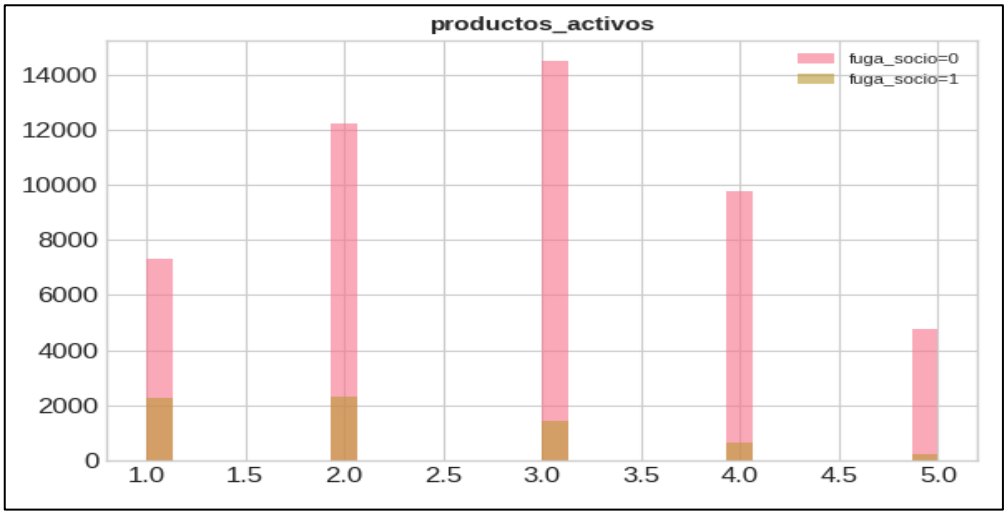


Nota. Distribución de la variable antigüedad (meses).

La Figura 9 muestra la cantidad de productos activos de los socios de acuerdo con su condición de fuga. La variable presenta valores discretos entre uno y cinco productos, con una mayor concentración general entre dos y cuatro. En los socios fugados se aprecia una presencia relativamente mayor en las cantidades más bajas, mientras que la frecuencia disminuye conforme aumenta el número de productos. Este comportamiento permite considerar que el nivel de vinculación mediante productos activos podría contribuir a la clasificación, aunque debe analizarse juntamente con otras variables financieras y transaccionales.

Figura 9.

Distribución de fuga de socios por productos activos

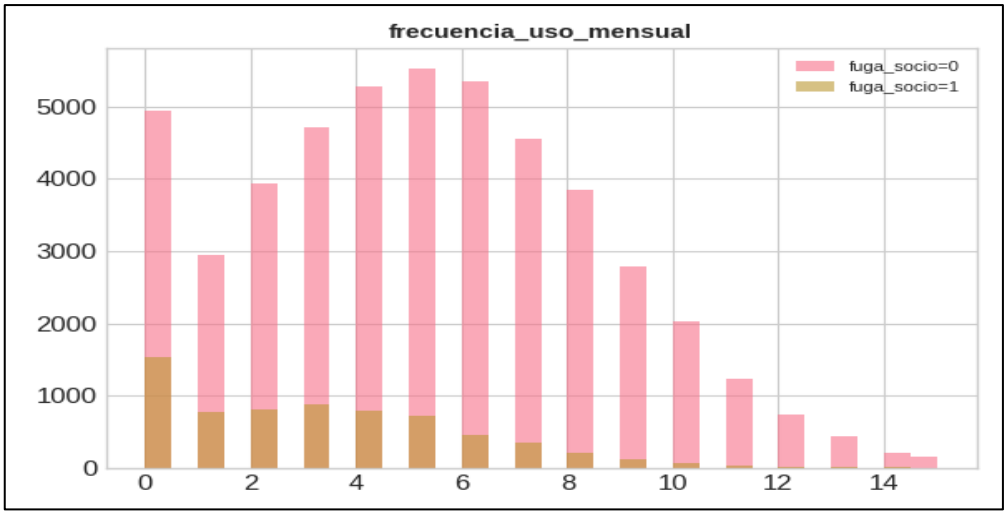


Nota. Distribución de la variable productos activos.

La Figura 10 presenta la frecuencia mensual de uso de los servicios según la condición de fuga de los socios. Los registros se concentran principalmente entre cero y diez usos mensuales, mientras que las frecuencias superiores son progresivamente menos comunes. Visualmente, los socios fugados presentan una mayor concentración en los niveles más bajos de uso y su frecuencia disminuye conforme aumenta la utilización mensual. Esta diferencia sugiere que una menor actividad podría aportar información relevante para la clasificación de la fuga, sin que ello implique por sí solo una relación causal.

Figura 10.

Distribución de fuga de socios por frecuencia de uso mensual

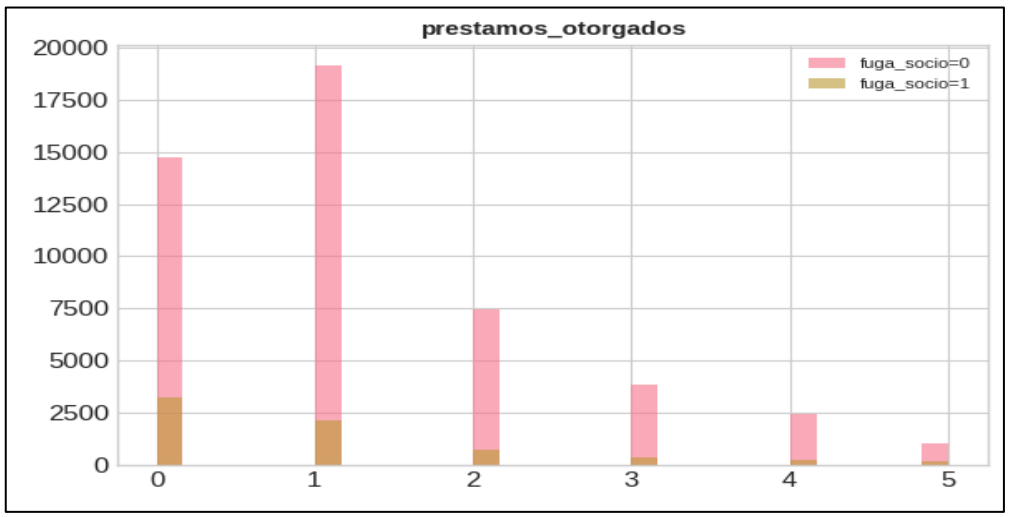


Nota. Distribución de la variable frecuencia de uso (mensual).

La Figura 11 muestra la cantidad de préstamos otorgados a los socios en función de su condición de fuga. La distribución es discreta y se concentra principalmente entre cero y dos préstamos, mientras que las frecuencias disminuyen a partir de tres. En ambas clases se observa una presencia importante de socios con uno o ningún préstamo, por lo que existe solapamiento entre sus distribuciones. En consecuencia, esta variable puede contribuir a representar el grado de vinculación crediticia, pero no ofrece una separación suficiente cuando se analiza de manera individual.

Figura 11.

Distribución de fuga de socios por préstamos otorgados

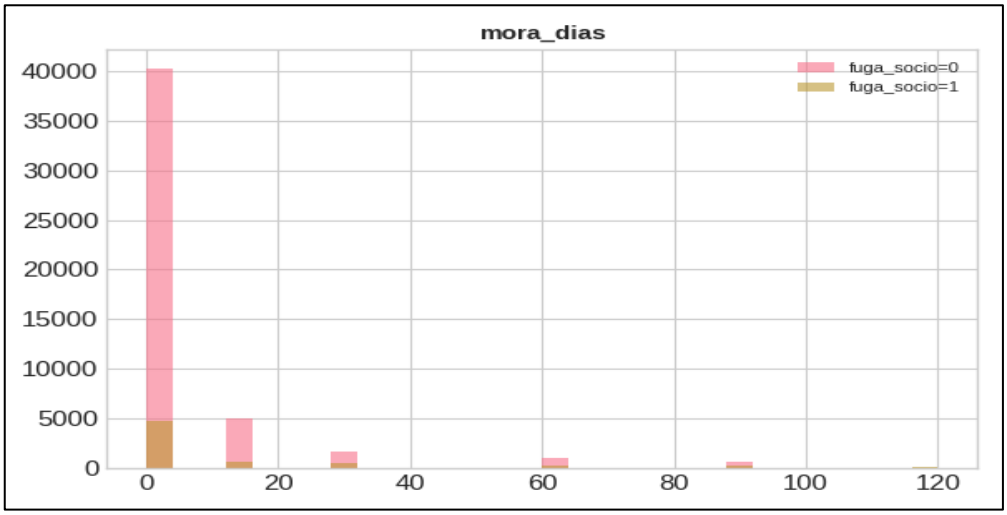


Nota. Distribución de la variable préstamos otorgados.

La Figura 12 presenta la distribución de los días de mora según la condición de fuga de los socios. En ambas clases se observa una concentración predominante en cero o pocos días de atraso y una cola extendida hacia valores elevados, lo que evidencia una distribución marcadamente asimétrica y la presencia de observaciones extremas. Debido a esta concentración, la ausencia de mora es una condición frecuente tanto entre los socios fugados como entre los no fugados. Por ello, la variable debe interpretarse juntamente con las demás características y recibir un tratamiento compatible con su asimetría durante el modelamiento.

Figura 12.

Distribución de fuga de socios por mora en días

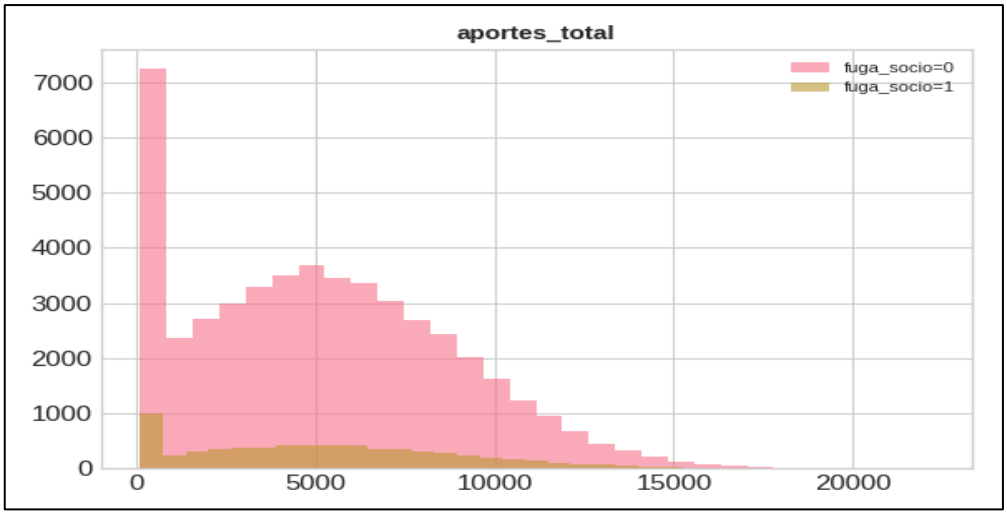


Nota. Distribución de la variable mora (días).

La Figura 13 muestra la distribución de los aportes totales de los socios de acuerdo con su condición de fuga. La mayor parte de los registros se concentra en los niveles bajos e intermedios, mientras que la frecuencia disminuye conforme aumenta el monto acumulado, generándose una cola hacia la derecha. Aunque las distribuciones de ambas clases presentan solapamiento, se observan diferencias en su concentración a lo largo de los montos registrados. Esta variable puede aportar información sobre el nivel de participación económica del socio, pero no debe interpretarse aisladamente como causa de la fuga.

Figura 13.

Distribución de fuga de socios por aportes totales

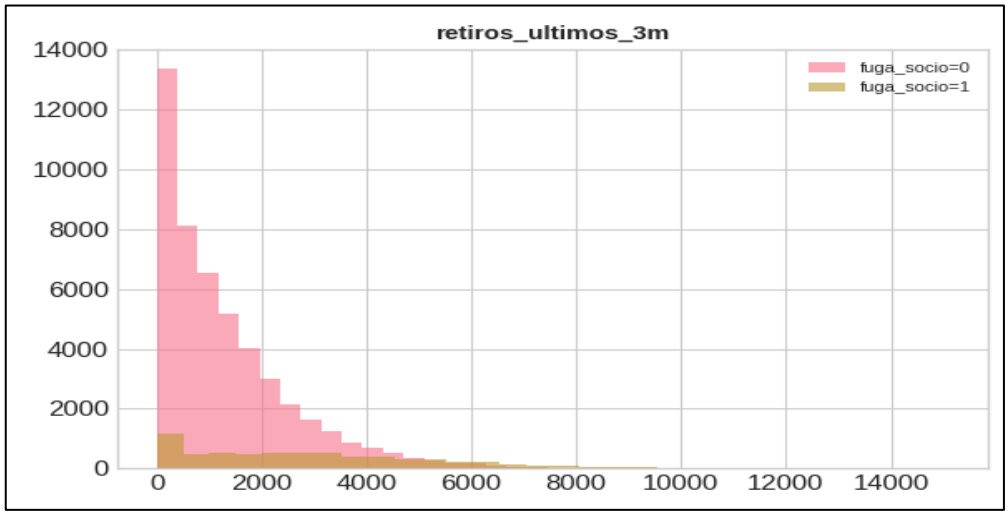


Nota. Distribución de la variable aportes en función del total.

La Figura 14 presenta la distribución de los retiros realizados durante los últimos tres meses según la condición de fuga. Los valores se concentran principalmente cerca de cero y disminuyen rápidamente conforme aumenta el monto retirado, mientras que un grupo reducido presenta valores considerablemente elevados. Este comportamiento evidencia una marcada asimetría positiva y la presencia de una cola prolongada. Por consiguiente, los retiros pueden proporcionar información sobre la actividad financiera reciente, aunque su distribución requiere considerar el efecto de los valores extremos en los modelos sensibles a la escala.

Figura 14.

Distribución de fuga de socios por retiros

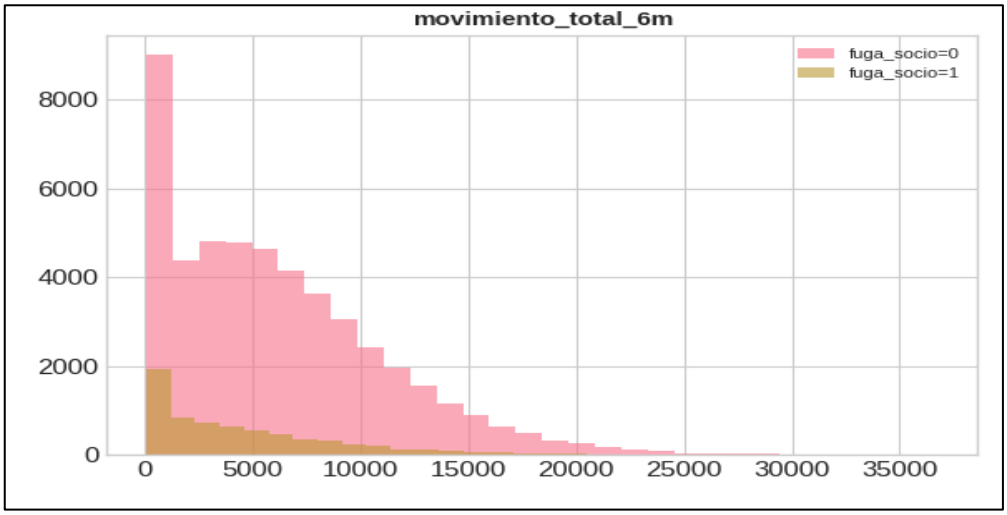


Nota. Distribución de la variable retiros (últimos 3 meses).

