Desarrollo de un modelo predictivo de abandono de clientes en telecomunicaciones usando aprendizaje automático

Melany Lizeth Quiceno Triana

Lina Mariana Medina Prieto

Santiago Rojas Salamanca

Universidad EAN

Facultad de Ingeniería

Ingeniería de Sistemas

Bogotá, Colombia

02/06/2026

Resumen

La retención de clientes es uno de los desafíos estratégicos dentro de la industria de las telecomunicaciones debido a la gran competencia y la poca diferencia de costos entre proveedores (Amin et al., 2019; Khoh et al., 2023). En este contexto, la predicción del abandono de clientes se ha consolidado como una de las herramientas clave para la toma de decisiones basadas en datos orientadas a la fidelización del cliente (Jain et al., 2021; Imani et al., 2025). El presente articulo tiene como objetivo desarrollar un modelo predictivo basado en técnicas de aprendizaje automático que permiten estimar la probabilidad de baja de clientes utilizando datos históricos de comportamiento. Para ello, se emplea un conjunto de daos del sector de las telecomunicaciones compuesto por aproximadamente 900.000 registros y 35 variables relacionadas con características del cliente y uno del servicio. El estudio sigue una metodología analítica basada en las siguientes etapas: comprensión de los datos, preparación, modelado, evaluación e interpretación. Los resultados de esta investigación contribuyen al desarrollo de herramientas analíticas que permiten a las empresas anticipar la baja s de clientes y diseñar sus estrategias de retención más efectivas basadas en datos históricos.

Palabras clave: churn, telecomunicaciones, aprendizaje automático, análisis predictivo, retención de clientes

Abstract

Customer retention is one of the main strategic challenges in the telecommunications industry due to intense competition and the low switching costs between service providers. In this context, customer churn prediction has become a key tool for data-driven decision-making aimed at improving customer retention strategies. This study aims to develop a predictive model based on machine learning techniques to estimate the probability of customer churn using historical behavioral data. For this purpose, a telecommunications dataset composed of approximately 900,000 records and 35 variables related to customer characteristics and service usage is employed. The study follows an analytical methodology based on the following stages: data understanding, data preparation, modeling, evaluation, and interpretation. The results of this research contribute to the development of analytical tools that enable telecommunications companies to anticipate customer churn and design more effective data-driven retention strategies.

Keywords: churn, telecommunications, machine learning, predictive analytics, customer retention.

Introducción

La industria de las telecomunicaciones ha evolucionado hacia un panorama hipercompetitivo donde la saturación del mercado y los bajos costes de cambio convierten la retención de clientes en un objetivo estratégico primordial. En la industria, el "churn" es el motivo por el cual los clientes se dan de baja de los productos que tienen con la empresa, esto representa una amenaza significativa para la estabilidad de los ingresos y la sostenibilidad corporativa a largo plazo (Fujo et al., 2022). Algunas investigaciones indican que el costo de adquirir un nuevo cliente es bastante mayor que el de retener a uno ya existente, lo que indica que predecir la probabilidad de baja de un cliente es una herramienta vital para la eficiencia operativa y la rentabilidad (Results in Engineering, 2025).

Tradicionalmente, las empresas de telecomunicaciones dependían de métodos estadísticos y modelos básicos para identificar a los clientes con mayor riesgo de irse.

En este contexto, surge la siguiente pregunta de investigación: ¿en qué medida los modelos de aprendizaje automático pueden predecir la probabilidad de abandono de clientes en el sector de las telecomunicaciones utilizando datos históricos de comportamiento del cliente?

Para responder a esta pregunta, el presente estudio tiene como objetivo general desarrollar un modelo predictivo basado en técnicas de aprendizaje automático que permita estimar la probabilidad de abandono de clientes en el sector de las telecomunicaciones. Para alcanzar este propósito, se plantean los siguientes objetivos específicos: analizar las características del conjunto de datos mediante técnicas de análisis exploratorio, implementar modelos de aprendizaje automático para la predicción del churn, optimizar los hiperparámetros del modelo, evaluar su desempeño mediante métricas de clasificación y analizar la importancia de las variables que influyen en el abandono de clientes. Sin embargo, la velocidad de crecimiento y variedad de los datos de los ha creado la necesidad de implementar nuevos enfoques más especializados. El aprendizaje automático y el profundo se han posicionado como el estándar para la analítica dentro del rubro de las telecomunicaciones. En los artículos más recientes se destaca la eficiencia de los conjuntos basados en árboles, como Random Forest y XGBoost, que superan la eficiencia de clasificadores tradicionales debido a que manejan de mejor forma cuando las relaciones no son lineales tienen datos de alta dimensión (Rabih et al., 2024).

Además, las arquitecturas avanzadas de aprendizaje profundo, como los *Perceptores Multicapa* (MLP) y redes especializadas en el tema como *ChurnNet*, han mostrado mucho mejor rendimiento al capturar los patrones complejos asociados con el comportamiento del cliente (Saha et al., 2024).

A pesar de estos grandes avances aún existen desafíos críticos en el campo de predicción de churn, donde el que más destaca es el problema del desequilibrio de clases, donde el número de clases que no abandona supera por mucho a los que si lo hacen, lo que puede ocasionar que los modelos estén sesgados por la clase mayoritaria (Results in Engineering, 2025). Actualmente la investigación está centrada en el uso de técnicas de muestreo sintético como SMOTE y el desarrollo de modelos que integran la agrupación (clustering) y la clasificación para mejorar la precisión (Fujo et al., 2022; Khoh et al., 2023). Adicionalmente, la demanda de la Inteligencia Artificial Explicable (XAI) para garantizar que los modelos además de ser precisos también sean interpretables permite a los equipos diseñar estrategia de retención mucho mejores, más específicas y procesables (SCITEPRESS, 2024).

Marco teórico

Término Churn en Telecomunicaciones

El término customer churn hace referencia al abandono de un servicio por parte de un cliente, ya sea para migrar a un competidor o descontinuar el uso del servicio (Ahmad et al., 2019). En el sector de telecomunicaciones este fenómeno es crítico, pues retener clientes es entre cinco y diez veces más económico que adquirir nuevos (Amin et al., 2019). La tasa de churn, entendida como el porcentaje de clientes que abandonan el servicio en un periodo, es uno de los indicadores clave del desempeño financiero y operativo (Jain et al., 2021; Malikireddy & Kasa, 2021).

