Factores que inciden en el “Churn” de una empresa de telecomunicaciones: Un análisis basado en Machine Learning

Carolina Gutiérrez Jaramillo

Director

Magíster Juan Alejandro Trujillo Posada

Tesis para optar al título de Magíster en Gestión Estratégica de la Información

Facultad de Ciencias e Ingeniería

Universidad de Manizales

Octubre de 2023

Resumen

La deserción de clientes -churn- es uno de los problemas estratégicos que deben abordarse de manera urgente por las empresas que prestan servicios en telecomunicaciones. Esto se explica en que retener un cliente existente es mucho más viable a nivel económico que convencer a clientes nuevos de que adquieran los servicios de dichas empresas. Por esto, el objetivo de esta tesis de maestría consiste en establecer los factores que inciden en el "churn" de una empresa que presta servicios de telecomunicaciones que opera en los municipios de Manizales, Villamaría y Chinchiná (Departamento de Caldas). A nivel práctico, esto permitió identificar estrategias viables para retener su cartera de clientes. En este estudio, se aplicaron técnicas de Machine Learning en Python que hace posible predecir las variables relevantes que inciden en el churn de la empresa de telecomunicaciones en los municipios señalados.

Palabras claves: churn, deserción de clientes, machine learning, telecomunicaciones

Abstract

A company's ability to reduce customer churn is one of the most important strategic issues it must address. A large part of this can be attributed to the fact that retaining an existing customer is more economically viable than convincing new customers to purchase the services of such companies. For this reason, this master's thesis aims to establish the factors that influence the churn of a company that provides telecommunications services operating in the municipalities of Manizales, Villamaría, and Chinchiná (Department of Caldas). On a practical level, this makes it possible to identify customer retention strategies. A Python-based Machine Learning techniques were used in this study. Therefore, it was possible to predict the relevant variables that affect the churn of the telecommunications company in the municipalities mentioned above.

Keywords: churn, customer churn, machine learning, telecommunications, customer defection

Tabla de Contenido

Lista de Figuras

Lista de Tablas

INTRODUCCIÒN

La tendencia de los clientes a abandonar el consumo o compra de una marca, lo que se conoce como “churn” en el idioma inglés, suele ser uno de los problemas más complejos que enfrentan las empresas. En particular, esta tasa de abandono o “churn” suele presentar unas tasas mayores en las empresas de telecomunicaciones, a razón de la alta saturación de la oferta de estos servicios en el mercado por la participación de diversas empresas. Por ende, es una problemática que incide de forma importante en el desempeño comercial y, por ende, financiero de las organizaciones de este sector.

El incremento del “churn”, por lo tanto, tiene una relación negativa con los tipos de desempeño mencionados, pues su incremento conlleva el decrecimiento de los ingresos y la rentabilidad de la organización. A partir de esto es que el “churn” se establece como uno de los problemas estratégicos más importantes que enfrentan las empresas del sector telecomunicaciones. En este orden de ideas, las empresas del sector telecomunicaciones deben establecer estrategias para retener a sus clientes y disminuir la tasa de churn.

Aunque las empresas del sector telecomunicaciones mantienen un monitoreo constante de sus tasas de churn, esto no es suficiente, a razón de que tales indicadores se consideran una variable dependiente que es resultado de un número importante de variables o factores que llevan al abandono de los clientes. Por lo tanto, es crucial que las organizaciones identifiquen las variables relevantes que inciden en el churn, de tal forma que puedan diseñar e implementar estratégicas orientadas a la retención de clientes. Esto último es relevante, dado que es mucho más costoso adquirir un cliente nuevo que retener un cliente existente. Por ende, una capacidad estratégica que deben desarrollar las organizaciones del sector telecomunicaciones consiste en la destreza para predecir el churn de acuerdo con las contingencias específicas de su modelo de negocio y de las condiciones del mercado.