La Figura 15 muestra la distribución del monto total de movimientos efectuados durante los últimos seis meses según la condición de fuga. En ambas clases se observa una mayor concentración en los montos bajos, seguida de una reducción gradual de la frecuencia a medida que aumenta el volumen de movimientos. La cola extensa hacia valores superiores evidencia asimetría positiva y la existencia de socios con una actividad transaccional considerablemente mayor. Esta variable permite representar el nivel reciente de actividad financiera y puede contribuir al modelo al combinarse con las demás características analizadas.

Figura 15.

Distribución de fuga de socios por movimientos



Nota. Distribución de la variable movimientos (últimos 6 meses).

- Hallazgo 1: Las variables financieras recientes y de vinculación bancaria presentan mayor asociación lineal con la fuga.

Las correlaciones absolutas más elevadas con la condición de fuga correspondieron a variables relacionadas con el comportamiento financiero reciente —retiros_ultimos_3m, frecuencia_uso_mensual, mora_dias y movimiento_total_6m— y con las características bancarias, representada por productos_activos. Estas variables registraron valores de |r| entre 0.12 y 0.35. En contraste, las variables de perfil y stock —edad, ingresos_mensuales, antiguedad_meses y aportes_total— presentaron correlaciones lineales prácticamente nulas (|r| < 0.01).

Los valores cercanos a cero de estas últimas variables indican que, consideradas individualmente, no muestran una asociación lineal apreciable con la condición de fuga. Esto no significa que sean irrelevantes, que exista igualdad de probabilidades entre los socios ni que las variables sean independientes de la fuga, pues podrían intervenir relaciones no lineales, interacciones con otras características o efectos condicionados por el comportamiento financiero reciente.

- Hallazgo 2: Ranking de correlaciones con fuga_socio

La tabla presenta las correlaciones de Pearson de las 10 variables numéricas con la variable objetivo, ordenadas de mayor a menor valor absoluto. Se identifican tres niveles de asociación.

Tabla 15.

Correlación de Pearson de variables numéricas con fuga_socio

Variable	|r| con fuga	Nivel	Interpretación
retiros_ultimos_3m	0.3480	Alto	Proporción de retiros recientes respecto de los aportes acumulados antes de la salida
frecuencia_uso_mensual	0.1964	Alto	Desenganche operativo con la cooperativa
productos_activos	0.1901	Alto	Menor vinculación = menor barrera de salida
mora_dias	0.1520	Alto	Morosidad asociada a exclusión o retiro voluntario
movimiento_total_6m	0.1155	Alto	Pico de actividad previo a la fuga
prestamos_otorgados	0.0885	Moderado	Préstamo activo como ancla institucional
ingresos_mensuales	0.0080	Nulo	Asociación lineal prácticamente nula con la fuga
edad	0.0038	Nulo	Asociación lineal prácticamente nula con la fuga
antiguedad_meses	0.0031	Nulo	Asociación lineal prácticamente nula con la fuga
aportes_total	0.0005	Nulo	Asociación lineal prácticamente nula con la fuga

Nota: Verde = alto poder discriminatorio, amarillo = moderado, sin color = nulo.

- Hallazgo 3: Tres patrones asociados con la condición de fuga

Del análisis de distribuciones se identificaron tres patrones de conducta que caracterizan a los socios clasificados como fugados:

Retiros crecientes. Los socios fugados se concentran en los valores altos de retiros_ultimos_3m (superiores a S/ 4,000), evidenciando un proceso de proporción de retiros recientes respecto de los aportes acumulados. Este es el indicador más potente (r = 0.35) y motivó la creación de la variable derivada ratio_retiro_aportes, que resultó ser el predictor más importante del modelo según SHAP.

Baja frecuencia de uso. Los churners se agrupan en frecuencias de 0 a 3 operaciones mensuales, mientras que son casi inexistentes entre quienes realizan 8 o más operaciones. La baja frecuencia de uso presentó asociación con la condición de fuga.

Escasa vinculación con productos. Los socios con 1 o 2 productos activos presentan proporción significativamente mayor de fuga que aquellos con 4 o 5 productos. La vinculación múltiple genera un mayor costo de cambio percibido que actúa como barrera natural de retención.

- Hallazgo 4: Variables nulas individualmente generan features potentes al combinarse

Las variables aportes_total (r = 0.0005) y antiguedad_meses (r = 0.0031), que individualmente no discriminan entre clases, resultaron fundamentales como componentes de variables derivadas. El ratio ratio_retiro_aportes (retiros_ultimos_3m / aportes_total) se convirtió en el predictor más importante del modelo final según SHAP, y intensidad_movimiento (movimiento_total_6m / antiguedad_meses) también aportó poder predictivo significativo. Este resultado subraya que el Feature Engineering basado en conocimiento del dominio cooperativo es esencial: las relaciones entre variables capturan dinámicas financieras que los valores absolutos por sí solos no revelan.

4.4. Análisis de variables categóricas

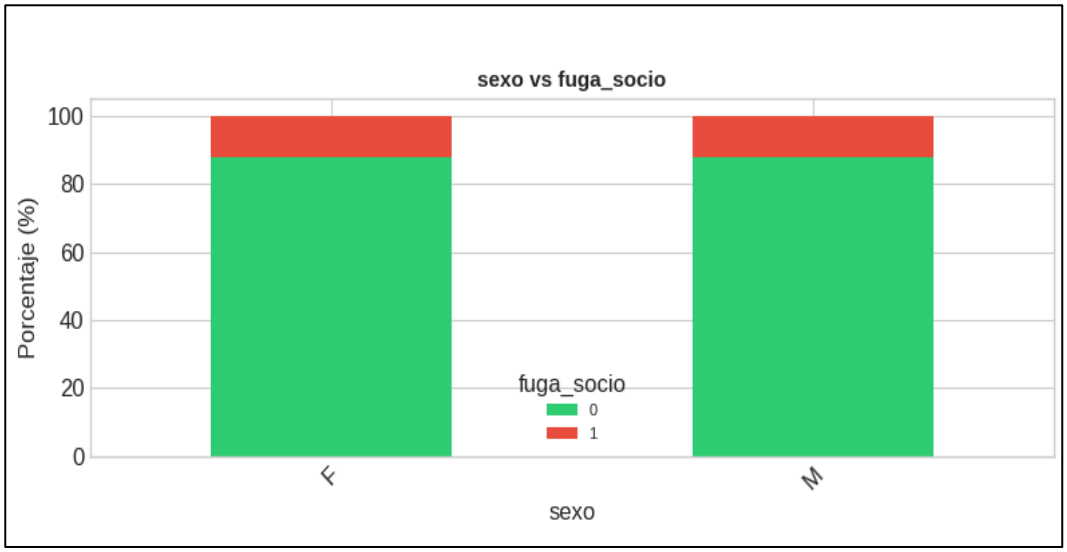
Se analizaron las 6 variables categóricas del dataset mediante gráficos de barras apiladas al 100%, donde cada barra representa la proporción de socios activos (fuga_socio=0, verde) y fugados (fuga_socio=1, rojo) dentro de cada categoría. Este tipo de visualización permite comparar directamente las tasas de churn entre categorías, independientemente del tamaño de cada grupo. La tasa global de churn del dataset es de 12.2%, valor que sirve como línea base para evaluar si alguna categoría se desvía significativamente.

La Figura 16 presenta la distribución porcentual de la condición de fuga según el sexo de los socios. Se observa que la proporción de socios fugados es muy similar entre las categorías femenino y masculino, pues en ambos casos representa aproximadamente la décima parte de los registros. Por tanto, la variable sexo no muestra visualmente diferencias marcadas entre sus categorías y su capacidad de

clasificación deberá evaluarse juntamente con las demás características incorporadas en el modelo.

Figura 16.

Distribución de fuga de socios por sexo

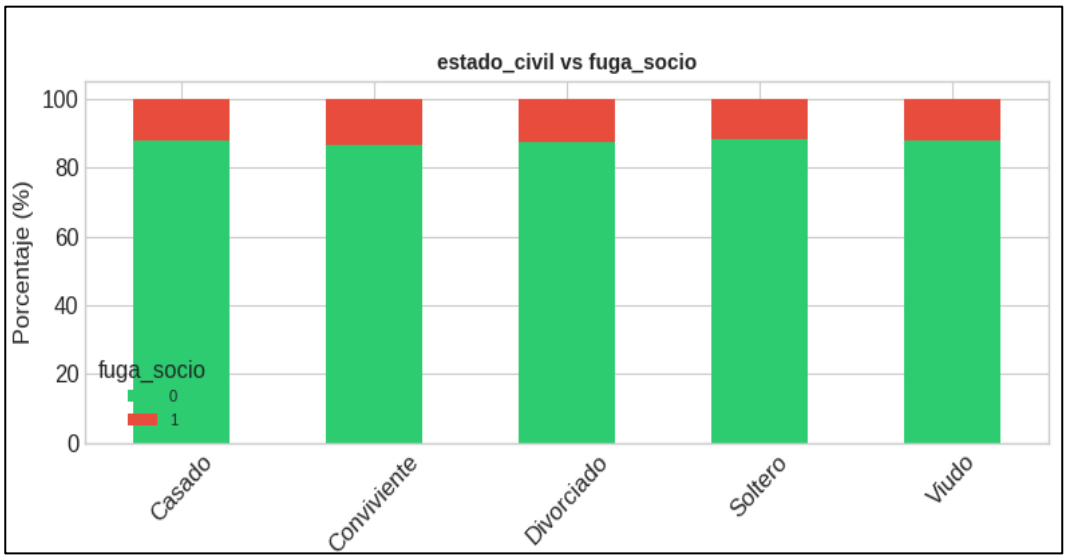


Nota. Distribución de la variable sexo.

La Figura 17 muestra la distribución porcentual de la fuga de socios según su estado civil. En todas las categorías predomina ampliamente la condición de no fuga, mientras que la proporción de socios fugados presenta variaciones reducidas. La categoría conviviente registra visualmente una proporción ligeramente mayor de fuga; sin embargo, las diferencias entre casados, convivientes, divorciados, solteros y viudos no son suficientemente pronunciadas para establecer una separación clara mediante esta variable de manera individual.

Figura 17.

Distribución de fuga de socios por estado civil

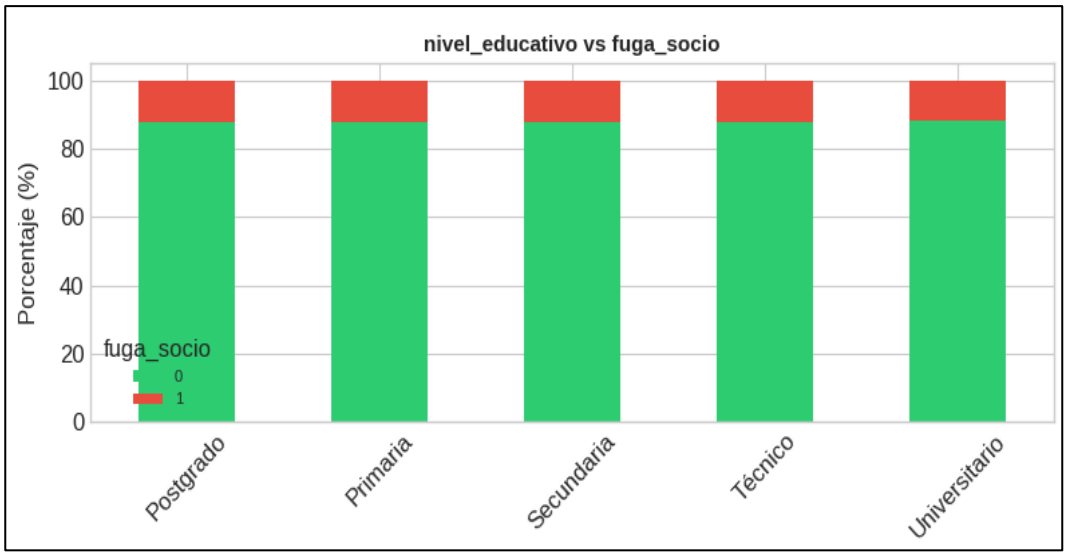


Nota. Distribución de la variable estado civil.

La Figura 18 presenta la distribución porcentual de la condición de fuga de acuerdo con el nivel educativo de los socios. Las categorías de postgrado, primaria, secundaria, formación técnica y educación universitaria muestran proporciones muy semejantes de socios fugados y no fugados. En consecuencia, no se advierte visualmente una categoría educativa con un comportamiento claramente diferenciado, por lo que esta característica debe analizarse en combinación con las demás variables del modelo.

Figura 18.

Distribución de fuga de socios por nivel educativo

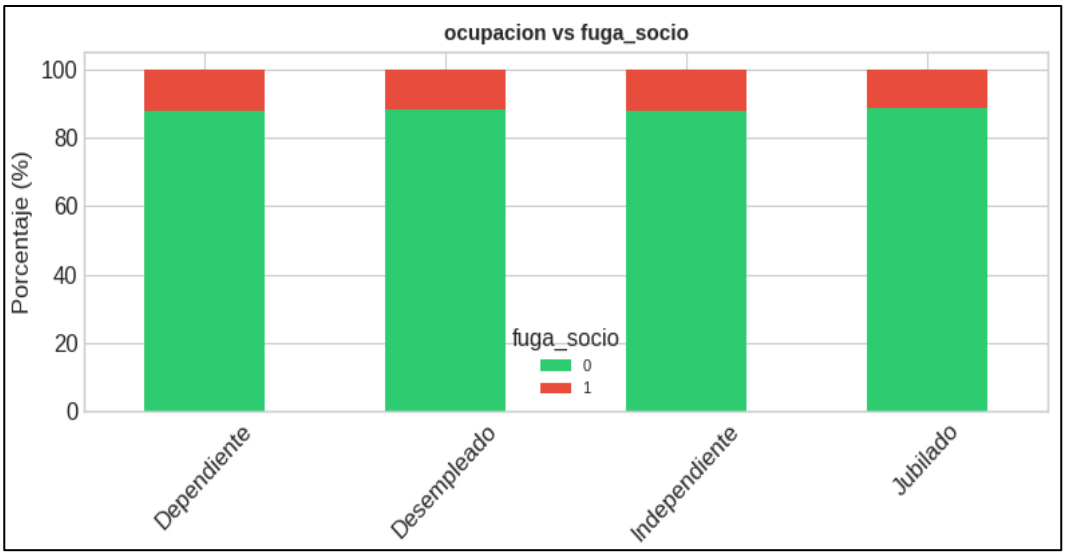


Nota. Distribución de la variable nivel educativo.