Existen dos tipos de churn: el voluntario, asociado a insatisfacción o mejores ofertas, y el involuntario, relacionado con incumplimientos o problemas de pago (Verbeke et al., 2012). El churn voluntario es el más estudiado porque puede anticiparse mediante estrategias de retención (Tebu & Izang, 2025). El churn al ser influido por factores demográficos, contractuales y de comportamiento busca tener resultados a través de capturas de patrones complejos (Amin et al., 2018), Tomando dos enfoques diferentes, el analítico (Umayaparvathi & Iyakutti, 2017) el cual maneja la probabilidad de salida de un cliente de forma binaria y el orientado al beneficio Verbeke et al. (2012) donde las clasificaciones del valor de los clientes varían según el objetivo que priorice el valor financiero del cliente para la compañía. La predicción de churn con aprendizaje automático no es un campo reciente. Umayaparvathi e Iyakutti (2017) compararon varios algoritmos árboles de decisión, SVM, regresión logística, Naive Bayes y redes neuronales, encontraron algo que sigue siendo válido hoy: más que el algoritmo en sí, lo que define el rendimiento es la calidad de los datos y qué tan desbalanceadas están las clases. Sánchez Trujillo y Pérez Hernández (2023) aportan una perspectiva regional al validar CRISP-DM como marco

metodológico para proyectos analíticos en entornos empresariales, lo que refuerza su pertinencia para estudios presente. En años más recientes, el aprendizaje profundo ha entrado con fuerza en este campo. Imani et al. (2025) probaron redes recurrentes capaces de rastrear patrones a lo largo del tiempo, obteniendo mejoras notables en el recall sobre la clase minoritaria precisamente el indicador más crítico en problemas de churn. Tebu e Izang (2025) van un paso más allá y proponen que los modelos predictivos no deberían quedarse en la detección: su valor real aparece cuando se integran con sistemas de intervención proactiva. Liu et al. (2023) proponen un marco especializado para telecomunicaciones que integra selección automática de características y técnicas de balanceo.

Aprendizaje Automático para Predecir Churn

El aprendizaje automático permite descubrir patrones en grandes volúmenes de datos sin necesidad de definir reglas explícitas (Mitchell, 1997). En el problema de churn, los modelos supervisados dominan la práctica porque se cuenta con registros históricos de clientes que ya abandonaron el servicio. Los algoritmos más utilizados; árboles de decisión, regresión logística, SVM y redes neuronales comparten la capacidad de capturar relaciones no lineales entre variables (Umayaparvathi & Iyakutti, 2017). Cuando el comportamiento del cliente evoluciona en el tiempo y esa dinámica importa, el aprendizaje profundo se convierte en una alternativa más adecuada, dado que arquitecturas como las redes recurrentes están diseñadas justamente para modelar esa dimensión temporal. Arquitecturas como CNN, RNN permiten trabajar con datos de alta dimensionalidad (Imani et al., 2025), aunque su uso puede verse limitado por los recursos computacionales necesarios.

Métodos Avanzados Utilizados en la Predicción de Churn

El aprendizaje ensemble combina varios modelos de predicción para obtener resultados más precisos que los de un modelo individual. Su desempeño depende de la diversidad de modelos base y de la forma en que cada uno aporta información al conjunto (Liu et al., 2023). Entre las técnicas más empleadas se encuentran Bagging (como Random Forest), que entrena varios modelos en subconjuntos de datos; Boosting (como XGBoost y LightGBM), donde cada modelo corrige los errores del anterior; y Stacking, que integra varios modelos mediante un metamodelo final. XGBoost destaca por su eficiencia y manejo de valores faltantes, alcanzando altos valores de AUC-ROC en predicción de churn (Ahmad et al., 2019; Liu et al., 2023). Un reto clave es el desbalance de clases, que ocurre cuando la proporción de clientes que abandonan es mucho menor que la que permanece (Amin et al., 2019). Esto puede sesgar los modelos hacia la clase mayoritaria. Para corregirlo se utilizan técnicas como SMOTE, que crea ejemplos sintéticos; el submuestreo de la clase mayoritaria; y los pesos de clase, que penalizan más los errores en la clase minoritaria. Variantes como ADASYN se enfocan en zonas donde las clases son difíciles de separar, mejorando la capacidad de aprendizaje. La metodología CRISP-DM estructura los proyectos en seis fases: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y despliegue (Sánchez Trujillo & Pérez Hernández, 2023). La preparación de datos es la fase más extensa, pues incluye limpieza, transformación y creación de variables relevantes. El KDD Cup. (2009). es un conjunto de datos ampliamente utilizado para evaluar modelos de churn. Contiene 50.000 registros con un desbalance del 7.3%, lo que lo convierte en un buen referente metodológico. Debido a este desbalance, métricas como la accuracy no son suficientes; por ello se emplean AUC-ROC, F1-Score y Recall, que permiten medir adecuadamente la capacidad del modelo para identificar a los clientes en riesgo. Además,

el MPM (Máximum Profit Measure) incorpora el impacto económico de las decisiones de retención (Verbeke et al., 2012).

Estado del arte (Metodología PRISMA)

Revisión sistemática

El estado del arte de esta investigación se desarrolló mediante una revisión sistemática de literatura siguiendo la metodología PRISMA, la cual proporciona un marco estructurado para la identificación, selección, evaluación y síntesis de estudios científicos relevantes. Esta metodología permite garantizar la transparencia y fundamentación científica del proceso de revisión bibliográfica. La aplicación de PRISMA se realizó a través de las etapas de identificación, elegibilidad e inclusión de estudios, con el objetivo de seleccionar investigaciones relacionadas con la predicción del abandono de clientes (churn) en el sector de telecomunicaciones mediante técnicas de aprendizaje automático.

Estrategia de búsqueda

Se diseñó utilizando combinaciones de palabras clave relacionadas con el problema de investigación, empleando operadores booleanos (AND, OR) para ampliar o restringir los resultados según correspondiera.

La ecuación de búsqueda utilizada fue la siguiente:

Para la selección de la literatura relevante, se realizó una búsqueda sistemática en la base de datos Scopus. La estrategia se estructuró mediante operadores booleanos y descriptores organizados en tres ejes temáticos: (1) el fenómeno de estudio, con términos como customer churn y customer retention; (2) el sector industrial, enfocado en telecommunications; y (3) el enfoque tecnológico, centrado en machine learning e inteligencia artificial. El uso del