Ante el problema planteado, las técnicas en analítica de datos son fundamentales para identificar los factores o variables que inciden en el churn. En la literatura reciente, se ha encontrado que técnicas como XGBoost y los árboles de decisión son utilizadas para desarrollar modelos predictivos que permitan establecer los factores que causan el churn. Estas herramientas permiten generar conocimiento accionable para formular estrategias que reduzcan las tasas de abandono de los clientes en las empresas del sector telecomunicaciones, lo que permitirá mejorar su desempeño tanto comercial como financiero. Por lo anterior, el objetivo de esta tesis consiste en establcer los factores que inciden en el "churn" (abandono de clientes) de una empresa que presta servicios de telecomunicaciones que opera en los municipios de Manizales, Villamaría y Chinchiná (Departamento de Caldas). Para lograr este objetivo se llevó a cabo un proceso de Machine learning, en el programa Python, siguiendo las recomendaciones de la literatura previa sobre CRISP-DM.

Este documento se estructura de la manera siguiente: primero, se plantea el problema de investigación; segundo, se desarrolla todo el componente teórico que sustenta dicho problema; tercero, se establecen los lineamientos metodológicos de la tesis; cuarto, se presentan los resultados principales; y finalmente, estos resultados son discutidos en línea con los objetivos y la pregunta de investigación.

CAPÍTULO 1: PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACIÓN

1.1 Antecedentes

En el contexto de la predicción del Churn en el sector de las telecomunicaciones, se ha llevado a cabo una serie de investigaciones para comprender uno de los desafíos más complejos en este sector, el cual es altamente competitivo: retener a los clientes. En particular, la predicción con éxito del Churn es esencial para el éxito financiero y operacional de las empresas de telecomunicaciones. Ante este reto, diversos estudios han explorado estrategias y técnicas para abordar este problema desde la perspectiva de la analítica de datos. En este apartado se identifican estudios recientes que han abordado esta problemática en el sector telecomunicaciones.

En primer lugar, el estudio de Senyürek y Alp (2023) se destaca como un esfuerzo significativo para construir un modelo que permita entender por qué los suscriptores abandonan los servicios de una empresa de telecomunicaciones. Para lograr este objetivo, de acuerdo con los autores, a nivel práctico-científico, se han empleado técnicas de aprendizaje automático, como la regresión logística, la red neuronal artificial, el bosque aleatorio y el método de aumento. Estas técnicas se aplicaron para estimar a los posibles suscriptores propensos al Churn en una empresa de telecomunicaciones. Al examinar los resultados de esta investigación, se encontró que el método de aumento (boosting) demostró ser más preciso y exitoso en comparación con otras técnicas. En este estudio se encontró que los factores clave que contribuyen al Churn incluyen el período restante hasta la finalización del contrato, la antigüedad del cliente (tenure), la preferencia por el operador elegido por familiares cercanos y la calidad de la red de comunicaciones.

En segundo lugar, Saha et al. (2023) llevaron a cabo un estudio centrado en desarrollar un método de predicción de Churn a partir del Deep Learning. Su objetivo principal era lograr una predicción más precisa del porcentaje de Churn de los clientes sin comprometer las ganancias de la industria de las telecomunicaciones. Para esto, exploraron estrategias de Deep Learning y

probaron una serie de técnicas, desde Ensemble Learning hasta Redes Neuronales Convolucionales. En dos conjuntos de datos, uno del sudeste asiático y otro del mercado estadounidense, encontraron que las Redes Neuronales Convolucionales (CNN) y las Redes Neuronales Artificiales (ANN) arrojaron resultados superiores a otras técnicas. Este estudio demostró que el mejoramiento de la propuesta de valor, junto con una estrategia de bajo costo, inciden en la retención de los usuarios.

En tercer lugar, Brmez and Znidaršic (2019) realizaron un estudio para establecer la predicción de Churn basada en enfoques tradicionales de minería de datos. Su estudio se basó en datos de un operador móvil europeo y resaltó la importancia de detectar patrones a partir de información contractual de clientes, datos de tráfico, facturas, datos de CRM y registros de atención al cliente. Los autores identificaron características importantes para la selección de atributos relevantes en la prevención del Churn.