La Figura 19 muestra la distribución porcentual de la fuga según la ocupación de los socios. En las categorías dependiente, desempleado, independiente y jubilado, la proporción de socios no fugados es ampliamente mayor y presenta valores similares. Las diferencias observadas en la proporción de fuga son mínimas, lo cual indica que la ocupación no ofrece visualmente una separación marcada entre ambas clases cuando se examina de forma aislada.

Figura 19.

Distribución de fuga de socios por ocupación

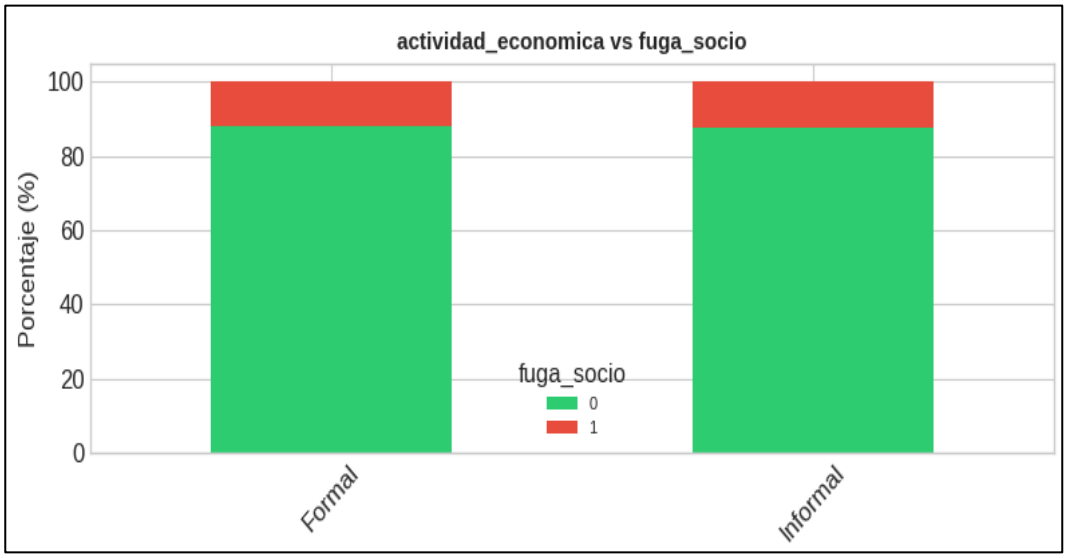


Nota. Distribución de la variable ocupación.

La Figura 20 presenta la distribución porcentual de la condición de fuga según la actividad económica formal o informal de los socios. Ambas categorías muestran una composición prácticamente equivalente, con predominio de los socios no fugados y una proporción cercana a la décima parte correspondiente a los fugados. Esta similitud sugiere que el tipo de actividad económica no permite distinguir claramente las clases por sí solo, aunque podría aportar información al interactuar con otras características financieras, demográficas o transaccionales.

Figura 20.

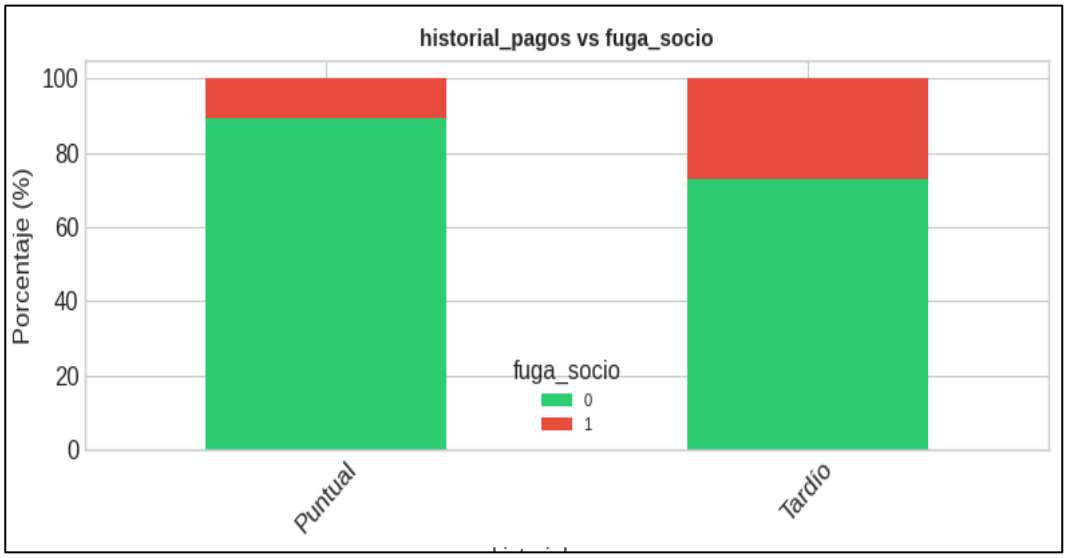
Distribución de fuga de socios por actividad económica



Nota. Distribución de la variable actividad económica.

La Figura 21 muestra la distribución porcentual de la condición de fuga según el historial de pagos de los socios. Se observa una diferencia apreciable entre las categorías analizadas: entre los socios con historial puntual, la fuga representa aproximadamente una décima parte de los registros; mientras que, entre aquellos con historial tardío, la proporción se aproxima a la cuarta parte. Esta diferencia evidencia una asociación descriptiva entre el historial de pagos y la condición de fuga, por lo que la variable podría contribuir a la capacidad clasificadora del modelo. No obstante, el resultado no debe interpretarse como evidencia de que la tardanza en los pagos cause la renuncia del socio.

Figura 21.
Distribución de fuga de socios por historial de pagos



Nota. Distribución de la variable historial de pagos.

- Hallazgo 1: Cinco de seis variables categóricas no discriminan la fuga

Las variables sexo, estado_civil, nivel_educativo, ocupacion y actividad_economica presentan tasas de churn prácticamente uniformes en todas sus categorías, oscilando entre el 11% y el 13%, sin desviaciones relevantes respecto al promedio global de 12.2%. Los gráficos de barras apiladas muestran proporciones visualmente idénticas (aproximadamente 88% verde / 12% rojo) en cada categoría, confirmando la ausencia de poder discriminatorio.

- Hallazgo 2: historial_pagos es la única variable categórica con poder discriminatorio

La variable historial_pagos presenta dos categorías: Puntual y Tardío. A diferencia de las cinco variables anteriores, aquí se observa una diferencia visible en los gráficos de barras apiladas: los socios con historial Tardío muestran una proporción de fuga notablemente mayor (franja roja más amplia) que los socios con historial Puntual, cuya franja roja es considerablemente menor.

Entrenamiento y Validación Cruzada de los Modelos

Tabla 16.

Tiempo de entrenamiento y métricas iniciales por modelo

Modelo	Accuracy	Precision	Recall	Tiempo (s)
Logistic Regression	0.8742	0.4910	0.8455	5.0
Random Forest	0.9210	0.6578	0.7347	425.1
Gradient Boosting	0.9232	0.6934	0.6637	775.3
XGBoost	0.9271	0.7210	0.6571	25.7
LightGBM	0.9272	0.7249	0.6504	53.0

Nota. Estos tiempos fueron obtenidos teniendo en cuenta los hiper parámetros declarados en la tabla 13.

Logistic Regression alcanza un recall elevado (0.8455), pero con una precisión muy baja (0.4910), lo que significa que detecta muchos churners reales, aunque genera una cantidad elevada de falsas alarmas. Los modelos basados en árboles presentan el patrón inverso: precisiones entre 0.6578 y 0.7249, y valores de recall entre 0.6504 y 0.7347. Esta diferencia es esperada: la regresión logística captura solo la estructura lineal de los datos, insuficiente para discriminar casos mixtos en un problema con interacciones no lineales como la fuga de socios.

4.5. Evaluación Comparativa de Modelos

La evaluación comparativa se realizó sobre el conjunto de test (11,087 registros, nunca visto durante el entrenamiento) con cinco perspectivas complementarias: tabla comparativa de métricas, gráfico de barras por métrica, curvas ROC y Precision-Recall, matrices de confusión, y box plots de estabilidad en validación cruzada. El criterio primario de selección fue el F1-Score sobre el conjunto de test, con ROC-AUC como desempate.

Tabla comparativa de métricas

La tabla consolida las métricas de desempeño de los cinco modelos sobre el conjunto de test. Los resultados de la validación cruzada de 5 folds se presentan posteriormente para el modelo seleccionado.

Tabla 17.

Desempeño comparativo de los modelos sobre el conjunto de test

Modelo	Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR - AUC
Random Forest	0.9210	0.6578	0.7347	0.6941	0.9076	0.7154
XGBoost	0.9271	0.7210	0.6571	0.6875	0.9036	0.7088
LightGBM	0.9272	0.7249	0.6504	0.6856	0.9049	0.7158
Gradient Boosting	0.9232	0.6934	0.6637	0.6782	0.9021	0.6892
Logistic Regression	0.8742	0.4910	0.8455	0.6212	0.9162	0.7365

Nota. Comparativa de los modelos según bajo indicadores.

Los resultados presentados en la Tabla 17 muestran que la regresión logística obtuvo el mayor valor de PR-AUC (0.7365), seguida de LightGBM (0.7158), Random Forest (0.7154), XGBoost (0.7088) y Gradient Boosting (0.6892). Esto indica que la regresión logística presentó el mejor desempeño global en la relación entre precisión y sensibilidad para la clase fuga. Sin embargo, también registró la menor precisión (0.4910) y el menor F1-score (0.6212), lo que evidencia una mayor cantidad de falsos positivos.

Random Forest obtuvo un PR-AUC de 0.7154 y alcanzó el mayor F1-score (0.6941), con una precisión de 0.6578 y una sensibilidad de 0.7347. Por esta razón, aunque no presentó el mayor PR-AUC de manera individual, fue seleccionado como modelo final al ofrecer el mejor equilibrio entre precisión y sensibilidad, aspecto relevante para identificar a los socios con riesgo de fuga sin incrementar excesivamente las clasificaciones positivas incorrectas.

En consecuencia, la selección de Random Forest no se sustentó exclusivamente en el PR-AUC, sino en la evaluación conjunta de las métricas, principalmente el F1-score, complementada con accuracy, precision, recall, ROC-AUC y PR-AUC.

Curvas ROC y Precision-Recall

Las curvas ROC y Precision-Recall evalúan el desempeño de los modelos a lo largo de todos los thresholds posibles, sin limitarse al valor por defecto de 0.5. Los cuatro modelos basados en árboles producen curvas ROC casi superpuestas, con

valores AUC entre 0.9021 y 0.9076, lo que indica una capacidad discriminante similar. En la curva Precision-Recall, Regresión Logística obtuvo el mayor PR-AUC (0.7365), seguida de LightGBM (0.7158) y Random Forest (0.7154). En consecuencia, estas curvas no sustentan que Random Forest domine la región de recall alto; su selección se fundamenta en el mayor F1-score, de acuerdo con el criterio principal definido.

4.6. Matrices de confusión

Las matrices de confusión sobre el conjunto de test (threshold = 0.50) permiten cuantificar los aciertos y errores en términos absolutos: verdaderos positivos (TP, churners detectados), falsos positivos (FP, falsas alarmas), verdaderos negativos (TN, socios activos correctamente clasificados) y falsos negativos (FN, churners no detectados). Para la cooperativa, los FN son el error más costoso: un socio que se fuga sin ser detectado representa la pérdida total de su valor futuro.

Tabla 18.

Matrices de confusión por modelo

Modelo	TP	FP	TN	FN	Churners detectados (%)	Falsas alarmas (%)
Random Forest	994	517	9,217	359	73.5%	5.3%
XGBoost	889	344	9,390	464	65.7%	3.5%
LightGBM	880	334	9,400	473	65.0%	3.4%
Gradient Boosting	898	397	9,337	455	66.4%	4.1%
Regresión logística	1,144	1,186	8,548	209	84.6%	12.2%

Nota. (total clase 0 = 9,734 | total clase 1 = 1,353)

Random Forest detecta 994 de los 1,353 socios que efectivamente se fugaron, equivalente al 73.5% del universo de riesgo, a cambio de 517 falsas alarmas sobre 9,734 socios activos (5.3% de tasa de falsas alarmas). En términos operativos, esto implica que por cada 100 socios señalados por el modelo como de alto riesgo, aproximadamente 66 son churners reales y 34 son falsas alarmas. Los modelos de boosting producen menos falsas alarmas, pero pierden entre 96 y 114 churners adicionales respecto a Random Forest.

Estabilidad en validación cruzada

Los box plots del F1-Score en los cinco folds muestran la estabilidad comparativa de los modelos basados en árboles. Random Forest presenta la caja más estrecha; adicionalmente, obtuvo la media de F1 más alta (0.6939), con una desviación estándar de 0.0016 puntos entre folds. Esta baja variabilidad sugiere estabilidad interna entre los pliegues evaluados, pero no demuestra por sí sola la ausencia de sobreajuste ni garantiza el desempeño futuro sobre nuevos datos. Logistic Regression obtuvo la media de F1 más baja (0.6311), lo que evidencia un desempeño inferior según el criterio principal dentro del experimento realizado.

4.7. Selección del Modelo Óptimo

Modelo seleccionado: Random Forest

En base al criterio de selección definido previamente (F1-Score como métrica primaria, ROC-AUC como desempate y Recall como indicador de cobertura), Random Forest resulta ser el modelo óptimo entre las cinco alternativas evaluadas. Sus métricas sobre el conjunto de test son: Accuracy = 0.9210, Precision = 0.6578, Recall = 0.7347, F1-Score = 0.6941, ROC-AUC = 0.9076 y PR-AUC = 0.7154. En validación cruzada obtiene F1 = 0.6939 ± 0.0016 y ROC-AUC = 0.9128 ± 0.0033, confirmando la consistencia de los resultados reportados sobre el conjunto de test.

Tabla 19.

Reporte de clasificación del modelo seleccionado Random Forest

Clase	Precision	Recall	F1-Score	Soporte
0 (No Churn)	0.96	0.95	0.95	9,734
1 (Churn)	0.66	0.73	0.69	1,353
Accuracy			0.92	11,087
Macro avg	0.81	0.84	0.82	11,087
Weighted avg	0.93	0.92	0.92	11,087

Nota. El modelo seleccionado es Random Forest - threshold = 0.50

4.8. Interpretabilidad mediante SHAP

Esta sección responde al tercer objetivo específico de la investigación, orientado a interpretar las predicciones del modelo de mejor desempeño mediante SHAP, este permite responder tres preguntas operativamente relevantes: qué variables son más influyentes en promedio para todo el modelo, cómo afecta cada

variable a cada socio individual, y si la relación entre una variable y la probabilidad de fuga es positiva o negativa. Se aplicó TreeExplainer (optimizado para modelos basados en árboles) sobre una muestra de 1,000 socios del conjunto de test, suficiente para obtener estimaciones estables de importancia global.