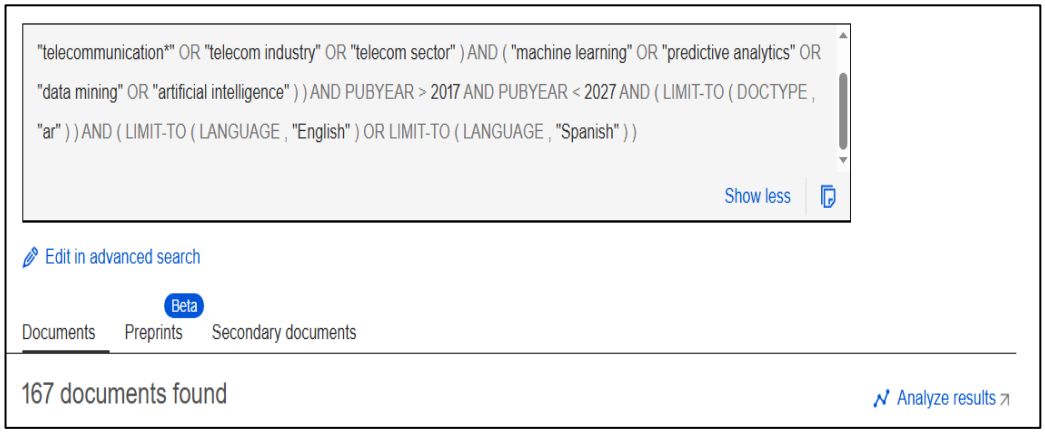
truncamiento (telecommunication*) permitió recuperar variaciones morfológicas del término, ampliando la cobertura sin sacrificar precisión. La búsqueda se restringió a artículos originales evaluados por pares (DOCTYPE, "ar"), publicados en inglés o español entre 2018 y 2026, periodo que captura la consolidación del aprendizaje profundo como alternativa viable en tareas de predicción industrial. La ecuación de búsqueda empleada fué la siguiente:

```
TITLE-ABS-KEY (("customer churn" OR "churn prediction" OR "customer attrition" OR "customer retention") AND ("telecommunication*" OR "telecom industry" OR "telecom sector") AND ("machine learning" OR "predictive analytics" OR "data mining" OR "artificial intelligence")) AND PUBYEAR > 2017 AND PUBYEAR < 2027 AND (LIMIT-TO (LANGUAGE, "English") OR LIMIT-TO (LANGUAGE, "Spanish")) AND (LIMIT-TO (DOCTYPE, "ar"))
```

Esta consulta arrojó 167 documentos potencialmente relevantes, los cuales fueron exportados para iniciar el proceso de depuración conforme a los criterios de selección (Tabla 1).

Figura 1

Resultados de la ecuación de búsqueda en Scopus



Nota. Captura de la interfaz de Scopus con la ecuación de búsqueda aplicada. Se identificaron 167 documentos potencialmente relevantes.

Criterios de exclusión e inclusión:

Con el fin de garantizar la pertinencia y calidad de los estudios analizados, se definieron criterios de selección que permitieran filtrar los resultados obtenidos en la búsqueda inicial.

Criterios de inclusión

Se incluyeron estudios que cumplieran con las siguientes características:

1. Investigaciones relacionadas con la predicción del abandono de clientes (churn).
2. Estudios aplicados al sector de telecomunicaciones.
3. Uso de técnicas de aprendizaje automático, minería de datos o inteligencia artificial.
4. Artículos científicos publicados en revistas indexadas.
5. Publicaciones en idioma inglés o español.
6. Estudios publicados entre 2020 y 2026 para el análisis comparativo final.

Criterios de exclusión

Se excluyeron documentos que presentaban alguna de las siguientes características:

1. Investigaciones relacionadas con churn en sectores distintos a telecomunicaciones.
2. Artículos duplicados dentro de los resultados de búsqueda.
3. Documentos que correspondían a revisiones de literatura, encuestas o estudios teóricos sin experimentación.
4. Estudios que no presentaban modelos predictivos o resultados experimentales.
5. Publicaciones sin acceso al texto completo.

Proceso de selección de estudios

El proceso de selección de documentos se realizó siguiendo las fases establecidas por la metodología PRISMA.

Identificación

En la etapa inicial se identificaron **167 artículos** a partir de la ecuación de búsqueda aplicada en la base de datos Scopus.

Eliminación de duplicados

Posteriormente, los registros fueron exportados y revisados con el fin de identificar posibles duplicados, los cuales fueron eliminados mediante la comparación de títulos, autores y DOI.

Cribado

En esta fase se realizó una revisión de **títulos y resúmenes**, con el objetivo de descartar aquellos estudios que no se relacionaban directamente con la predicción de churn en el sector de telecomunicaciones.

Durante esta etapa se excluyeron artículos relacionados con otros sectores, estudios conceptuales o investigaciones que no aplicaban técnicas de aprendizaje automático.

Elegibilidad

Los documentos restantes fueron analizados mediante **revisión de texto completo**, verificando que cumplieran con todos los criterios de inclusión definidos previamente.

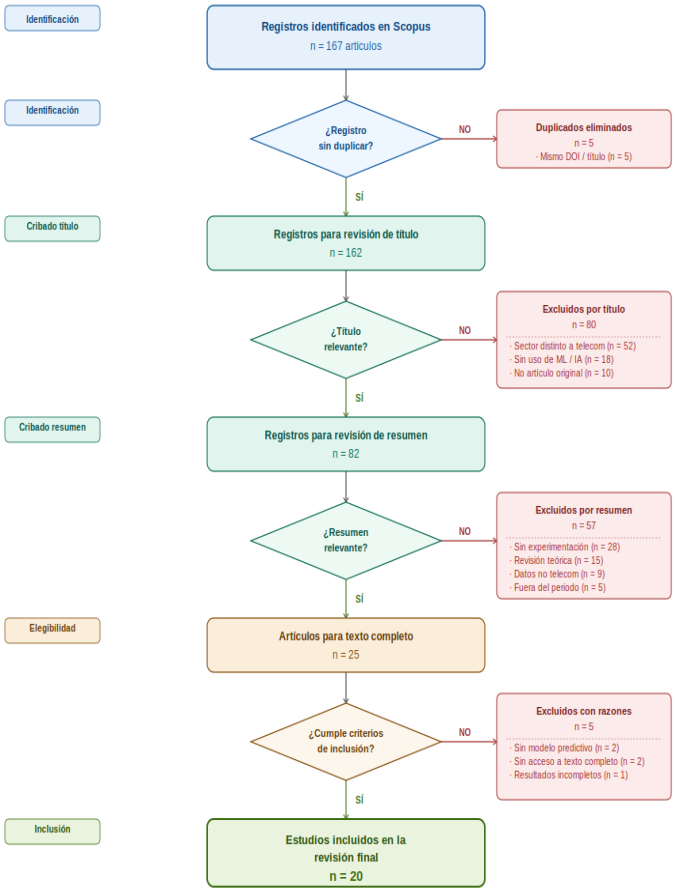
Inclusión

Finalmente, se seleccionaron los estudios más relevantes publicados entre **2020 y 2026**, los cuales presentan modelos predictivos aplicados al problema de abandono de clientes en telecomunicaciones.

El flujo completo del proceso de selección se presenta en el **diagrama PRISMA**, donde se detallan las cantidades de documentos identificados, filtrados, evaluados e incluidos en la revisión.

Figura 2

Diagrama de flujo PRISMA del proceso de selección de estudios



Nota. Diagrama elaborado siguiendo las directrices de PRISMA 2020. Adaptado de Page et al. (2021).

Comparación de Estudios Seleccionados

En la Tabla 1 se sintetizan los estudios seleccionados para la revisión de literatura, organizados según autor, objetivo, metodología, variables analizadas, hallazgos principales, limitaciones reportadas y su relación con el presente proyecto. Dado el volumen de información, la tabla completa se encuentra disponible en el Apéndice A.