En cuarto lugar, Amin et al. (2019) desarrollaron un enfoque novedoso para la predicción de Churn en la industria de las telecomunicaciones. Reconocieron el desafío de que los clientes propensos al Churn y los no propensos a menudo presentan características similares. Introdujeron una metodología basada en la certeza del clasificador y el concepto de distancia para dividir el conjunto de datos en zonas con alta y baja certeza, obteniendo resultados prometedores.

En quinto lugar, Dumitrache et al. (2020) destacaron la importancia de los clientes en la estabilidad financiera de la industria de las telecomunicaciones y enfocaron su estudio en la identificación de perfiles de clientes propensos al Churn en Rumania. Utilizaron la técnica Violin Plot para definir a los clientes de prepago con alto riesgo de Churn, caracterizándolos por su inactividad, bajos valores de recarga y opciones adicionales.

En sexto lugar, Haridasan et al. (2023) encontraron en su estudio que la retención de clientes es un desafío en diversos sectores, incluyendo las telecomunicaciones. Su enfoque se centró en el

diseño de un modelo de predicción de Churn utilizando un algoritmo de optimización aritmética y un modelo SBLSTM. Obtuvieron resultados apropiados en términos de precisión y F1-score.

En séptimo lugar, Melian et al. (2022) analizaron el comportamiento de Churn en una muestra de más de 10,000 clientes de una empresa de telecomunicaciones, destacando la importancia de predecir el Churn a través de la minería de datos y la asignación de un "chum score" a cada cliente. Este indicador permite realizar una gestión oportuna del cliente, de tal modo que se pueda evitar su abandono.

En octavo lugar, Saleh y Saha (2023) exploraron los factores que influyen en el Churn en la industria de las telecomunicaciones danesa, enfocándose en la relación entre estos factores y las estrategias de retención. Su investigación se basó en cinco algoritmos de aprendizaje automático y reveló la importancia de la calidad del servicio, la satisfacción del cliente, las actualizaciones de planes de suscripción y la cobertura de la red en referencia a reducir el Churn.

Por último, Khoh et al. (2023) realizaron su estudio alrededor de la importancia de predecir el Churn para reducirlo, también en la industria de telecomunicaciones. Los autores desarrollaron un sistema de predicción de Churn utilizando un enfoque de ensamblaje optimizado que demostró un rendimiento prometedor con una precisión del 84% y una puntuación F1 del 83.42%.

De acuerdo con los antecedentes anteriores, la predicción de Churn en la industria de las telecomunicaciones es una problemática que se ha estudiado de forma recurrente en la investigación reciente. Los estudios revisados muestran que tanto enfoques tradicionales como técnicas de aprendizaje automático avanzado, incluidas las redes neuronales convolucionales, pueden facilitar la predicción exitosa del Churn. Estos estudios contribuyen con un panorama integral y en constante evolución en la búsqueda de soluciones exitosas en la industria de las telecomunicaciones. Es coherente plantear que las tecnologías en análisis de datos ofrecen grandes oportunidades para que las empresas del sector de telecomunicaciones puedan predecir el churn.

1.2 Formulación del problema de investigación

1.2.1 Pregunta de investigación

La pregunta de investigación que guía el desarrollo de esta tesis de maestría es: ¿Cuáles son los factores que inciden en el "churn" (abandono de clientes) de una empresa que presta servicios de telecomunicaciones en los municipios de Manizales, Villamaría y Chinchiná (Departamento de Caldas)?

1.2.2 Objetivo general

El objetivo general de esta investigación consiste en: Establecer los factores que inciden en el "churn" (abandono de clientes) de una empresa que presta servicios de telecomunicaciones que opera en los municipios de Manizales, Villamaría y Chinchiná (Departamento de Caldas).

1.2.3 Objetivos específicos

- Desarrollar un análisis exploratorio para identificar las variables que se relacionan con el "churn"-abandono del servicio de la empresa de telecomunicaciones-.