El análisis de interpretabilidad se aplicó sobre el mismo pipeline reportado en las secciones de evaluación comparativa y selección del modelo óptimo (Random Forest base, umbral de decisión 0.50), de modo que existe correspondencia directa entre las métricas presentadas, el modelo interpretado mediante SHAP y el modelo persistido para uso posterior.

Ranking de importancia global

La siguiente tabla presenta las 10 features con mayor valor absoluto medio de SHAP sobre la muestra analizada. Un valor SHAP absoluto alto significa que la variable desplaza sustancialmente la probabilidad de fuga predicha, independientemente del signo del desplazamiento.

Tabla 20.

Top 10 features según importancia

#	Feature	Mean |SHAP|	Categoría
1	ratio_retiro_aportes	0.172930	Derivada (descapitalización)
2	productos_activos	0.072599	Características bancarias
3	prestamos_otorgados	0.055430	Características bancarias
4	frecuencia_uso_mensual	0.054502	Comportamiento financiero
5	retiros_ultimos_3m	0.052443	Comportamiento financiero
6	riesgo_mora	0.040938	Derivada (riesgo crediticio)
7	nivel_educativo_Universitario	0.034175	Demográfica
8	nivel_educativo_Secundaria	0.029171	Demográfica
9	nivel_educativo_Técnico	0.028802	Demográfica
10	estado_civil_Soltero	0.024757	Demográfica

Nota. Importancia global obtenida mediante SHAP para el modelo Random Forest seleccionado.

Para los socios con fuga, la variable retiros_ultimos_3m se calculó considerando exclusivamente los retiros efectuados durante los tres meses anteriores a la fecha de renuncia voluntaria. Se excluyeron los movimientos

correspondientes a la liquidación, devolución o retiro final de aportes generado como consecuencia de la renuncia. Para los socios sin fuga, se consideraron los retiros efectuados durante los tres meses anteriores a la fecha de corte del estudio, establecida al 31 de diciembre de 2024.

ratio_retiro_aportes fue calculado como el monto de retiros de los últimos tres meses divididos entre los aportes acumulados disponibles hasta la fecha de referencia. La variable representa la intensidad de descapitalización reciente del socio.

El predictor más importante es ratio_retiro_aportes, variable derivada construida durante el feature engineering como el cociente entre retiros de los últimos 3 meses y el total acumulado de aportes del socio. Esta variable supera con amplitud al resto, confirmando que la proporción de retiros recientes respecto de los aportes acumulados relativa (cuánto está retirando el socio en proporción a lo que tiene acumulado) es una señal mucho más potente que cualquiera de las variables originales por separado. Ni retiros_ultimos_3m ni aportes_total, tomadas individualmente, alcanzan ese nivel de poder predictivo.

Este hallazgo respalda la utilidad metodológica del feature engineering basado en conocimiento del dominio cooperativo: las relaciones entre variables capturan dinámicas financieras que los valores absolutos no revelan. El mismo patrón se observa con riesgo_mora (posición 6), otra variable derivada que combina mora_dias con el número de préstamos otorgados al socio en su historial.

Las variables productos_activos, prestamos_otorgados y frecuencia_uso_mensual ocupan posiciones destacadas y son interpretables operativamente: un socio con pocos productos activos, con una relación crediticia limitada y con baja frecuencia de uso carece de anclas institucionales que incrementen su costo de cambio. Estos son los socios más vulnerables ante ofertas competitivas o insatisfacciones puntuales. Las variables demográficas (nivel educativo y estado civil) aparecen en las posiciones 7–10 pero con magnitud notablemente menor, confirmando el hallazgo del EDA: En conjunto, el ranking muestra que las variables derivadas, las características bancarias y los comportamientos financieros recientes tuvieron mayor contribución predictiva que las características demográficas y socioeconómicas.

En conjunto, el análisis realizado mediante SHAP permitió identificar e interpretar las variables con mayor influencia en la predicción del riesgo de fuga, cumpliendo de esta manera con el **objetivo específico 3** de la investigación. Se evidenció que la variable **ratio_retiro_aportes** presentó el mayor valor medio absoluto de SHAP, destacándose como la variable con mayor contribución predictiva dentro del modelo, seguida por productos_activos, prestamos_otorgados y frecuencia_uso_mensual.

Estos resultados corresponden al modelo Random Forest, seleccionado como el de mejor desempeño según el criterio establecido, con un F1-score de 0.6941 y un ROC-AUC de 0.9076. Estas métricas describen el desempeño predictivo del modelo. De manera complementaria, SHAP permite estimar la contribución de las variables a sus predicciones; por tanto, F1-score y ROC-AUC no validan por sí mismos la confiabilidad del análisis interpretativo.

4.9. Discusión de resultados

Los resultados obtenidos evidencian que los modelos evaluados presentaron diferentes niveles de desempeño para clasificar la fuga voluntaria de los socios de la Cooperativa de Ahorro y Crédito Santa María Magdalena. De acuerdo con el F1-score establecido como criterio principal de comparación, Random Forest obtuvo el mejor resultado, con un valor de 0,6941, una precisión de 0,6578 y un recall de 0,7347. Estas métricas muestran que el modelo alcanzó un equilibrio moderado entre la identificación de los socios fugados y la confiabilidad de las alertas generadas. En términos prácticos, logró detectar aproximadamente 73 de cada 100 socios que efectivamente presentaron fuga; asimismo, alrededor de 66 de cada 100 alertas positivas correspondieron a casos reales. Por consiguiente, Random Forest presentó el mejor balance comparativo entre precisión y sensibilidad dentro del experimento realizado, aunque todavía produjo falsas alertas y dejó sin identificar una parte de los casos de fuga.

La matriz de confusión de Random Forest registró 994 verdaderos positivos, 517 falsos positivos, 9217 verdaderos negativos y 359 falsos negativos. Los falsos positivos representan socios clasificados como posibles fugados que, en realidad, no abandonaron la cooperativa. Su inclusión en campañas de retención podría ocasionar costos innecesarios por comunicaciones, incentivos o atenciones personalizadas. Por su parte, los falsos negativos corresponden a socios que presentaron fuga, pero que no fueron identificados por el modelo, por lo que no

habrían recibido una intervención preventiva. Aunque la tasa de falsas alarmas fue de aproximadamente 5,3 % respecto de los socios no fugados, los 517 casos representan una cantidad operativamente relevante. Por ello, el desempeño del modelo debe analizarse considerando no solo los porcentajes obtenidos, sino también las consecuencias institucionales de cada tipo de error.

La elección entre priorizar la reducción de falsos positivos o falsos negativos depende de los objetivos y recursos de la cooperativa. Si el costo de perder a un socio es mayor que el costo de contactarlo preventivamente puede resultar conveniente privilegiar el recall y aceptar una mayor cantidad de falsas alertas. En cambio, si las campañas de retención requieren incentivos económicos elevados o recursos humanos limitados, será necesario incrementar la precisión para concentrar las intervenciones en los socios con mayor riesgo estimado. En consecuencia, el valor de las métricas no determina por sí solo la utilidad operativa del modelo, sino que debe complementarse con un análisis de costos y con la definición institucional del error que resulte más perjudicial.

Al comparar Random Forest con XGBoost, se observó que ninguno de los algoritmos fue superior en todas las métricas. Random Forest alcanzó un F1-score de 0,6941 y un recall de 0,7347, mientras que XGBoost obtuvo un F1-score aproximado de 0,6875 y un recall de 0,6571. Sin embargo, XGBoost presentó una mayor precisión, con 0,7210 frente a 0,6578 de Random Forest, y una menor tasa de falsas alarmas, con aproximadamente 3,5 % frente a 5,3 %. En valores absolutos, Random Forest identificó correctamente a 994 socios fugados y dejó sin detectar a 359; por otro lado, XGBoost identificó a 889 y no detectó a 464. Esto significa que Random Forest detectó 105 casos adicionales de fuga, aunque produjo 173 falsos positivos más que XGBoost. Debido a que el F1-score fue definido como criterio principal, Random Forest fue seleccionado, pero su ventaja se limita al equilibrio alcanzado entre precisión y recall y no implica superioridad absoluta en todas las condiciones de uso.

El F1-score de 0,6941 obtenido en esta investigación es inferior al 0,9143 reportado por Arotinco y Gutiérrez (2025), quienes aplicaron Random Forest para predecir la deserción de clientes de aseguradoras peruanas de pólizas vehiculares. En su validación con datos históricos, dichos autores alcanzaron una precisión de 88,89 %, un recall de 94,12 %, un F1-score de 91,43 % y un AUC de 92,71 %. Posteriormente, durante la validación en condiciones operativas, informaron una

exactitud de 97,08 %, un recall de 97,76 % y un AUC de 97,34 %. Si bien ambos estudios utilizan Random Forest y estudian un fenómeno de abandono, los valores obtenidos no son directamente equiparables debido a que se desarrollaron en dominios, poblaciones y condiciones experimentales diferentes.

La diferencia de 22,02 puntos porcentuales entre ambos F1-scores no demuestra necesariamente una deficiencia del modelo desarrollado en la presente investigación. Arotinco y Gutiérrez (2025) analizaron clientes de seguros vehiculares, cuya permanencia puede estar asociada con variables propias de las pólizas, como renovaciones, primas, siniestros, vencimientos, cobertura y características del vehículo. En cambio, el presente estudio examinó socios de una cooperativa de ahorro y crédito mediante variables demográficas, financieras y transaccionales, como aportes, retiros, frecuencia de uso, cantidad de productos y movimientos recientes. La definición operativa de la deserción, el nivel de desbalance, la calidad y cantidad de los datos, los periodos analizados, la selección de variables y los procedimientos de validación también pueden producir diferencias importantes en el desempeño.

Asimismo, las métricas dependen de la separabilidad existente entre las clases. Un conjunto de datos en el que los clientes que desertan presentan patrones claramente distintos puede alcanzar un F1-score elevado, mientras que otro conjunto con mayor solapamiento entre socios fugados y no fugados representa un problema de clasificación más complejo. En la base analizada en esta investigación se observaron distribuciones superpuestas en distintas características y una contribución limitada de las variables demográficas, lo que restringe la información disponible para separar ambas clases. Por consiguiente, exigir que el modelo alcance exactamente el 91,43 % reportado por otro estudio implicaría asumir que ambos conjuntos de datos poseen la misma estructura, calidad, variables, prevalencia y dificultad de clasificación, supuesto que no se encuentra metodológicamente sustentado.

En ese sentido, no resulta pertinente modificar o repetir el experimento con el propósito específico de igualar una métrica externa. El objetivo del entrenamiento y ajuste de un modelo predictivo debe ser mejorar su capacidad de generalización sobre datos no utilizados durante el aprendizaje, no alcanzar un valor previamente reportado en otro contexto. Orientar sucesivos ajustes hasta aproximarse al F1-score del antecedente podría aumentar el riesgo de sobreajuste, especialmente si las

decisiones metodológicas se realizan en función del conjunto de prueba. La comparación con Arotinco y Gutiérrez (2025) debe utilizarse como referencia para analizar las diferencias entre contextos y procedimientos, pero no como un umbral que el presente modelo deba reproducir obligatoriamente. La métrica obtenida debe valorarse de acuerdo con la complejidad del dataset propio, el protocolo de evaluación aplicado y las necesidades operativas de la cooperativa.

A pesar de estas diferencias, los resultados coinciden parcialmente con Arotinco y Gutiérrez (2025) en cuanto al comportamiento competitivo de Random Forest para estudiar la deserción. En ambas investigaciones, este algoritmo permitió representar relaciones no lineales e interacciones entre las características disponibles. No obstante, mientras que el antecedente alcanzó métricas superiores en el sector asegurador, en la presente investigación Random Forest presentó un desempeño más moderado y acompañado de un número relevante de falsos positivos y falsos negativos. Esta diferencia confirma que el rendimiento de un algoritmo no depende exclusivamente de su arquitectura, sino también del dominio, la representatividad de los datos, las características predictoras, la definición de la variable objetivo y el procedimiento utilizado para entrenarlo y evaluarlo.

Los resultados también presentan coincidencias con Bernachea et al. (2021) y Gallardo (2021), quienes aplicaron técnicas de machine learning para predecir la fuga en telecomunicaciones y servicios SaaS, respectivamente. En estos antecedentes, los algoritmos basados en árboles mostraron capacidad para representar interacciones complejas y relaciones no lineales. En la presente investigación, Random Forest, XGBoost, LightGBM y Gradient Boosting alcanzaron resultados relativamente cercanos, lo que evidencia que no existió una ventaja amplia de un único algoritmo. La diferencia entre Random Forest y XGBoost en el F1-score fue de aproximadamente 0,0066, por lo que la selección del primero responde al criterio establecido, pero debe interpretarse con cautela. Además, la precisión de 0,6578 demuestra que todavía existe margen para reducir las falsas alertas.

La aplicación de SHAP permitió analizar la contribución de las variables a las estimaciones generadas por el modelo seleccionado. Los resultados mostraron que las variables derivadas, las características bancarias y los comportamientos financieros ocuparon las primeras posiciones del ranking global, mientras que las características demográficas y socioeconómicas presentaron contribuciones

comparativamente menores. El principal predictor fue ratio_retiro_aportes, variable derivada que alcanzó un valor medio absoluto SHAP de 0.1729. A continuación, se ubicaron productos_activos (0.0726) y prestamos_otorgados (0.0554), pertenecientes a las características bancarias, seguidos por frecuencia_uso_mensual (0.0545) y retiros_ultimos_3m (0.0524), correspondientes al comportamiento financiero. Este resultado guarda relación con Khalilbayli y Sertkaya (2025), quienes sostienen que los valores de Shapley facilitan la comprensión de las decisiones generadas por modelos complejos y contribuyen a su transparencia. No obstante, la importancia obtenida mediante SHAP debe interpretarse en términos predictivos y no como evidencia causal. Una contribución elevada indica que la variable influyó de manera relevante en las estimaciones del modelo dentro del conjunto de datos analizado, pero no demuestra que sea responsable directa de la fuga del socio.

En relación con lo anterior, el ratio entre retiros y aportes debe entenderse como una asociación predictiva entre el comportamiento financiero del socio y la clasificación realizada por el modelo. No corresponde afirmar que un incremento en esta relación provoque necesariamente el abandono, porque pueden intervenir factores no incorporados en la base de datos. Del mismo modo, una menor frecuencia de uso o una cantidad reducida de productos activos pueden funcionar como señales para priorizar la revisión de determinados socios, pero no constituyen por sí solas pruebas de una decisión futura de renuncia. Por ello, la información obtenida mediante SHAP debe complementar el análisis institucional y no reemplazar la evaluación individual de los casos.