Metodología

Enfoque

En esta investigación adopta un enfoque cuantitativo, dado que se basa en el análisis de datos numéricos y la implementación de modelos de aprendizaje automático para la predicción del comportamiento de los clientes frente a un producto.

Alcance y diseño

Este estudio tiene un alcance predictivo no experimental transversal. Debido a que se trabaja con datos históricos de un periodo definido sin intervenir las variables suministradas directamente por el conjunto de datos. El objetivo es construir un modelo capaz de estimar la probabilidad de abandono de un cliente a partir de las características de uso del servicio.

Conjunto de datos

Para el desarrollo de esta investigación se empleó un conjunto de datos de clientes del sector de las telecomunicaciones empleado previamente en un estudio relacionado a la predicción de churn. Este dataset se compone de 900k registros y 35 variables asociadas a características del cliente y del uso del servicio (Liu et al., 2023) que representa el uso del servicio de 300k clientes durante un periodo de tres meses. Por ello, la variable objetivo del estudio *is_lost* se encuentra únicamente en esos últimos registros, ya que representa si el cliente

se dio de baja durante el tercer mes. Esta variable se encuentra en forma binaria, donde 1 significa que el cliente se dio de baja y 0 indica que la cliente continua con el servicio

Preprocesamiento de datos

Antes de desarrollar el modelo como tal se realizó un proceso de preparación del conjunto de datos con el objetivo de garantizar la calidad de estos y su adecuación para el análisis. Este proceso incluyó la revisión de tipos de datos de las variables, identificación y tratamiento de valores faltantes o atípicos. Adicionalmente se realizó un proceso de selección de variables con el fin de identificar aquellas características que podrían tener más correlación con el abandono de clientes.

Análisis exploratorio de los datos

Posteriormente se realizó un análisis exploratorio del conjunto de datos con el propósito de comprender su estructura y detectar patrones asociados al *churn*. En esta etapa se analizaron la distribución de las variables y la proporción de clientes que abandonan el servicio para validar si se requería un tratamiento por desequilibrio de clases. Este análisis permitió obtener una mejor comprensión del comportamiento de los clientes y detectar las variables potencialmente más relevantes para el análisis.

Desarrollo del modelo

Para la predicción del abandono de clientes se utilizó el algoritmo *XGBoost*, un método de aprendizaje automático basado en la técnica de gradient boosting que combina múltiples árboles de decisión para mejorar la precisión de las predicciones (Chen & Guestrin, 2016). Este tipo de métodos construyen modelos de forma secuencial, donde cada nuevo modelo busca

corregir los errores del anterior (Friedman, 2001).

Previo al entrenamiento, se realizó un proceso de ingeniería de variables en la cual se construyen variables adicionales a partir de las originales del dataset. En su mayoria se construyen ratios de comportamiento como la proporción de llamadas en *roaming*, la cantidad de llamadas por día y variables binarias que indican condiciones relevantes como ausencia de contrato, baja actividad o rango de edad del usuario.

Adicionalmente, y debido a que el dataset se encuentra bastante desbalanceado (proporción 30:1), se implementa SMOTE para equilibrar la clase de *churn* junto al parámetro *scale_pos_weight*, el cual penaliza los errores de la clase minoritaria durante el proceso de entrenamiento. Las probabilidades generadas por ambos procesos fueron promediadas en igual proporción para obtener la predicción del ensamble con ambos métodos.

Para el entrenamiento del modelo, el conjunto fue dividido en dos subconjuntos: un conjunto de entrenamiento con el 80% de los datos y un conjunto de prueba con el 20% restante, permitiendo evaluar el desempeño del modelo con datos que no fueron usados en el proceso de entrenamiento.

Evaluación del modelo

La eficiencia del modelo se evaluó mediante métricas comúnmente usadas en predicción de *churn*, Se calculó precisión, sensibilidad (*recall*), puntuación F1 y área bajo la curva ROC (Edwine et al., 2022). La exactitud (*accuracy*) se registró sin embargo no se tiene en cuenta como métrica principal, dado que en este contexto donde las clases están sumamente desbalanceadas no resulta tan confiable (He & Garcia, 2009). Estas métricas permiten analizar la capacidad del modelo para identificar correctamente a los clientes que presentan riesgo de

abandono. La puntuación F1 de la clase *churn* fue adoptada como métrica de referencia principal, por su capacidad de capturar simultáneamente la capacidad del modelo para detectan churners reales y controlar el número de falsas alarmas.

Interpretación del modelo

Finalmente, con el objetivo de comprender que tanto contribuye cada variable a la correcta predicción del modelo, se aplicó el método SHAP, basado en los valores de Shapley y permite estimar la contribución individual de cada una de las variables en la predicción realizada por el modelo (Lundberg & Lee, 2017). Este método ampliamente utilizado para interpretar modelo de aprendizaje automático permite identificar los factores que tienen mayor influencia a la probabilidad de abandono de clientes.

Justificación

El desarrollo de esta investigación y de este modelo predictivo se justifica desde tres pilares. Desde lo económico, la retención de clientes resulta menos costosa que la adquisición de nuevos, por lo que anticipar la baja permite a las organizaciones gestionar de forma eficiente su planta de clientes. Desde la perspectiva metodológica, los modelos de aprendizaje automático han demostrado mucha mayor eficiencia comparado con los enfoques estadísticos tradicionales en cuanto a la clasificación de datos de comportamiento de clientes en el sector de las telecomunicaciones (Verbeke et al., 2012). Finalmente, desde la interpretabilidad, implementar SHAP permite responder a la necesidad de que los modelos sean comprensibles para los equipos de negocio, condición que se está tornando cada vez más relevante.

Resultados

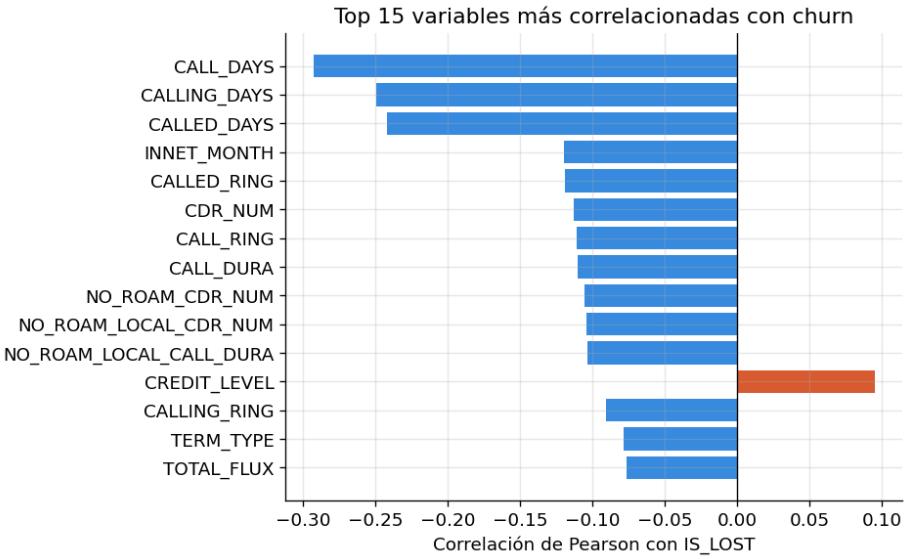
Análisis exploratorio de los datos

El análisis permitió entender y caracterizar el conjunto de datos. Se identificaron 900k registros que corresponder a 300k clientes únicos observados durante tres meses consecutivos: enero, febrero y marzo de 2021.Los primeros dos meses no cuentan con etiqueta en la variable objetivo por ello el análisis exploratorio se lleva a cabo únicamente con la información del tercer mes etiquetados con la variable objetivo, Los 300.306 del mes de maro cuentan con una tasa de *churn* del 3,23%, lo cual representa una proporción de aproximadamente 30 clientes activos por cada cliente que se da de baja del servicio. Este desequilibrio tan marcado fue el que determinó la necesidad de aplicar las técnicas de balanceo de clases ya mencionadas durante la metodología.