- Definir un modelo de clasificación de clientes que permita identificar los que son más y menos propensos al "churn".

- Formular acciones o estrategias empresariales para mejorar la retención de los clientes de la empresa de servicios de telecomunicaciones en las ciudades de Manizales, Chinchiná y Villamaría en el departamento de Caldas (Colombia).

1.3 Justificación

El contexto de negocios actual está caracterizado por su complejidad, especialmente al número de actores empresariales que se encuentran en un mismo sector o industria. En esencia, en los mercados se encuentran un gran número de competidores que ofertan los mismos productos y servicios, lo que, a la postre, genera que la oferta sea mayor que la demanda. Esta característica es

lo que se denomina hipercompetencia, pues dichos mercados, gracias a la posibilidad que tienen las empresas de adquirir conocimientos y tecnologías, no presentan barreras de entrada para nuevas empresas.

Ante la problemática planteada, los consumidores, especialmente en el sector de servicios, están expuestos a los estímulos promocionales -especialmente de estrategias de low cost- de diversas empresas, lo que produce la decisión de abandonar su compromiso con una empresa para buscar mayor satisfacción o menor costo en otra. Como resultado, las organizaciones tienen que desarrollar respuestas ágiles que permitan tanto la atracción como la retención tanto de clientes nuevos como de los existentes.

En particular, la identificación de las necesidades de los clientes y el mejoramiento de la solución o propuesta de valor que la empresa les ofrece es esencial para atraer y retener clientes nuevos. Por esto, la analítica de datos juega un papel crucial para comprender dichas necesidades y establecer soluciones de valor que satisfagan los clientes. La analítica es crucial, puesto que hace posible la generación de conocimiento valioso para reformar la propuesta de la empresa y, de este modo, facilita la adaptación de la empresa al mercado. Por esto este trabajo es relevante, pues en la empresa de servicios de telecomunicaciones en la que se desarrolla no se han implementado técnicas de analítica de datos que permitan identificar las causas del abandono de los clientes de la región del suroccidente del departamento de Caldas (Colombia).

Adicional a lo anterior, la predicción del churn de clientes es un fenómeno que ha recibido una amplia atención en la literatura previa. En particular, el big data analytics (Shirazi & Mohammadi, 2019; Zdravevski et al., 2020), el Deep learning (Mishra & Reddy, 2017) y el text analytics (Pustokhina et al., 2021) han demostrado ser estrategias efectivas en establecer las variables que inciden en el churn de clientes. Por lo anterior, la analítica de datos se ha convertido en una herramienta fundamental, desde sus diversas aplicaciones, para predecir el churn de clientes.

Adicionalmente, se ha convertido en parte de la plataforma tecnológica de los departamentos de marketing de las empresas del sector servicios.

Por ende, este trabajo se justifica en que, a nivel de avance tecnológico, existen las herramientas adecuadas para predecir el Churn en la empresa de telecomunicaciones estudiada, la cual requiere de acciones urgentes de retención de clientes ante la pérdida frecuente de éstos y el deterioro de algunos de sus indicadores tanto operacionales como financieros. Por último, este trabajo es novedoso, a razón de que es una iniciativa sin precedentes en la empresa estudiada, desde la perspectiva de la utilización de las técnicas en analítica de datos que se utilizan en esta tesis.

CAPÌTULO 2: REFERENTE TEÓRICO

2.1 Introducción

El estudio del churn, la gestión estratégica de la información y la metodología CRISP-DM, en conjunto, se ha convertido en un área de interés creciente para académicos y profesionales. Estos tres elementos, aunque distintos en naturaleza, se combinan ante la necesidad de las organizaciones de entender, predecir y gestionar la retención de clientes utilizando la información de manera estratégica. Por lo anterior, se consideraron temáticas principales de este marco teórico.