La relevancia de las variables de comportamiento presenta semejanzas con el trabajo de Segura (2022), quien implementó un modelo de fuga en el sector de factoring y complementó la predicción con simulaciones de estrategias de retención. En ambos estudios, las interacciones recientes del usuario con los servicios aportaron información para la clasificación. Sin embargo, el antecedente evaluó también el posible impacto financiero de las medidas de fidelización, mientras que la presente investigación se concentró en comparar los modelos, identificar a los socios con riesgo estimado e interpretar las variables que contribuyeron a las predicciones. Por tanto, los resultados obtenidos permiten priorizar socios para su evaluación, pero no demuestran todavía que una acción específica de retención consiga disminuir la fuga.

La importancia del ratio entre retiros y aportes y de la intensidad de movimiento también amplía lo reportado por Márquez (2020), quien destacó la utilidad de construir indicadores a partir de registros operativos para anticipar la fuga en telecomunicaciones. Las variables derivadas sintetizan relaciones que no se encuentran expresadas directamente en los atributos originales y pueden facilitar la identificación de cambios en la vinculación financiera. No obstante, su contribución depende de la forma en que fueron calculadas, del periodo comprendido y de que toda la información utilizada estuviera disponible antes de la fuga. En consecuencia, estas características deberán evaluarse en periodos posteriores para determinar si conservan su importancia y estabilidad.

Por otra parte, las variables demográficas mostraron baja capacidad discriminativa individual y correlaciones reducidas o nulas con la condición de fuga. Este resultado no significa que dichas características sean irrelevantes en cualquier contexto. Una correlación cercana a cero indica ausencia de una relación lineal apreciable, pero no descarta relaciones no lineales o interacciones con otras variables. Random Forest puede utilizar puntos de corte y combinaciones que no quedan reflejados en una correlación simple. Además, la limitada contribución de estas características podría deberse a la homogeneidad de la población, a la distribución de sus categorías o a la insuficiente representación de algunos grupos. Por consiguiente, los resultados respaldan una mayor importancia relativa de las características bancarias, los comportamientos financieros y las variables derivadas dentro del conjunto analizado, pero no justifican eliminar automáticamente las características demográficas de futuras evaluaciones.

Desde el punto de vista metodológico, el desbalance de las clases fue tratado mediante SMOTE dentro de cada pliegue de entrenamiento de la validación cruzada estratificada, de acuerdo con las recomendaciones de Elreedy y Atiya (2019). Este procedimiento evitó que los registros sintéticos fueran incorporados a los pliegues de validación o al conjunto de prueba y redujo el riesgo de transferencia de información. Sin embargo, SMOTE se aplicó después de la codificación one-hot, por lo que la interpolación pudo producir valores fraccionarios en variables originalmente binarias. Esta condición constituye una limitación del pipeline y deberá contrastarse en futuras investigaciones con alternativas como SMOTENC, ponderación de clases o diferentes técnicas de remuestreo.

El procedimiento empleado permitió comparar los algoritmos bajo condiciones homogéneas, pero las métricas obtenidas no deben considerarse una garantía de desempeño futuro. El uso de una partición aleatoria estratificada conserva la proporción de las clases, aunque no reproduce necesariamente los cambios temporales que pueden presentarse en un entorno operativo. Por ello, una mejora metodológica consiste en efectuar una validación temporal fuera de muestra, entrenando el modelo con periodos anteriores y evaluándolo con registros posteriores. Este procedimiento permitiría establecer con mayor rigor si las relaciones aprendidas se mantienen frente a variaciones en el comportamiento de los socios y en las condiciones institucionales.

Otra limitación fue la aplicación del umbral predeterminado de 0,50 y la ausencia de calibración probabilística. En estas condiciones, las salidas del modelo deben interpretarse como puntajes de priorización relativa y no como probabilidades calibradas de que un socio vaya a presentar fuga. Un umbral distinto podría incrementar el recall o la precisión, pero su selección deberá realizarse en un conjunto de validación independiente y mediante criterios vinculados con los costos institucionales. No sería metodológicamente adecuado ajustar el umbral utilizando el conjunto de prueba utilizado para la evaluación comparativa y selección del modelo, porque ello produciría estimaciones optimistas de desempeño.

También debe considerarse que la búsqueda de hiperparámetros no incluyó los parámetros asociados con el sobremuestreo y que no se incorporaron variables de tendencia construidas mediante diferentes ventanas temporales. La comparación de la actividad de los socios en periodos de tres, seis y doce meses, así como el cálculo de variaciones mensuales o pendientes, podría representar con mayor precisión la disminución progresiva del vínculo con la cooperativa. Sin embargo, la incorporación de nuevas características debe efectuarse únicamente con información disponible antes del evento de fuga y evaluarse mediante un protocolo que evite filtraciones de información.

Las limitaciones del conjunto de datos también influyen en la diferencia respecto de otros antecedentes. La base correspondió a registros históricos de una sola cooperativa y no incorporó variables externas como satisfacción del socio, calidad percibida del servicio, reclamos, ofertas de la competencia, condiciones económicas regionales o indicadores macroeconómicos. Estos factores podrían explicar parte del abandono que las variables internas no consiguen representar.

Además, el comportamiento observado en la Cooperativa de Ahorro y Crédito Santa María Magdalena puede diferir del correspondiente a aseguradoras, empresas de telecomunicaciones, servicios SaaS u otras cooperativas. Por esta razón, los resultados deben delimitarse a la población y al periodo estudiados.

En términos operativos, Random Forest puede emplearse inicialmente como una herramienta de priorización y no como un mecanismo autónomo para decidir la entrega de beneficios o incentivos. Su recall de 0,7347 indica que identificó aproximadamente a 73 de cada 100 socios que presentaron fuga; sin embargo, su precisión de 0,6578 significa que cerca de 34 de cada 100 alertas fueron incorrectas. Las predicciones deberían complementarse con una revisión institucional que considere el historial del socio, su valor para la cooperativa, el costo de la intervención y la posibilidad de contacto. Asimismo, antes de su implementación será necesario evaluar si las acciones de retención aplicadas a los casos priorizados producen una reducción real de la fuga.

En síntesis, Random Forest obtuvo el mayor F1-score dentro de los modelos evaluados, con un valor de 0,6941, además del recall más elevado entre los modelos basados en árboles, con 0,7347. No obstante, su precisión de 0,6578, los 517 falsos positivos y los 359 falsos negativos evidencian que la clasificación estuvo acompañada de errores que deben considerarse antes de una aplicación operativa. El F1-score inferior al 91,43 % reportado por Arotinco y Gutiérrez (2025) puede explicarse por las diferencias en el sector, la definición de deserción, la población, las variables disponibles, el nivel de separación entre las clases y el protocolo experimental. Por ello, el resultado del antecedente constituye una referencia comparativa y no una métrica que deba ser igualada de manera obligatoria. La finalidad metodológica debe ser obtener estimaciones generalizables y reproducibles en los datos propios de la cooperativa, evitando ajustes orientados únicamente a alcanzar un valor externo. Finalmente, SHAP permitió identificar las características con mayor contribución a las predicciones, pero sin establecer relaciones causales; en consecuencia, la implementación del modelo requerirá validación temporal, calibración, evaluación de costos y seguimiento de su desempeño.

Capítulo V

Conclusiones y recomendaciones

5.1. Conclusiones

- En relación con el objetivo general, se logró desarrollar y comparar modelos de machine learning para la predicción del riesgo de fuga de socios, utilizando una base de 55,432 registros elegibles, conformados por 48,669 socios no fugados y 6,763 socios con renuncia voluntaria formal. El modelo con mayor F1-score fue Random Forest, el cual alcanzó una accuracy de 0.9210, precision de 0.6578, recall de 0.7347, F1-score de 0.6941, ROC-AUC de 0.9076 y PR-AUC de 0.7154. En la validación cruzada, el modelo obtuvo un F1-score promedio de 0.6939 ± 0.0016 y un ROC-AUC de 0.9128 ± 0.0033. A nivel operativo, el modelo identificó correctamente 994 de 1,353 socios fugados (73.5%), aunque generó 517 falsas alarmas sobre 9,734 socios activos.
- En relación con el objetivo específico 1, se determinó que, entre las características bancarias evaluadas, productos_activos y prestamos_otorgados presentaron la mayor contribución a las predicciones del modelo, ocupando respectivamente las posiciones 2 y 3 del ranking global SHAP, con valores medios absolutos de 0.0726 y 0.0554. Por su parte, la antigüedad presentó una baja asociación lineal individual con la condición de fuga ($|r| = 0.0031$). Estos resultados indican que las características relacionadas con el nivel de vinculación mediante productos y créditos aportaron mayor información al modelo que la antigüedad considerada individualmente.
- En relación con el objetivo específico 2, los comportamientos financieros con mayor contribución a la clasificación fueron el ratio entre retiros y aportes, la frecuencia de uso mensual, los retiros realizados durante los últimos tres meses y el riesgo de mora. El ratio_retiro_aportes presentó la mayor importancia global del modelo, con un valor medio absoluto SHAP de 0.1729, mientras que frecuencia_uso_mensual, retiros_ultimos_3m y riesgo_mora alcanzaron valores de 0.0545, 0.0524 y 0.0409, respectivamente. Estos resultados evidencian una mayor contribución de los comportamientos financieros y de los indicadores derivados dentro del conjunto de variables analizado, sin que ello implique una relación causal con la fuga.

- En relación con el objetivo específico 3, el análisis mediante SHAP permitió identificar que las variables con mayor contribución global a las predicciones de Random Forest fueron ratio_retiro_aportes (0.1729), productos_activos (0.0726), prestamos_otorgados (0.0554), frecuencia_uso_mensual (0.0545) y retiros_ultimos_3m (0.0524). Asimismo, riesgo_mora ocupó la sexta posición con 0.0409. Los resultados muestran que las variables derivadas, las características bancarias y los comportamientos financieros tuvieron mayor presencia entre los predictores principales que las características demográficas, permitiendo interpretar las estimaciones generadas por el modelo sin atribuir relaciones causales.

5.2. Recomendaciones

- Se recomienda a la Cooperativa de Ahorro y Crédito Santa María Magdalena implementar progresivamente el modelo predictivo Random Forest como herramienta de apoyo para la identificación de socios con riesgo de renuncia voluntaria, considerando su desempeño en las métricas de clasificación evaluadas. Para ello, se sugiere realizar una fase piloto que permita ajustar el umbral de decisión según la capacidad operativa de la cooperativa, considerando tanto la identificación correcta de socios fugados como la generación de falsas alarmas. Asimismo, se propone evaluar periódicamente su desempeño con nuevos registros, a fin de verificar la estabilidad de su capacidad predictiva antes de su incorporación definitiva a los procesos institucionales.
- Se recomienda a la Cooperativa de Ahorro y Crédito Santa María Magdalena fortalecer el seguimiento de la vinculación financiera de sus socios, priorizando el monitoreo del número de productos activos y préstamos otorgados, debido a su contribución relevante en las predicciones del modelo. Esta información puede utilizarse para identificar perfiles con diferentes niveles de vinculación y orientar estrategias de atención y fidelización, considerando que la importancia predictiva de estas variables no implica necesariamente que el incremento de productos o créditos reduzca directamente la probabilidad de fuga.
- Se recomienda a la Cooperativa de Ahorro y Crédito Santa María Magdalena establecer mecanismos de monitoreo preventivo del comportamiento financiero de sus socios, considerando indicadores como la proporción entre retiros y aportes, la frecuencia de uso mensual, los retiros realizados durante los últimos tres meses y el riesgo de mora. A partir de estos indicadores, se sugiere evaluar

la implementación de alertas tempranas que permitan identificar patrones asociados con las predicciones de riesgo de fuga y orientar oportunamente las acciones de seguimiento y retención, sin asumir que dichos comportamientos constituyen causas directas de la desvinculación.

- Se recomienda a la Cooperativa de Ahorro y Crédito Santa María Magdalena incorporar herramientas de interpretabilidad como SHAP en el análisis de las predicciones generadas por el modelo Random Forest, priorizando la revisión de las variables con mayor contribución, como la proporción entre retiros y aportes, el número de productos activos, los préstamos otorgados, la frecuencia de uso mensual y los retiros recientes. Su aplicación permitiría comprender y evaluar la contribución de estas variables a las estimaciones del modelo, facilitando la interpretación de los resultados y su utilización como información de apoyo para la toma de decisiones institucionales, sin atribuir relaciones causales a los predictores identificados.

Referencias bibliográficas

Alvarez, J. (2022). *Perfil de la fuga de clientes nuevos con servicio de pago haberes de un banco peruano utilizando árboles de decisión*. [Tesis de pregrado, Universidad Nacional Agraría La Molina] Repositorio Lamolina https://repositorio.lamolina.edu.pe/items/130d91af-e7ce-4ee1-8a77-e35c33a3cc9c

Arotinco, J. & Gutierrez, W. (2025). *Modelo de predicción de deserción de clientes en el sector de aseguradoras peruanas de pólizas vehiculares mediante el uso del algoritmo Random Forest*. [Tesis de pregrado, Universidad Peruana de Ciencias Aplicadas] Repositorio UPC https://repositorioacademico.upc.edu.pe/handle/10757/684271

Barreno, J.; Alfaro, G.; Saltos, P. & Striseo, D. (2024). Análisis del uso de blockchain en auditoría financiera: Impacto en la transparencia, seguridad y eficiencia de los procesos contables. Un estudio en educación superior. *Ciencia aplicada em procesos educativos*, 4(5) 1-19. https://www.revistasocialfronteriza.com/ojs/index.php/rev/article/view/500

Bernachea, C. et al. (2021). *Técnica de Machine Learning para el cálculo de la probabilidad de fuga de los clientes de la empresa Bitel*. [Trabajo de suficiencia profesional, Universidad Esan] Repositorio Esan: https://repositorio.esan.edu.pe/items/e863066f-20b5-49ad-bcca-6a8e55f6b57d

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). *Machine learning explainability for external stakeholders* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2007.05408

Biecek, P., & Burzykowski, T. (2021). Explanatory model analysis: Explore, explain, and examine predictive models. *Chapman & Hall/CRC*. https://ema.drwhy.ai/

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

Brownlee, J. (2020). *Machine learning mastery with Python*. Machine Learning Mastery. https://machinelearningmastery.com/machine-learning-with-python/

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD*. https://arxiv.org/abs/1603.02754

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, 6. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems, 33*, 17212–17223. https://proceedings.neurips.cc/paper/2020/hash/c7bf0b7c1a86d5eb3be2c722cf2cf746-Abstract.html

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://jair.org/index.php/jair/article/view/11192

Forero, W. & Negre, F. (2024). Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática. *RIED Revista Iberoamericana de Educación a Distancia*, *27*(1) 1-34. https://www.redalyc.org/journal/3314/331475280017/331475280017.pdf

Gallardo, C. (2021). *Modelo de predicción de fuga de clientes en empresa SaaS de inteligencia logística*. [Tesis de pregrado, Universidad de Chile] Repositorio uchile https://repositorio.uchile.cl/handle/2250/183445