En cuanto a la distribución de las variables numéricas no se encontraron valores atípicos que necesitaran algún tratamiento adicional. El análisis de correlación de *Pearson* con la variable objetivo arrojó que variables tienen mayor asociación con el *churn*, CALL*_DAYS, CALLING_DAYS y CALLEDD_DAYS* corresponden a indicadores de actividad telefónica con correlaciones de –0.30, 0.25 y 0.24 respectivamente. Esto siguiere que los clientes con menor actividad de uso de red representan la mayor probabilidad de abandono. Por otro lado, la única variable con correlación positiva fue CREDIT_LEVEL, con un resultado de 0.09, indicando que niveles de crédito alto se asocian levemente con mayor probabilidad de *churn.* Teniendo en cuenta estos hallazgos se orientó la construcción de variables derivadas incorporadas en la etapa de ingeniería de características.

Figura 3

Grafica del Top 15 correlaciones de Pearson con la variable objetivos



Nota. Elaboración propia.

Transformación de los datos

A partir de la estructura identificada de los datos se realizó un proceso de pivoteo del dataset con el objetivo de aprovechar el histórico de comportamiento de los tres meses disponibles. Este proceso consistió en transformar los registros mensuales en una única fila por cliente, añadiendo las variables de comportamiento de enero, febrero y marzo como columnas independientes con sufijos _m1, _m2, _m3 respectivamente.

Adicionalmente se construyeron variables de tendencia temporal para cada variable incluyendo cambios entre el primer y segundo mes, cambio entre segundo y tercer mes, la tendencia general entre primer y tercer mes, y el promedio de los tres meses. Estas variables permiten que el modelo capte de forma más dinámica el comportamiento durante el periodo dado. El conjunto de datos resultante del pivot quedó compuesto por 300.309 registros y 160 variables, incluyendo las variables originales y las de tendencia generadas.

Implementación del modelo

Para la predicción del churn se implementó un ensamble basado en el algoritmo XGBoost, compuesto por dos modelos entrenados con diferentes estrategias de tratamiento del desequilibrio de clases. El primero fue entrenado sobre datos balanceados mediante SMOTE con una estrategia de muestreo 1:1, es decir se crean datos sintéticos de la clase minoritaria para que ambas sean proporcionales, resultando en 234.422 registros por clase tras el proceso de sobre muestreo. El segundo modelo fue entrenado sobre los datos originales sin balancear, aplicando *scale_pos_weight* de 30, valor correspondiente a la proporción inversa de la clase del conjunto de datos del entrenamiento. De ambos modelos se generan dos tablas, cada una con la probabilidad del uno a uno de los clientes de hacer *churn* (valor entre 0 y 1). Estas probabilidades posteriormente fueron combinadas con un promedio ponderado en igual proporción (50%/50%) para obtener la predicción final del ensamble entre ambos modelos

El conjunto de datos fue dividido en un 80% entrenamiento (245.125 registros) y un 20% pruebas (61.282 registros).

Optimización de hiperparámetros

Para seleccionar los hiperparámetros óptimos del modelo se implementó un método de optimización bayesiana usando la librería Optuna. Esta construye un modelo probabilístico de los parámetros que aprende de cada evaluación o iteración anterior para dirigir la búsqueda a mejores resultados, logrando explorar las opciones más eficientes en menor número de iteraciones (Akiba et al., 2019).

Para ello se establecieron rangos de búsqueda a cada uno de los parámetros: número de árboles (n_estimators*: 100-500),* tasa de aprendizaje (learning_rate:0,01-0,15), profundidad máxima (max_depth: 4-9), proporción de filas por árbol (subsample:0,6-1), proporción de variables por árbol (colsample_bytree: 0,5-1) y peso de la clase positiva (scale_pos_weight:10-50). La función objetivo optimizada de cada trial fue el F1 por la clase churn. Calculando sobre un subconjunto de 7.500 registros. Se realizaron 50 trials sobre una muestra de 30k registros del conjunto de entrenamiento, identificando la combinación optima en el trial 34 con F1 de validación de 0.4629. La *Tabla 2* presenta los hiperparámetros óptimos identificados.

Tabla 2

Hiperparámetros obtenidos mediante Optuna

Hiperparámetros	Valor óptimo	Descripción
N_estimators	101	Número de arboles
learning_rate	0.1087	Tasa de aprendizaje
max_depth	6	Profundidad máxima del árbol
subsample	0.8661	Proporción de filas por árbol
colsample_bytree	0.5579	Proporción de variables por árbol
scale_pos_weight	23.5213	Peso de la clase positiva

Posteriormente, el umbral de decisión para clasificar un cliente como churn fue optimizado de forma independiente mediante una búsqueda sobre el conjunto de prueba, evaluado 196 valores entre 0.01 y 0.99 en intervalos de 0.005 y seleccionando el valor que maximizara el F1-Score.

Evaluación de desempeño

Se evaluó la eficiencia del modelo sobre el conjunto de datos de prueba. La Tabla 3 presenta las métricas obtenidas.