Galvez, F. & Uribe, J. (2024). *Aumento de la tasa de retención de clientes Visa en una entidad financiera de Lima, utilizando la metodología PDCA*. [Tesis de pregrado, Universidad Peruana de Ciencias Aplicadas] Repositorio UPC https://repositorioacademico.upc.edu.pe/handle/10757/675232

García, J. (2021). *Comparativa de técnicas de balanceo de datos Aplicación a un caso real para la predicción de fuga de clientes*. [Tesis de maestria, Universidad de Oviedo] Repositorio Institucional de la Universidad de Oviedo https://digibuo.uniovi.es/dspace/handle/10651/60629

García, O.; Zorrilla, A.; Briseño, A. & Arango, E. (2021). Actitud financiera, comportamiento financiero y conocimiento financiero en méxico. *Revista Cuadernos de Economía*, 40(83) 431- 457. http://www.scielo.org.co/scielo.php?pid=S0121-47722021000200431&script=sci_arttext

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media. https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

Goyo, J. (2024). *Implementación de un modelo de segmentación de socios para la identificación de capacidades de pago y riesgo de morosidad de los socios de la Cooperativa de Ahorro y Crédito Jardín Azuayo*. [Tesis de pregrado, Pontificia Unviersidad Católica del Ecuador] Repositorio PUCE https://repositorio.puce.edu.ec/items/83a44bdf-d805-4070-8845-512d10589fcb

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning* (2nd ed.). Springer. https://hastie.su.domains/ElemStatLearn/

Hermitaño, J. (2022). Aplicación de Machine Learning en la gestión de riesgo de crédito financiero: Una revisión sistemática. *Revista Interfases*, *1*(15) 160-178. https://dialnet.unirioja.es/servlet/articulo?codigo=9039554

Hossin, M., & Sulaiman, M. N. (2020). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*.

https://www.researchgate.net/publication/275224157_A_Review_on_Evaluation_Metrics_for_Data_Classification_Evaluations

International Journal of Creative Research Thoughts. (2021). Review on deep learning and machine learning. *IJCRT, 9*(5). https://www.ijcrt.org/papers/IJCRT2105026.pdf

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data—Recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 245–251). IEEE. https://doi.org/10.1109/ACII.2013.47

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khalilbayli, F., & Sertkaya, C. (2025). Enhancing Customer Churn Prediction in the Finance Sector through Explainable AI and Machine Learning . *International Journal of Advanced Natural Sciences and Engineering Researches, 9*(10). https://as-proceeding.com/index.php/ijanser/article/view/2827

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica, 31*(3), 249–268. https://www.informatica.si/index.php/informatica/article/view/148

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2020). Consistent individualized feature attribution for tree ensembles. *Nature Machine Intelligence, 2*, 56–67. https://www.nature.com/articles/s42256-019-0138-9

Marquez, F. (2020). *Identificación de clientes que realizaron fuga de equipos móviles en una empresa de telecomunicaciones utilizando el algoritmo Random Forest*. [Tesis de pregrado, Universidad Nacional Agraria La Molina] Repositorio Lamolina https://repositorio.lamolina.edu.pe/items/c4754761-cf29-4ce2-878b-efe76829a718

Molnar, C. (2022). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

Moreno, W. (2025). *Clusterización y aplicación de modelos predictivos para la identificación de factores que conlleven a una desafiliación de una cooperativa financiera*. [Tesis de pregrado, Universidad Nacional Abierta y a Distancia] Repositorio UNAD https://repository.unad.edu.co/handle/10596/67241

Osman, Y. (2021). *Customer churn prediction using machine learning* [Tesis de maestría, KTH Royal Institute of Technology]. https://www.diva-portal.org/smash/get/diva2:1574424/FULLTEXT01.pdf

Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. https://arxiv.org/abs/2010.16061

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *Measurement: Sensors, 25*, 100621. https://www.sciencedirect.com/science/article/pii/S2666603023000143

Probst, P., Wright, M. N., & Boulesteix, A. L. (2020). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery, 9*(3). https://onlinelibrary.wiley.com/doi/10.1002/widm.1301

Quintero, L. (2023). *Aplicación de Machine Learning a un modelo tradicional de Prevención y detección de fraude en entidad financiera proyectado a periodos trimestrales*. [Tesis de pregrado, Universidad de La Salle] Repositorio lasalle https://ciencia.lasalle.edu.co/items/c469c878-0f16-4731-ac3e-ecb55c9d8df9

Raschka, S., & Mirjalili, V. (2020). *Python machine learning* (3rd ed.). Packt Publishing. https://github.com/rasbt/python-machine-learning-book-3rd-edition

Reyes, E. (2022). Metodología de la investigación científica (1er ed.). Page Publishing. ISBN 978-1-64334-601-4

Saito, T., & Rehmsmeier, M. (2020). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*, 160. https://link.springer.com/article/10.1007/s42979-021-00592-x

Segura, J. (2022). *Desarrollo de un modelo de predicción de fuga de clientes y diseño de experimento para la aplicación de estrategias de fidelización en factoring*. [Tesis de pregrado, Universidad de Chile] Repositorio uchile https://repositorio.uchile.cl/handle/2250/186972

Shalev-Shwartz, S., & Ben-David, S. (2021). *Understanding machine learning: From theory to algorithms*. Cambridge University Press. https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/

Sharifani, K. & Amini, M. (2023). Machine Learning and Deep Learning: A Review of Methods and Applications. *World Information Technology and Engineering Journal*, *10*(7) 3897-3904. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723

Siblini, W., Fréry, J., He-Guelton, L., Oblé, F., & Wang, Y.-Q. (2020). Master your metrics with calibration. In M. R. Berthold, A. Feelders, & G. Krempl (Eds.), *Advances in intelligent data analysis XVIII* (Lecture Notes in Computer Science, Vol. 12080, pp. 457–469). Springer. https://doi.org/10.1007/978-3-030-44584-3_36

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375830

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2020). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*. https://www.researchgate.net/publication/329722807_The_area_under_the_precision-recall_curve_as_a_performance_metric_for_rare_binary_events

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192.

https://www.emerald.com/aci/article/17/1/168/6011/Classification-assessment-methods

Vásquez, K. & Ramirez, S. (2025). Estrategias de captación y retención del cliente en el sector del consumo masivo (retail) dentro del Distrito Metropolitano de Quito. *MQRInvestigar Revista Multidisciplinaria de Investigación Científica*, 9(1) 1-25. https://mqrinvestigar.com/2025/index.php/mqr/article/view/307

Vicente, R. (2023). *Diseño y desarrollo de un modelo predictivo de fuga para un segmento de clientes en el mercado de foodservice en Agrosuper S.A. utilizando herramientas de machine learning*. [Tesis de pregrado, Universidad de Chile] Repositorio uchile https://repositorio.uchile.cl/handle/2250/193177

Vizcaíno, P.; Cedeño, R. & Maldonado, I. (2023). Metodología de la investigación científica: guía práctica. *Ciencia Latina Revista Científica Multidisciplinar, 7*(4) 9723-9762. https://www.ciencialatina.org/index.php/cienciala/article/view/7658

Vowels, L. M., Vowels, M. J., Carnelley, K. B., & Kumashiro, M. (2023). A machine learning approach to predicting perceived partner support from relational and individual variables. *Social Psychological and Personality Science, 14*(5). 617-629. https://doi.org/10.1177/19485506221114982

Zhou, Z. H., & Liu, X. Y. (2020). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEEXplore, 1*(18), 63-77. https://ieeexplore.ieee.org/document/1549828/citations?tabFilter=patents#citations

Anexos

Anexo 1. Matriz de consistencia

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES	METODO DE INVESTIGACIÓN
PROBLEMA GENERAL ¿Qué modelo de machine learning de clasificación supervisada, desarrollado a partir de características bancarias y comportamientos financieros de los socios, presenta el mejor desempeño para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025? **PROBLEMAS ESPECÍFICOS** a) ¿Qué características bancarias de los socios tienen mayor capacidad predictiva para predecir el riesgo de fuga en la	**OBJETIVO GENERAL** Desarrollar y comparar modelos de machine learning de clasificación supervisada, basados en características bancarias y comportamientos financieros de los socios, para identificar el modelo con mejor desempeño en la predicción del riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025. **OBJETIVOS ESPECÍFICOS** a) Determinar las características bancarias de los socios con mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y	**HIPÓTESIS GENERAL** En el presente estudio no se plantea una hipótesis debido a la naturaleza descriptiva-comparativa con propósito predictivo, esta se orienta a identificar y caracterizar los factores asociados a la fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena, empleando técnicas de aprendizaje automático (Machine Learning) para analizar grandes volúmenes de datos históricos.	**VARIABLE OBJETIVO** Y: Fuga de socios **VARIABLES PREDICTORAS** X1: Características bancarias del socio X2: Comportamiento financiero del socio X3: Características demográficas	**TIPO DE INVESTIGACIÓN** Aplicada **NIVEL DE INVESTIGACIÓN** Descriptivo-Comparativo con propósito predictivo **DISEÑO** No experimental **POBLACIÓN** Conformada por los 55,432 registros elegibles de socios de la CACSMM, clasificados en socios no fugados y socios con renuncia voluntaria formal. **MUESTRA** Se tomó el total de la población debido al acceso a la información.

Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025? b) ¿Qué comportamientos financieros de los socios tienen mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025? c) ¿Qué variables explican con mayor peso las predicciones del modelo de mejor desempeño, según el análisis de interpretabilidad SHAP, en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025?	Crédito Santa María Magdalena, 2025. b) Determinar los comportamientos financieros de los socios con mayor capacidad predictiva para predecir el riesgo de fuga en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025. c) Interpretar las predicciones del modelo de mejor desempeño mediante el análisis de importancia de variables SHAP, para identificar los factores con mayor influencia en la predicción del riesgo de fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena, 2025.	**TÉCNICA** Análisis documental y extracción automatizada de registros **INSTRUMENTO** Ficha de registro digital y ficha de procesamiento de variables derivadas	

Anexo 2. Matriz de operacionalización de variables

Tipo de variable	Variable	Definición operacional	Dimensión	Indicadores	Forma de medición	Escala	Instrumento	Fuente
Variable objetivo	Fuga de socios	Se midió según el registro de renuncia voluntaria del socio en BESTERP.	Condición de fuga	Fuga del socio	Registro de renuncia voluntaria: 0 = no fuga y 1 = fuga.	Nominal dicotómica	Ficha de registro digital de datos históricos	Sistema BESTERP
Variable predictora	Características demográficas y socioeconómicas de los socios	Se midieron mediante los datos demográficos y socioeconómicos registrados en BESTERP.	Perfil demográfico y socioeconómico	Edad	Años cumplidos por el socio en la fecha de corte.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Sexo	Sexo registrado: masculino o femenino.	Nominal dicotómica	Ficha de registro digital de datos históricos	Sistema BESTERP
				Estado civil	Categoría registrada: casado, conviviente, soltero, divorciado o viudo.	Nominal politómica	Ficha de registro digital de datos históricos	Sistema BESTERP
				Nivel educativo	Máximo nivel alcanzado: primaria, secundaria, técnico,	Ordinal	Ficha de registro digital de datos históricos	Sistema BESTERP

					universitario o posgrado.			
				Ocupación	Condición registrada: dependiente, desempleado, independiente o jubilado.	Nominal politómica	Ficha de registro digital de datos históricos	Sistema BESTERP
				Actividad económica	Clasificación de la actividad económica: formal o informal.	Nominal dicotómica	Ficha de registro digital de datos históricos	Sistema BESTERP
				Ingresos mensuales	Monto del ingreso mensual declarado, expresado en soles.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
	Características bancarias de los socios	Se midieron mediante los datos de vinculación y productos financieros registrados en BESTERP.	Vinculación institucional y productos financieros	Productos activos	Número de productos financieros activos del socio.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Préstamos otorgados	Número de préstamos otorgados al socio.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Antigüedad	Meses transcurridos	Razón	Ficha de registro digital	Sistema BESTERP

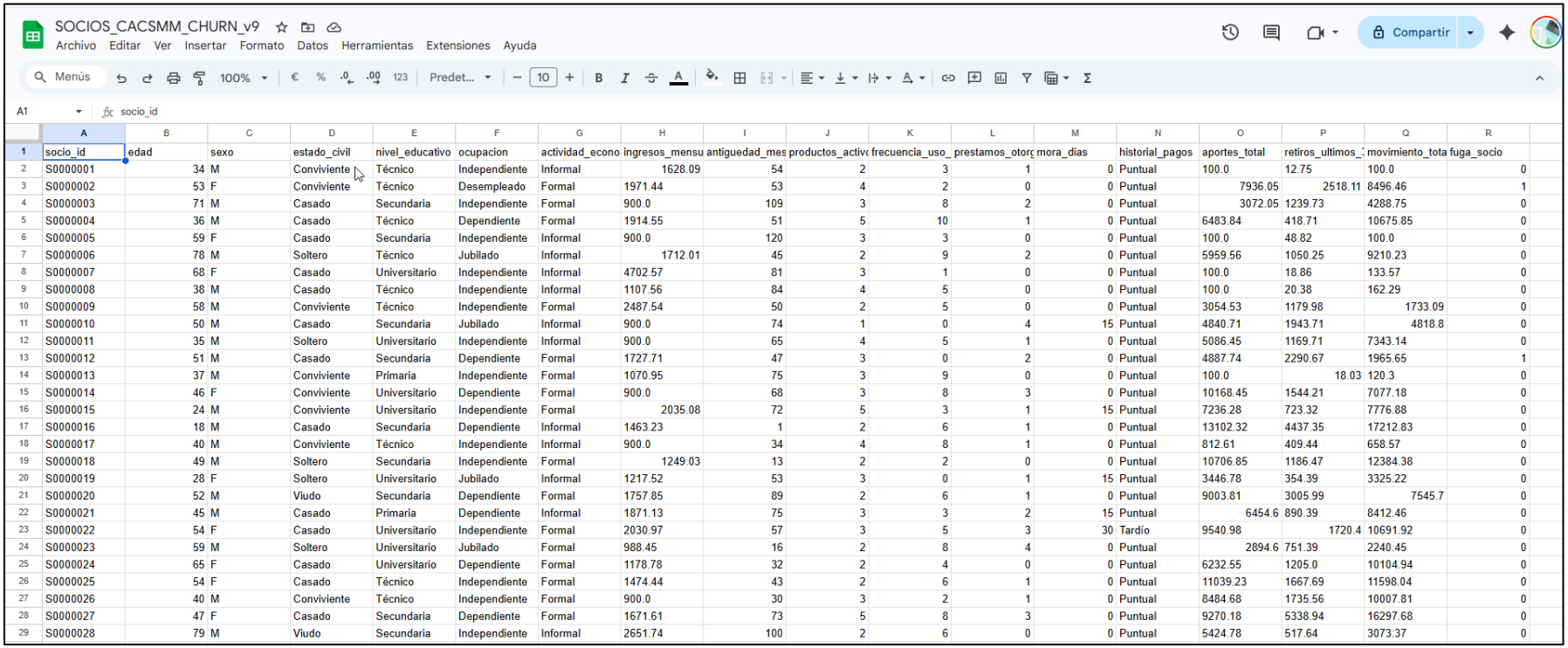
					desde la afiliación del socio.		de datos históricos	
	Comportamientos financieros de los socios	Se midieron mediante los datos transaccionales, de pago y de riesgo registrados en BESTERP.	Actividad transaccional y riesgo financiero	Historial de pagos	Clasificación crediticia al cierre mensual: Puntual = Normal; Tardío = distinta de Normal. Con varias obligaciones, se tomó la de mayor riesgo.	Nominal dicotómica	Ficha de registro digital de datos históricos	Sistema BESTERP
				Frecuencia de uso mensual	Número promedio de transacciones realizadas mensualmente.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Retiros de los últimos tres meses	Monto total retirado durante los últimos tres meses, expresado en soles.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Movimiento total de los últimos seis meses	Suma de depósitos y retiros de los últimos seis	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP

					meses, expresada en soles.			
				Días de mora	Número de días de mora acumulados en préstamos vigentes.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
				Aportes acumulados	Monto total de aportes acumulados por el socio, expresado en soles.	Razón	Ficha de registro digital de datos históricos	Sistema BESTERP
Variable predictora derivada	Comportamientos financieros de los socios	Se midieron mediante indicadores calculados a partir de las variables originales.	Variables derivadas mediante ingeniería de características	Ratio retiro/aportes	Retiros de los últimos tres meses divididos entre los aportes totales.	Razón	Ficha de procesamiento de variables derivadas	Dataset procesado en Python
				Intensidad de movimiento	Movimiento total de los últimos seis meses dividido entre la antigüedad en meses.	Razón	Ficha de procesamiento de variables derivadas	Dataset procesado en Python
				Uso por producto	Frecuencia de uso mensual dividida entre	Razón	Ficha de procesamiento	Dataset procesado en Python

					el número de productos activos.		de variables derivadas	
				Riesgo de mora	Días de mora x (el número de préstamos otorgados +1).	Razón	Ficha de procesamiento de variables derivadas	Dataset procesado en Python
				Ratio ingreso/movimiento	Ingreso mensual proyectado a seis meses dividido entre el movimiento total del mismo periodo.	Razón	Ficha de procesamiento de variables derivadas	Dataset procesado en Python

Nota. El desempeño de los modelos fue evaluado mediante Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC y matriz de confusión.

Anexo 3. Dataset



SOCIOS_CACSMM_CHURN_v9

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda

Compartir

A1 socio_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	socio_id	edad	sexo	estado_civil	nivel_educativo	ocupacion	actividad_econo	ingresos_mensu	antiguedad_mes	productos_activ	frecuencia_uso_	prestamos_otor	mora_dias	historial_pagos	aportes_total	retiros_ultimos_	movimiento_tota	fuga_socio
2	S0000001	34	M	Conviviente	Técnico	Independiente	Informal	1628.09	54	2	3	1	0	Puntual	100.0	12.75	100.0	0
3	S0000002	53	F	Conviviente	Técnico	Desempleado	Formal	1971.44	53	4	2	0	0	Puntual	7936.05	2518.11	8496.46	1
4	S0000003	71	M	Casado	Secundaria	Independiente	Formal	900.0	109	3	8	2	0	Puntual	3072.05	1239.73	4288.75	0
5	S0000004	36	M	Casado	Técnico	Dependiente	Formal	1914.55	51	5	10	1	0	Puntual	6483.84	418.71	10675.85	0
6	S0000005	59	F	Casado	Secundaria	Independiente	Informal	900.0	120	3	3	0	0	Puntual	100.0	48.82	100.0	0
7	S0000006	78	M	Soltero	Técnico	Jubilado	Informal	1712.01	45	2	9	2	0	Puntual	5959.56	1050.25	9210.23	0
8	S0000007	68	F	Casado	Universitario	Independiente	Informal	4702.57	81	3	1	0	0	Puntual	100.0	18.86	133.57	0
9	S0000008	38	M	Casado	Técnico	Independiente	Informal	1107.56	84	4	5	0	0	Puntual	100.0	20.38	162.29	0
10	S0000009	58	M	Conviviente	Técnico	Independiente	Formal	2487.54	50	2	5	0	0	Puntual	3054.53	1179.98	1733.09	0
11	S0000010	50	M	Casado	Secundaria	Jubilado	Informal	900.0	74	1	0	4	15	Puntual	4840.71	1943.71	4818.8	0
12	S0000011	35	M	Soltero	Universitario	Independiente	Informal	900.0	65	4	5	1	0	Puntual	5086.45	1169.71	7343.14	0
13	S0000012	51	M	Casado	Secundaria	Dependiente	Formal	1727.71	47	3	0	2	0	Puntual	4887.74	2290.67	1965.65	1
14	S0000013	37	M	Conviviente	Primaria	Independiente	Formal	1070.95	75	3	9	0	0	Puntual	100.0	18.03	120.3	0
15	S0000014	46	F	Conviviente	Universitario	Dependiente	Formal	900.0	68	3	8	3	0	Puntual	10168.45	1544.21	7077.18	0
16	S0000015	24	M	Conviviente	Universitario	Independiente	Formal	2035.08	72	5	3	1	15	Puntual	7236.28	723.32	7776.88	0
17	S0000016	18	M	Casado	Secundaria	Dependiente	Informal	1463.23	1	2	6	1	0	Puntual	13102.32	4437.35	17212.83	0
18	S0000017	40	M	Conviviente	Técnico	Independiente	Informal	900.0	34	4	8	1	0	Puntual	812.61	409.44	658.57	0
19	S0000018	49	M	Soltero	Secundaria	Independiente	Formal	1249.03	13	2	2	0	0	Puntual	10706.85	1186.47	12384.38	0
20	S0000019	28	F	Soltero	Universitario	Jubilado	Informal	1217.52	53	3	0	1	15	Puntual	3446.78	354.39	3325.22	0
21	S0000020	52	M	Viudo	Secundaria	Dependiente	Formal	1757.85	89	2	6	1	0	Puntual	9003.81	3005.99	7545.7	0
22	S0000021	45	M	Casado	Primaria	Dependiente	Informal	1871.13	75	3	3	2	15	Puntual	6454.6	890.39	8412.46	0
23	S0000022	54	F	Casado	Universitario	Independiente	Formal	2030.97	57	3	5	3	30	Tardío	9540.98	1720.4	10691.92	0
24	S0000023	59	M	Soltero	Universitario	Jubilado	Formal	988.45	16	2	8	4	0	Puntual	2894.6	751.39	2240.45	0
25	S0000024	65	F	Casado	Universitario	Dependiente	Formal	1178.78	32	2	4	0	0	Puntual	6232.55	1205.0	10104.94	0
26	S0000025	54	F	Casado	Técnico	Independiente	Formal	1474.44	43	2	6	1	0	Puntual	11039.23	1667.69	11598.04	0
27	S0000026	40	M	Conviviente	Técnico	Independiente	Formal	900.0	30	3	2	1	0	Puntual	8484.68	1735.56	10007.81	0
28	S0000027	47	F	Casado	Secundaria	Dependiente	Formal	1671.61	73	5	8	3	0	Puntual	9270.18	5338.94	16297.68	0
29	S0000028	79	M	Viudo	Secundaria	Independiente	Informal	2651.74	100	2	6	0	0	Puntual	5424.78	517.64	3073.37	0

Anexo 4. Código

```
# ============================================================
# 4.4b FEATURE ENGINEERING — VARIABLES DERIVADAS
# ============================================================

n_before = df_model.shape[1]

# --- Ratio de retiro sobre aportes (señal de descapitalización) ---
if 'retiros_ultimos_3m' in df_model.columns and 'aportes_total' in df_model.columns:
    df_model['ratio_retiro_aportes'] = (
        df_model['retiros_ultimos_3m'] / df_model['aportes_total'].replace(0, np.nan)
    ).fillna(0)
    print('  ✅ ratio_retiro_aportes = retiros_3m / aportes_total')

# --- Intensidad de movimiento (actividad relativa a antigüedad) ---
if 'movimiento_total_6m' in df_model.columns and 'antiguedad_meses' in df_model.columns:
    df_model['intensidad_movimiento'] = (
        df_model['movimiento_total_6m'] / df_model['antiguedad_meses'].replace(0, np.nan)
    ).fillna(0)
    print('  ✅ intensidad_movimiento = movimiento_6m / antigüedad')

# --- Uso relativo a productos (¿usa lo que tiene?) ---
if 'frecuencia_uso_mensual' in df_model.columns and 'productos_activos' in df_model.columns:
    df_model['uso_por_producto'] = (
        df_model['frecuencia_uso_mensual'] / df_model['productos_activos'].replace(0, np.nan)
    ).fillna(0)
    print('  ✅ uso_por_producto = frecuencia_uso / productos_activos')

# --- Riesgo crediticio compuesto ---
if 'mora_dias' in df_model.columns and 'prestamos_otorgados' in df_model.columns:
    df_model['riesgo_mora'] = df_model['mora_dias'] * (df_model['prestamos_otorgados'] + 1)
    print('  ✅ riesgo_mora = mora_dias × (prestamos + 1)')

# --- Ingreso relativo a movimiento (capacidad vs actividad) ---
if 'ingresos_mensuales' in df_model.columns and 'movimiento_total_6m' in df_model.columns:
    df_model['ratio_ingreso_movimiento'] = (
        df_model['ingresos_mensuales'] * 6 / df_model['movimiento_total_6m'].replace(0, np.nan)
    )
    print('  ✅ ratio_ingreso_movimiento = ingreso×6 / movimiento_6m')

n_after = df_model.shape[1]
print(f'\n📐 Features creadas: {n_after - n_before}')
print(f'📐 Shape actual: {df_model.shape}')
```

```
  ✅ ratio_retiro_aportes = retiros_3m / aportes_total
  ✅ intensidad_movimiento = movimiento_6m / antigüedad
  ✅ uso_por_producto = frecuencia_uso / productos_activos
  ✅ riesgo_mora = mora_dias × (prestamos + 1)
  ✅ ratio_ingreso_movimiento = ingreso×6 / movimiento_6m
```

```python
# ============================================================
# 4.5 CODIFICACIÓN DE VARIABLES CATEGÓRICAS
# ============================================================

cat_cols_final = df_model.select_dtypes(include=['object', 'category']).columns.tolist()
cat_cols_final = [c for c in cat_cols_final if c != 'fuga_socio']

if len(cat_cols_final) > 0:
    print('🏷️ Codificación:')
    for col in cat_cols_final:
        n_unique = df_model[col].nunique()
        if n_unique <= 10:
            # FIX: .astype(int) para evitar columnas bool que se pierden con select_dtypes
            dummies = pd.get_dummies(df_model[col], prefix=col, drop_first=True).astype(int)
            df_model = pd.concat([df_model, dummies], axis=1)
            df_model = df_model.drop(columns=[col])
            print(f'  ✅ {col} ({n_unique} cat): One-Hot → {dummies.shape[1]} columnas (dtype=int)')
        else:
            le = LabelEncoder()
            df_model[col] = le.fit_transform(df_model[col].astype(str))
            print(f'  ✅ {col} ({n_unique} cat): Label Encoding')

print(f'\n📐 Shape final: {df_model.shape}')
print(f'📋 Tipos de datos resultantes:')
print(f'   Numéricas (int/float): {len(df_model.select_dtypes(include=[np.number]).columns)}')
print(f'   Bool:                  {len(df_model.select_dtypes(include=["bool"]).columns)}')
print(f'   Object:                {len(df_model.select_dtypes(include=["object"]).columns)}')
```

```
🏷️ Codificación:
  ✅ sexo (2 cat): One-Hot → 1 columnas (dtype=int)
  ✅ estado_civil (5 cat): One-Hot → 4 columnas (dtype=int)
  ✅ nivel_educativo (5 cat): One-Hot → 4 columnas (dtype=int)
  ✅ ocupacion (4 cat): One-Hot → 3 columnas (dtype=int)
  ✅ actividad_economica (2 cat): One-Hot → 1 columnas (dtype=int)
  ✅ historial_pagos (2 cat): One-Hot → 1 columnas (dtype=int)

📐 Shape final: (55432, 30)
📋 Tipos de datos resultantes:
   Numéricas (int/float): 30
   Bool:                  0
   Object:                0
```

```python
# ============================================================
# 5.1 SEPARACIÓN FEATURES / TARGET
# ============================================================

# Codificar target si es categórico
if df_model[TARGET_COL].dtype == 'object':
    le_target = LabelEncoder()
    df_model[TARGET_COL] = le_target.fit_transform(df_model[TARGET_COL])
    print(f'🏈 Target codificado: {dict(zip(le_target.classes_, le_target.transform(le_target.classes_)))}')

X = df_model.drop(columns=[TARGET_COL])
y = df_model[TARGET_COL]

# FIX CRÍTICO: incluir TODAS las columnas numéricas y booleanas
# (pd.get_dummies puede generar bool en algunas versiones de pandas)
X = X.select_dtypes(include=[np.number, 'bool'])

# Convertir cualquier columna bool restante a int
bool_cols = X.select_dtypes(include=['bool']).columns
if len(bool_cols) > 0:
    X[bool_cols] = X[bool_cols].astype(int)
    print(f'⚠️ {len(bool_cols)} columnas bool convertidas a int')

# Rellenar cualquier nulo restante
X = X.fillna(0)

print(f'📊 Features (X): {X.shape}')
print(f'🎯 Target (y): {y.shape} — distribución: {dict(y.value_counts())}')

# Verificación: listar todas las features
print(f'\n📋 Features incluidas ({X.shape[1]}):')
for j, col in enumerate(X.columns, 1):
    print(f'   {j:2d}. {col} (dtype={X[col].dtype})')
```

```
📊 Features (X): (55432, 29)
🎯 Target (y): (55432,) — distribución: {0: np.int64(48669), 1: np.int64(6763)}

📋 Features incluidas (29):
    1. edad (dtype=int64)
    2. ingresos_mensuales (dtype=float64)
    3. antiguedad_meses (dtype=int64)
    4. productos_activos (dtype=int64)
    5. frecuencia_uso_mensual (dtype=int64)
    6. prestamos_otorgados (dtype=int64)
    7. mora_dias (dtype=int64)
    8. aportes_total (dtype=float64)
    9. retiros_ultimos_3m (dtype=float64)
   10. movimiento_total_6m (dtype=float64)
   11. ratio_retiro_aportes (dtype=float64)
   12. intensidad_movimiento (dtype=float64)
```

```python
# ============================================================
# 6.1 DEFINICIÓN DE MODELOS
# ============================================================
models = {
    'Logistic Regression': ImbPipeline([
        ('scaler', StandardScaler()),
        ('smote', SMOTE(random_state=RANDOM_STATE)),
        ('clf', LogisticRegression(max_iter=1000, random_state=RANDOM_STATE))
    ]),
    'Random Forest': ImbPipeline([
        ('smote', SMOTE(random_state=RANDOM_STATE)),
        ('clf', RandomForestClassifier(n_estimators=300, max_depth=15, min_samples_split=5,
                                       min_samples_leaf=2, random_state=RANDOM_STATE, n_jobs=-1))
    ]),
    'Gradient Boosting': ImbPipeline([
        ('smote', SMOTE(random_state=RANDOM_STATE)),
        ('clf', GradientBoostingClassifier(n_estimators=200, max_depth=5, learning_rate=0.1,
                                           subsample=0.8, random_state=RANDOM_STATE))
    ]),
    'XGBoost': ImbPipeline([
        ('smote', SMOTE(random_state=RANDOM_STATE)),
        ('clf', XGBClassifier(n_estimators=300, max_depth=6, learning_rate=0.1,
                              subsample=0.8, colsample_bytree=0.8,
                              use_label_encoder=False, eval_metric='logloss',
                              random_state=RANDOM_STATE, n_jobs=-1))
    ]),
    'LightGBM': ImbPipeline([
        ('smote', SMOTE(random_state=RANDOM_STATE)),
        ('clf', LGBMClassifier(n_estimators=300, max_depth=6, learning_rate=0.1,
                               subsample=0.8, colsample_bytree=0.8,
                               random_state=RANDOM_STATE, n_jobs=-1, verbose=-1))
    ]),
    #'SVM': ImbPipeline([
    #    ('scaler', StandardScaler()),
    #    ('smote', SMOTE(random_state=RANDOM_STATE)),
    #    ('clf', SVC(kernel='rbf', probability=True, random_state=RANDOM_STATE))
    #]),
}

print(f'✅ {len(models)} modelos definidos:')
for name in models:
    print(f'   • {name}')
```

```
✅ 5 modelos definidos:
   • Logistic Regression
   • Random Forest
   • Gradient Boosting
   • XGBoost
   • LightGBM
```

```python
# ============================================================
# 6.2 ENTRENAMIENTO Y VALIDACIÓN CRUZADA
# ============================================================

results = {}

print('=' * 70)
print('ENTRENAMIENTO DE MODELOS')
print('=' * 70)

for name, pipeline in models.items():
    print(f'\n🔄 Entrenando: {name}...')
    start_time = time()

    try:
        # Validación cruzada
        cv_f1 = cross_val_score(pipeline, X_train, y_train, cv=cv, scoring='f1', n_jobs=-1)
        cv_roc = cross_val_score(pipeline, X_train, y_train, cv=cv, scoring='roc_auc', n_jobs=-1)

        # Entrenar en todo el train set
        pipeline.fit(X_train, y_train)

        # Predicciones en test
        y_pred = pipeline.predict(X_test)
        y_proba = pipeline.predict_proba(X_test)[:, 1]

        elapsed = time() - start_time

        results[name] = {
            'y_pred': y_pred, 'y_proba': y_proba,
            'accuracy': accuracy_score(y_test, y_pred),
            'precision': precision_score(y_test, y_pred, zero_division=0),
            'recall': recall_score(y_test, y_pred, zero_division=0),
            'f1': f1_score(y_test, y_pred, zero_division=0),
            'roc_auc': roc_auc_score(y_test, y_proba),
            'pr_auc': average_precision_score(y_test, y_proba),
            'cv_f1_mean': cv_f1.mean(), 'cv_f1_std': cv_f1.std(),
            'cv_roc_mean': cv_roc.mean(), 'cv_roc_std': cv_roc.std(),
            'train_time': elapsed, 'pipeline': pipeline
        }

        #print(f'  ✅ F1={results[name]["f1"]:.4f} | ROC-AUC={results[name]["roc_auc"]:.4f} '
         #   f'| CV-F1={cv_f1.mean():.4f}±{cv_f1.std():.4f} | ⏱ {elapsed:.1f}s')
        print(f'  ✅ accuracy={results[name]["accuracy"]:.4f} | precision-AUC={results[name]["precision"]:.4f} '
              f'| recall={results[name]["recall"]:.4f} | ⏱ {elapsed:.1f}s')

    except Exception as e:
        print(f'  ❌ Error: {str(e)}')

print(f'\n✅ {len(results)} modelos entrenados exitosamente')
```