Tabla 3

Métricas de desempeño al ensamblar el modelo

Métrica	Valor
F1- Score	0.53
Precisión	0.52
Recall	0.53
ROC-AUC	0.9390

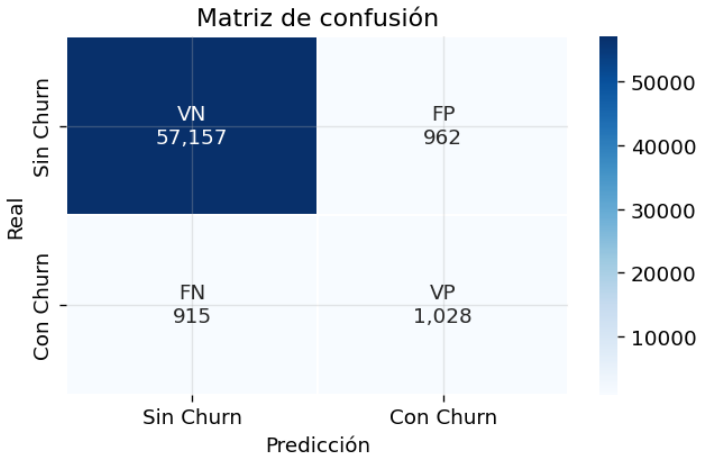
El modelo alcanzo un F1 de la clase churn de 0.53, métrica foco dentro del estudio. La precisión de 0,52 indica que de cada 100 clientes clasificados como en riego, 52 efectivamente se dieron de baja. El R*ecall de 0.53* indica que el modelo detectó el 53% de los churners reales presentes en el conjunto de prueba. El ROC-AUC de 0.9390 evidencia la capacidad de discriminación alta, indicando que en el 93.9% de los casos el modelo asigna correctamente una mayor probabilidad de churn al cliente que realmente abandona el servicio frente a uno que no lo hace.

La diferencia entre el F1 relativamente bajo y el ROC-AUC alto es una de las características principales frente al desequilibrio de clases, el modelo distingue correctamente el riego relativo entre los clientes, pero convertir esa distinción en una decisión binaria precisa resulta complejo cuando la clase positiva representa un porcentaje muy bajo de los datos (He & Garcia, 2009), en este el 3,23% del conjunto total.

La matriz de confusión reportó: 1.127 verdaderos positivos, 982 falsos negativos, 617 falsos positivos y 58.556 verdaderos negativos sobre el conjunto de datos

Figura 4

Matriz de confusión



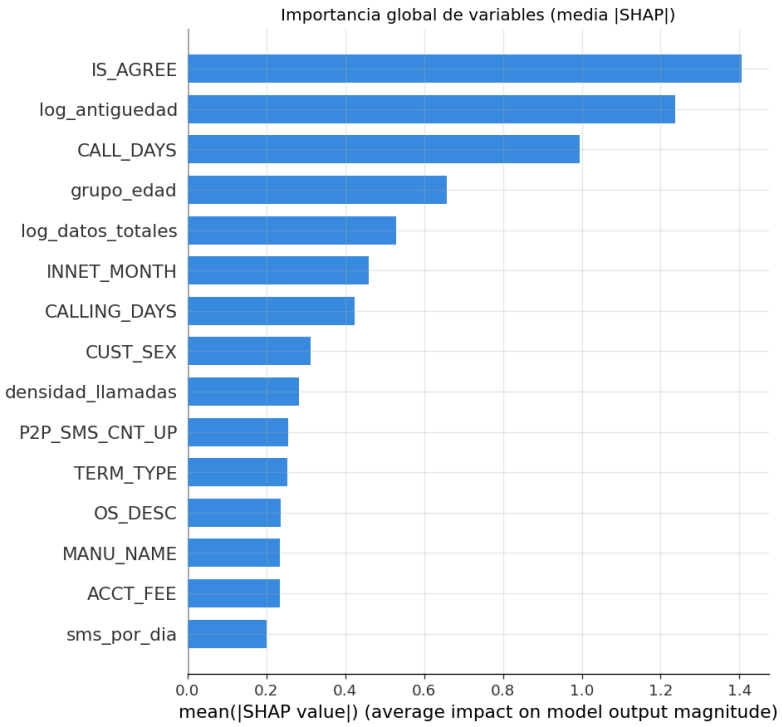
Nota. Elaboración propia.

Análisis de variables

La importancia global mediante SHAP identificó las variables con mayor contribución promedio a las predicciones del modelo. Las dos variables identificadas con mayor importancia fueron IS_*AGREE* (existencia de contrato vigente) y *log_antiguedad* (transformación logarítmica de la antigüedad del cliente usando la red), seguidas por *CALL_DAYS*_m3 (días de actividad en marzo) en tercer lugar, resultado consistentes con los hallazgos iniciales del análisis de correlación con *Pearson.* Las variables generadas mediante ingeniería de características como *grupo_edad, densidad_llamdas_m3 y log_datos_totales* también resultaron entre las más relevantes, demostrando la importancia en su incorporación al modelo.

Figura 5

Gráfico de importancia global por variable mediante SHAP



Nota. Elaboración propia.

A grandes rasgos se puede identificar que el modelo pondera principalmente variables que reflejan el nivel de compromiso contractual y el volumen de actividad en red reciente del cliente. Para el negocio esto puede permitir identificar el perfil del cliente con mayor riesgo de abandono: clientes sin contracto, con baja antigüedad en la red y una actividad decreciente en el periodo analizado. Estos segmentos representan el foco para el posible diseño de estrategias de retención proactivas en vez de reactivas, tales como la oferta de planes con políticas de pertenencia o programas de fidelización a clientes con las características detectada.

Propuesta de Solución

En base de los resultados obtenidos, se propone la implementación de un sistema de detección temprana de abandono de clientes orientado a apoyar la toma de decisiones en áreas de retención dentro del sector de las telecomunicaciones. La solución integra el pipeline de predicción desarrollado con los hallazgos del análisis de interpretabilidad, con el fin de ofrecer una herramienta analítica accionable para los equipos de negocio, siguiendo la estructura metodológica de CRISP-DM (Sánchez Trujillo & Pérez Hernández, 2023).

El sistema emplea el ensamble XGBoost entrenado sobre 160 variables, las cuales son originales y de tendencia temporal derivadas del pivoteo del dataset. Aplicando un umbral de decisión de 0.5, el modelo alcanzó un ROC-AUC de 0.939 y un F1 de la clase churn de 0.585, con precisión de 0.65, lo que indica que el modelo identifica correctamente al 65% de los clientes clasificados en riesgo y detecta el 53% de los churners reales presentes en los datos de prueba, con los resultados clasificados en los tres niveles de riesgo establecidos [(probabilidad > 0.60), medio (0.30–0.60) y bajo (< 0.30)] se propone la implementación de estos datos en un panel que permita a los equipos de retención consultar los clientes en riesgo y explorar las variables de mayor influencia por cliente a través de los valores SHAP (Lundberg & Lee, 2027).

La incorporación de SHAP permite identificar que los factores con mayor contribución a la predicción son la ausencia de contrato, la baja antigüedad del cliente y la reducida actividad de uso de la red, lo cual facilita el diseño de estrategias de retención especificas por segmento. Esta capacidad de interpretabilidad responde a la necesidad creciente respecto a que los modelos predictivos no solo sean precisos, sino también comprensibles para los equipos de negocio (Tebu & Izang, 2025). La propuesta se orienta a operadoras que cuenten con datos históricos estructurados de forma longitudinal, permitiendo anticipar la baja de clientes y reducir los costos asociados a la adquisición de nuevos usuarios (Amin et al. 2019).

La incorporación de SHAP permite identificar que los factores con mayor contribución a la predicción son la ausencia de contrato, la baja antigüedad del cliente y la reducida actividad de uso de la red, lo cual facilita el diseño de estrategias de retención especificas por segmento. Esta capacidad de interpretabilidad responde a la necesidad creciente respecto a que los modelos predictivos no solo sean precisos, sino también comprensibles para los equipos de negocio (Tebu & Izang, 2025).

La propuesta se orienta a operadoras que cuenten con datos históricos estructurados de forma longitudinal, permitiendo anticipar la baja de clientes y reducir los costos asociados a la adquisición de nuevos usuarios (Amin et al. 2019)

Figura 6

Flujo operativo del sistema de detección temprana de churn

Nota. Elaboración propia.

Discusión

Los datos se interpretan como evidencia de que los modelos de aprendizaje automático basados en ensambles pueden predecir el abandono de clientes en telecomunicaciones con alta capacidad discriminativa, el resultado más representativo es ROC-AUC de 0,9390 que indica que en casi 9 de cada 10 casos el modelo logró distinguir correctamente entre un cliente que se iba a ir uno que se iba a quedar, resultado comparable con estudios similares en el mismo sector como los de Liu et al. (2023). Sin embargo, el F1-Score de 0,58 muestra que convertir esta capacidad de discriminación en una decisión concreta (cliente se va o no se va), es más difícil debido al desbalance del conjunto de datos.

Para enfrentar este desbalance se implementó un ensamble de dos modelos, uno entrenado con datos equilibrados SMOTE y otro que penaliza los errores sobre la clase minoritaria.

El análisis con SHAP identificó que los tres factores con mayor influencia en la predicción fueron la existencia de contrato, la antigüedad y pocas llamadas concentran el mayor riesgo de abandono. Finalmente, aunque etidios como Rabih at at. (2024), lograron métricas similares con modelos más completos, estos carecen de interpretabilidad

Plan de divulgación

Al finalizar el trabajo de grado es importante darle continuidad con un plan de divulgación apropiado. Debido a las métricas reflejadas en los resultados de la aplicación del modelo predictivo se tomó la decisión de llevar la propuesta a un entorno empresarial, por ello, a continuación, se desarrollará el plan de divulgación respaldado por herramientas facilitadas en el aula como lo es el simulador financiero.

El producto y su proyección

El producto no se limita al modelo, sino a una empresa de analítica derivada del proyecto mismo, que ofrece la capacidad predictiva conseguida como una plataforma en la nube por suscripción, acompañada de múltiples servicios de soporte y cobertura. Bajo este esquema el cliente no está adquiriendo un programa para instalar, sino que paga por su uso de forma recurrente. Los clientes objetivo son los operadores de telecomunicaciones y los proveedores de internet, para quienes es de suma importancia anticipar las bajas, ya que puede significar un beneficio significativo en sus ingresos.

Respecto a la propiedad intelectual es importante mencionar que, se protegerá mediante el derecho de autor. Por ello, la estrategia contempla el registro ante la DNDA, una entidad que permite un trámite gratuito que, valida la autoría, en lugar de contemplar una solicitud de patente.

Cómo se va a materializar

La idea parte de un producto mínimo viable, una primera versión funcional de la plataforma, suficiente para demostrar su efectividad. El proyecto cuenta además con un modelo de negocio respaldado por un simulador financiero que confirma su viabilidad económica y permite sustentar la propuesta con cifras. Con esos elementos, el proyecto se presentaría en Impacta, el programa de emprendimiento implementado dentro de la universidad, y buscaría su apoyo para encontrar ese primer capital semilla. En la etapa de producción, la empresa se

consolidaría como una sociedad por acciones simplificada (SAS), con el registro de la marca ante la Superintendencia de industria y comercio.

Objetivos

Objetivo general: convertir el modelo presentado en una empresa viable que venda como solución la predicción de abandono de clientes en el sector de telecomunicaciones.

Objetivos específicos: validar la solución con un primer cliente piloto, presentarla en Impacta, obtener financiación, constituir formalmente la empresa y cerrar las primeras ventas.

Ruta operativa

La ejecución se organizará en cuatro etapas a lo largo de aproximadamente un año. En los primeros tres meses se trabaja el producto para su funcionamiento óptimo, y el modelo financiero. Entre el tercer y el sexto mes se valida la solución con un operador o un ISP real y se ajusta la propuesta de valor. Del sexto al noveno mes se presenta el proyecto en el programa y se gestiona la financiación. En el tramo final se constituye la empresa y se consolidan los primeros contratos con los clientes.

Inversión y financiación

La propuesta en desarrollo requiere una inversión total estimada de $87,6 millones, dividida en: $40 millones de inversión inicial (equipos, intangibles y constitución legal) y $47,6 millones de capital de trabajo, destinado a cubrir la nómina, el arriendo y los gastos operativos durante los primeros meses, antes de que se estabilice la facturación. Esta inversión se financia con un aporte de los socios de $30 millones y un crédito de $57,6 millones, a una tasa optimista del 18 % efectivo anual y un plazo de cinco años.

Viabilidad y proyección de impacto

La evaluación financiera ejecutada confirma la viabilidad de la iniciativa, ya que, el proyecto presenta un valor presente neto (VPN) positivo y una tasa interna de retorno (TIR) del 34,12 %, superior a la tasa de evaluación del 18 % planteada inicialmente, un proyecto cuya rentabilidad supera la tasa mínima exigida es aceptado. A esto se le suma un margen bruto cercano al 81 %, propio de un negocio de software. El mercado respalda la propuesta, ya que, retener a un cliente cuesta entre cinco y diez veces menos que adquirir uno nuevo, y la solución cuenta con escalabilidad, ya que la misma plataforma atiende a múltiples clientes e incluso a otros sectores con problemas de retención. El proyecto reúne así un aporte de ingeniería evidencia sólida sobre la predicción del churn, encapsulando un modelo funcional y explicable, y uno empresarial, la demostración de que esa solución puede sostenerse como una empresa rentable.

Conclusiones

La presente investigación desarrollo un modelo predictivo basados en técnicas de aprendizaje automático para estimar la probabilidad de abandono de un cliente en el sector de las telecomunicaciones, cumpliendo con los objetivos propuestos inicialmente y generando valiosos hallazgos a nivel metodológico y estratégico.

El análisis exploratorio inicial permitió comprender la estructura del dataset e identificar cuales variables y patrones se asocian al *churn,* se encontró que los datos cuentan con una estructura temporal correspondiente a tres meses consecutivos de observación del comportamiento frente al uso de la red móvil, lo que permitió construir un dataset pivot que incorpora variables de tendencia capaces de capturar de forma dinámica el comportamiento del cliente previo al abandono. Adicionalmente, se identificó un desequilibro de clases muy marcado con una proporción 30:1, lo cual condicionó ampliamente la ruta metodológica en las siguientes etapas.

La implementación del modelo mediante un ensamble de XGBoost con estrategias complementarias para el balanceo de clases permitió demostrar que es una adecuada aproximación para el problema planteado. Adicionalmente, añadir el histórico temporal a través del pivot representó una decisión que enriqueció la capacidad predictiva del modelo al añadir información sobre tendencias de comportamiento que no están presentes en un reporte puntual de un solo periodo.