8. 🏆 Selección del Modelo Óptimo

Criterio: Mayor F1-Score en test, con ROC-AUC como desempate. Se prioriza Recall para detectar el máximo de churners posibles.

```
# ============================================================
# 8.1 MEJOR MODELO
# ============================================================

best_model_name = metrics_df['F1-Score'].idxmax()
best = results[best_model_name]
best_pipeline = best['pipeline']

print('=' * 70)
print(f'🏆 MODELO ÓPTIMO: {best_model_name}')
print('=' * 70)
print(f'''
📊 Métricas en Test:
   Accuracy:   {best["accuracy"]:.4f}
   Precision:  {best["precision"]:.4f}
   Recall:     {best["recall"]:.4f}
   F1-Score:   {best["f1"]:.4f}
   ROC-AUC:    {best["roc_auc"]:.4f}
   PR-AUC:     {best["pr_auc"]:.4f}

📊 Validación Cruzada:
   CV F1:      {best["cv_f1_mean"]:.4f} ± {best["cv_f1_std"]:.4f}
   CV ROC-AUC: {best["cv_roc_mean"]:.4f} ± {best["cv_roc_std"]:.4f}
''')

print('📋 Classification Report:')
print(classification_report(y_test, best['y_pred'], target_names=['No Churn', 'Churn']))
```

```
# ============================================================
# 9.1 CÁLCULO DE SHAP VALUES
# ============================================================

import shap

print(f'🔍 Calculando SHAP para {best_model_name}...')

# Extraer el clasificador del pipeline
if hasattr(best['pipeline'], 'named_steps'):
    clf = best['pipeline'].named_steps.get('clf', best['pipeline'])
else:
    clf = best['pipeline']

# Usar muestra para velocidad
X_sample = X_test[:1000]

# SHAP explainer
if 'Logistic' in best_model_name:
    explainer = shap.LinearExplainer(clf, X_train)
    shap_values = explainer.shap_values(X_sample)
else:
    explainer = shap.TreeExplainer(clf)
    shap_values = explainer.shap_values(X_sample)

# Seleccionar la clase positiva antes de generar cualquier visualización.
# Compatibilidad con versiones de SHAP que devuelven lista o arreglo 3D.
if isinstance(shap_values, list):
    shap_values = shap_values[1]
elif getattr(shap_values, 'ndim', 0) == 3:
    shap_values = shap_values[:, :, 1]

if getattr(shap_values, 'ndim', 0) != 2:
    raise ValueError(f'Formato SHAP inesperado: shape={getattr(shap_values, "shape", None)}')

# Summary plot para la clase positiva (fuga_socio = 1)
shap.summary_plot(shap_values, X_sample, show=False, max_display=15)
fig = plt.gcf()
fig.set_size_inches(8, 10)
plt.title(f'SHAP — {best_model_name}', fontweight='bold')
plt.tight_layout()
plt.savefig('shap_summary.png', dpi=150, bbox_inches='tight')
plt.show()
print(f'✅ SHAP values calculados: shape={shap_values.shape}')
```

UNSCH FACULTAD DE INGENIERÍA DE MINAS, GEOLOGÍA Y CIVIL

ACTA DE SUSTENTACIÓN DE TESIS Nº 79-2026-FIMGC

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

En la Universidad Nacional de San Cristóbal de Huamanga, en la ciudad de Ayacucho, en cumplimiento a la **Resolución Decanal No 438-2026-FIMGC-D**, a los **veinte días del mes de octubre de 2026**, siendo las **10:10 a.m.**, reunidos en el **Auditorio de la Escuela Profesional de Ingeniería Minas**, bajo la presidencia del **Dr. Ing. Víctor Félix FLORES MORENO**, y los miembros: **Dr. Manuel Avelino LAGOS BARZOLA** y **Dr. José Elías YAURI VIDALÓN**, actuando como secretario docente el **Ing. Saul Walter RETAMOZO FERNANDEZ**, para proceder a la sustentación de tesis para optar el **Título Profesional de Ingeniero de Sistemas**, del Bachiller en Ingeniería de Sistemas:

PAVEL WIGBERTO GARCIA ALVARADO

Quien presentó la tesis denominada:

"Predicción de fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena basado en machine learning, 2025"

Los señores miembros del jurado luego de expuesta la tesis y absueltas las preguntas, deliberaron y declararon:

Aprobado con 18 (dieciocho)

Siendo las **11:35 a.m.** del día **20 de octubre del 2026**, culmina el acto de sustentación de tesis, y en conformidad de lo actuado los miembros del jurado firmamos al pie del presente.

Dr. Ing. Víctor Félix FLORES MORENO
Presidente

Dr. José Elías YAURI VIDALÓN
Miembro

Dr. Manuel Avelino LAGOS BARZOLA
Miembro - Asesor

Ing. Saul Walter RETAMOZO FERNANDEZ
Secretario docente de la FIMGC

FACULTAD DE INGENIERÍA
DE MINAS Y CIVIL
Av. Independencia S/N
Ciudad Universitaria
Central Tel. 066 312510
Anexo 151

CONSTANCIA DE ORIGINALIDAD DE TRABAJO DE INVESTIGACIÓN

CONSTANCIA Nº 01-2026-CEMC-FIMGC/UNSCH

El que suscribe; responsable verificador de originalidad de trabajos de tesis de pregrado con el software Turnitin, en segunda instancia para las **Escuelas Profesionales** de la **Facultad de Ingeniería de Minas, Geología y Civil**; en cumplimiento a la **Resolución de Consejo Universitario Nº 039-2021-UNSCH-CU**, Reglamento de Originalidad de Trabajos de Investigación de la Universidad Nacional San Cristóbal de Huamanga y **Resolución Decanal Nº 507-2026-FIMGC-D**, deja constancia de originalidad de trabajo de investigación, que el/la Sr./Srta.

Nombres y Apellidos	:	Pavel Wigberto GARCIA ALVARADO
Escuela Profesional	:	INGENIERÍA DE SISTEMAS
Título de la Tesis	:	"Predicción de fuga de socios en la Cooperativa de Ahorro y Crédito Santa María Magdalena basado en machine learning, 2025"
Evaluación de la Originalidad	:	**9%** Índice de Similitud
Identificador de la entrega	:	3087119770

Por tanto, según los Artículos 12, 13 y 17 del Reglamento de Originalidad de Trabajos de Investigación, es **PROCEDENTE** otorgar la **Constancia de Originalidad** para los fines que crea conveniente.

En señal de conformidad y verificación se firma la presente constancia

Ayacucho, 11 de noviembre de 2026

Mg.Celia Edith Martinez Córdova

Con depósito para Sustentación y Tramites
C.c. Archivo

FACULTAD DE INGENIERIA DE MINAS, GEOLOGIA Y CIVIL
Av. Independencia S/N Ciudad Universitaria
Central Tel. 066 312510
Anexo 151

"Predicción de fuga de socios en la Cooperativa de Ahorro y Crédito Santa Maria Magdalena basado en machine learning, 2025"

por Pavel Wigberto GARCIA ALVARADO

Fecha de entrega: 11-nov-2026 04:18p. m. (UTC-0500)
Identificador de la entrega: 3087119770
Nombre del archivo: Predicción_de_fuga_de_socios_en_la_Cooperativa_de_Ahorro_y_Crédito_Santa_Maria_Magdalena_basado_en_machine_learning_2025.pdf (1.93M)
Total de palabras: 32655
Total de caracteres: 199680

"Predicción de fuga de socios en la Cooperativa de Ahorro y Crédito Santa Maria Magdalena basado en machine learning, 2025"

INFORME DE ORIGINALIDAD

9%	10%	4%	3%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	hdl.handle.net Fuente de Internet	1%
2	repositorioacademico.upc.edu.pe Fuente de Internet	1%
3	repositorio.esan.edu.pe Fuente de Internet	1%
4	purl.org Fuente de Internet	<1%
5	alicia.concytec.gob.pe Fuente de Internet	<1%
6	evaposgrado.puce.edu.ec Fuente de Internet	<1%
7	Submitted to Universidad Internacional del Ecuador Trabajo del estudiante	<1%
8	repositorio.usm.cl Fuente de Internet	<1%
9	repositorio.utdt.edu Fuente de Internet	<1%
10	Submitted to UNIBA Trabajo del estudiante	<1%
11	www.grafiati.com Fuente de Internet	<1%
12	crea.ujaen.es Fuente de Internet	<1%
13	dspace.ups.edu.ec Fuente de Internet	<1%
14	repositorio.upt.edu.pe Fuente de Internet	<1%
	repository.unad.edu.co	

15	Fuente de Internet	<1 %
16	laccei.org Fuente de Internet	<1 %
17	Submitted to Universidad Europea de Madrid Trabajo del estudiante	<1 %
18	uvadoc.uva.es Fuente de Internet	<1 %
19	Submitted to Universidad Nacional de San Cristóbal de Huamanga Trabajo del estudiante	<1 %
20	Submitted to University of Warwick Trabajo del estudiante	<1 %
21	1library.co Fuente de Internet	<1 %
22	repositorio.udh.edu.pe Fuente de Internet	<1 %
23	repositorio.upsc.edu.pe Fuente de Internet	<1 %
24	ciencia.lasalle.edu.co Fuente de Internet	<1 %
25	repositorio.continental.edu.pe Fuente de Internet	<1 %
26	Submitted to Universitat Oberta de Catalunya Trabajo del estudiante	<1 %
27	8eaa8e41-20bc-4582-8433-9286bea9c0e0.filesusr.com Fuente de Internet	<1 %
28	repositorio.autonoma.edu.pe Fuente de Internet	<1 %
29	Essam H. Houssein, Ibrahim A. Ibrahim, Amir Mostafa, Abdullah M. Albarrak, Mina Younan. "SMENN-hybrid: an efficient technique combining the synthetic minority oversampling technique with ensemble learning for diabetes prediction", Scientific Reports, 2025 Publicación	<1 %
30	core.ac.uk Fuente de Internet	<1 %
	reunir.unir.net	

31	Fuente de Internet	<1 %
32	www.revista.ister.edu.ec Fuente de Internet	<1 %
33	repositorio.uroosevelt.edu.pe Fuente de Internet	<1 %
34	repositorio.uasb.edu.ec Fuente de Internet	<1 %
35	worldwidescience.org Fuente de Internet	<1 %
36	Submitted to Universidad de Manizales Trabajo del estudiante	<1 %
37	Submitted to Corporación Universitaria Minuto de Dios,UNIMINUTO Trabajo del estudiante	<1 %
38	repositorio.umariana.edu.co Fuente de Internet	<1 %

Excluir citas	Activo	Excluir coincidencias	< 30 words
Excluir bibliografía	Activo		