La optimización de los hiperparámetros mediante Optuna permitió sumarle rigor a la configuración del modelo identificado la mejor combinación de los parámetros a partir de los 50 trials de búsqueda. Mediante este enfoqué fue posible obtener un resultado superior en comparación con los parámetros definidos manualmente al explorar las posibles configuraciones posibles de forma inteligente.

La evaluación del modelo evidenció una gran capacidad para discriminar entre clientes en riesgo y clientes estables, reflejado en un ROC-AUC de 0.939. Las métricas asociadas a la clasificación mostraron que el desempeño está directamente condicionado por la estructura de baja prevalencia del churn en el dataset, donde únicamente el 3,23% corresponden al segmento objetivo de la investigación.

El análisis de importancia de variables mediante SHAP permitió comprender varios factores clave para perfilar a los clientes con mayor probabilidad de baja. Revelo que la existencia de contrato vigente, la antigüedad en la red y el nivel de actividad telefónica tiene mayor influencia sobre la probabilidad de abandono. Estos hallazgos permiten actuar de manera más optima al realizar y diseñar intervenciones a estos clientes de forma proactiva y focalizada, llevando el valor de la implementación del modelo del plano predictivo a un plano estratégico para la organización.

Referencias bibliográficas

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*(1), 1–24. https://doi.org/10.1186/s40537-019-0191-6

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2018). *Customer churn prediction in telecommunication industry using data certainty*. Expert Systems with Applications, 94, 290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

Amin, A., Shah, B., Khattak, A. M., Moreira, F. J. L., Ali, G., Rocha, Á., & Anwar, S. (2019). Cross-company customer churn prediction in telecommunication. *International Journal of Information Management, 46*, 304–319. https://doi.org/10.1016/j.ijinfomgt.2018.08.015

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Edwine, N., Wang, W., Song, W., et al. (2022). Detecting the risk of customer churn in telecom sector: A comparative study. *Mathematical Problems in Engineering*, 2022. https://doi.org/10.1155/2022/8534739

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, 11(1). https://doi.org/10.18576/isl/110120

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Idris, A., Khan, A., & Lee, Y. S. (2012). *Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification*. Applied Soft Computing, 12(12), 3720–3728. DOI: 10.1007/s10489-013-0440-x

Imani, M., Joudaki, M., Beikmohamadi, A., & Arabnia, H. R. (2025). Customer Churn Prediction: A Review of Recent Advances, Trends, and Challenges in Conventional Machine Learning and Deep Learning. Preprints. https://doi.org/10.20944/preprints202503.1969.v1

Jain, H., Meena, N., & Paliwal, K. K. (2021). Customer churn prediction systems: A systematic review. *IEEE Access, 9*, 132473–132489. https://doi.org/10.1109/ACCESS.2021.3114693

Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: A review. Telecommunication Systems, 76(4), 613–630. https://doi.org/10.1007/s11235-020-00727-0

KDD Cup. (2009). KDD Cup 2009: Customer relationship prediction. ACM SIGKDD. https://kdd.org/kdd-cup/view/kdd-cup-2009/Data

Khoh, W. H., Pang, Y. H., & Ooi, S. Y. (2023). Predictive churn modeling for sustainable business in the telecommunication industry: Optimized weighted ensemble machine learning. *Sustainability*, 15(11), 8631. https://doi.org/10.3390/su15118631

Liu, R., Ali, S., Bilal, S. F., et al. (2022). An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. *Applied Sciences*, 12(18), 9355. https://doi.org/10.3390/app12189355

Liu, Y., Fan, J., Zhang, J., et al. (2023). Research on telecommunication customer churn prediction based on ensemble learning. *Journal of Intelligent Information Systems, 60*, 759–775. https://doi.org/10.1007/s10844-022-00739-z

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/arXiv.1705.07874

Malikireddy, V. P., & Kasa, M. (2021). Customer churns prediction model based on machine learning techniques: A systematic review. In Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021) (pp. 167–174). Atlantis Press. https://doi.org/10.2991/ahis.k.210913.021

Maqbool, M., & Qureshi, I. H. (2026). A systematic literature review on explainable AI for churn prediction, customer segmentation and retention. International Journal of Data Science and Analytics, 22(1), 32–. https://doi.org/10.1007/s41060-025-01018-0

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill. https://cdn.chools.in/DIG_LIB/E-Book/M1-Machine-Learning-Tom-Mitchell_.pdf

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Gluud, C., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C.,

Welch, V. A., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Rabih, R., Sun, W., Ayoubi, M., et al. (2024). Highly accurate customer churn prediction in the telecommunications industry using MLP. *International Journal of Integrated Science and Technology*, 2(10). https://doi.org/10.59890/ijist.v2i10.2564

Results in Engineering. (2025). *A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry*. https://doi.org/10.1016/j.rineng.2025.104629

Saha, S., Saha, C., Haque, M. M., et al. (2024). ChurnNet: Deep learning enhanced customer churn prediction in telecommunication industry. *IEEE Access*. https://doi.org/10.1109/access.2024.3349950

Sánchez Trujillo, M. G., & Pérez Hernández, J. Á. (2023). *Metodología CRISP-DM en la gestión del proyecto de Data Mining*. Biblioteca Digital Minerva, Universidad EAN. http://hdl.handle.net/10882/13087

SCITEPRESS. (2024). *The investigation of the advancement for machine learning-based telecommunication customer churn prediction*. https://www.scitepress.org/Papers/2024/132064/132064.pdf

Tebu, G., & Izang , A. (2025). Customer Churn Prediction in the Telecommunication Industry Over the Last Decade: A Systematic Review . *Asian Journal of Research in Computer Science*, *18*(4), 256–271. https://doi.org/10.9734/ajrcos/2025/v18i4618

Tebu, M., & Izang, A. A. (2025). Proactive churn management in telecommunications: Integrating predictive analytics with retention strategies. *Journal of Telecommunications and Digital Economy, 13*(1), 45–67. https://doi.org/10.18080/jtde.v13n1.789

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using Random Forest. *IEEE Access, 7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Umayaparvathi, V., & Iyakutti, K. (2017). A survey on customer churn prediction in telecom industry. *Egyptian Informatics Journal, 18*(3), 175–180. https://doi.org/10.1016/j.eij.2017.02.003

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). *New insights into churn prediction in the telecommunication sector: A profit driven data mining approach*. European Journal of Operational Research, 218(1), 211–229. DOI: 10.1016/j.ejor.2011.09.031

Apéndice A

Tabla 1

Comparación de estudios seleccionados

La tabla completa se encuentra disponible en el siguiente enlace:

Tabla x Comparación de estudios seleccionados

Nota. Elaboración propia basada en plantilla de Chacón Rivera, L. M. (s.f.). Material de clase, Universidad EAN.