El churn, por su parte, corresponde a un indicador (Key Performance Indicator) crucial para las empresas, puesto que, a menudo, es más costoso adquirir nuevos clientes que retener a los existentes. En el contexto del comercio electrónico (B2C: Business to consumer model), la predicción del churn es esencial para que las empresas formulen medidas eficaces de retención de clientes e implementen estrategias de mercadeo con éxito (Xiahou & Harada, 2022). Ante su relevancia como centro de este estudio, la conceptualización del Churn y sus tipologías también se abarcan en este marco teórico.

En la era de la revolución 4.0, los datos y la información se han convertido en uno de los activos más valiosos para las organizaciones. Así pues, la gestión estratégica de la información implica recolectar, analizar y utilizar la información de manera que apoye la toma de decisiones y la estrategia competitiva de la organización (Lemos et al., 2022). Por esto, para predecir el Churn, en esta tesis se utiliza la técnica del Machine Learning, la cual se explica en este capítulo.

Finalmente, en esta tesis de maestría, la metodología se aborda desde el enfoque CRISP-DM, que significa "Cross-Industry Standard Process for Data Mining". Esta es una metodología ampliamente reconocida para llevar a cabo proyectos de minería de datos y ciencia de datos. Sudharsan y Ganesh (2022), por ejemplo, lo han utilizado para predecir el churn en la industria de las telecomunicaciones. Así pues, también se describe el CRISP-DM dentro del desarrollo conceptual de esta tesis, a continuación.

2.2 Churn: el abandono de clientes

El churn, también conocido como tasa de abandono de clientes, es un fenómeno crítico en la industria de las telecomunicaciones (Xiahou & Harada, 2022). En este sector, el churn se refiere a la pérdida de clientes debido a la elección de otro proveedor de servicios (Grzybowski et al., 2021). Por ende, la retención de clientes y la prevención del churn son desafíos significativos en esta industria, dada la saturación del mercado y la intensa competencia entre operadores (Bugajev et al., 2022), lo cual es un resultado del avance tecnológico de la Industria 4.0.

2.2.1 Churn y la Agrupación de Servicios en Telecomunicaciones

La estrategia de agrupar múltiples servicios en comunicaciones se ha convertido en una estrategia ampliamente utilizada en la industria de las telecomunicaciones (Xiahou & Harada, 2022). Los operadores de telecomunicaciones fijas suelen ofrecer tarifas denominadas "triple-play", que constan de TV, línea fija e Internet de alta velocidad. En los años recientes, también se ha dado la

emergencia del fenómeno del "quadruple-play" en varios países, incluidos Francia, España, Corea del Sur y Japón. Estas ofertas incluyen servicios móviles además de los servicios triple-play (Sudharsan & Ganesh, 2022), lo que ha creado un mercado que experimenta una guerra de precios, en el que la estrategia es ofrecer muchos servicios a bajo costo (low cost strategy).

Respecto a la agrupación de servicios, Grzybowski et al. (2021) analizaron cómo el servicio combinado de lo fijo y lo móvil impacta en el churn de los consumidores. Estos autores utilizaron una base de datos completa de uno de los mayores operadores de servicios en telecomunicaciones de Europa. Particularmente, utilizaron datos de 9,6 millones de suscriptores de servicios de banda ancha fija y 14,2 millones de suscriptores de servicios móviles del mismo operador entre marzo de 2014 y febrero de 2015. De acuerdo con el análisis realizado por Grzybowski et al. (2021), aproximadamente el 8,4% de los suscriptores de servicios de banda ancha fija y alrededor del 11,5% de los consumidores móviles presentaron churn durante el período de un año (entre marzo de 2014 y febrero de 2015). Los resultados del estudio indican que los consumidores que agrupan servicios fijos y móviles del mismo proveedor tienen menos probabilidades de churn. Sin la agrupación de servicios fijos y móviles, el churn de los consumidores de banda ancha fija aumentaría del 8,4% al 9,2%, mientras que el churn de los consumidores en el mercado móvil aumentaría del 11,5% al 13,1%. En efecto, la agrupación de servicios fijos y móviles tiene un impacto más fuerte en la retención de consumidores en el mercado móvil que en el mercado de banda ancha fija. En este sentido, la agrupación es una tendencia del sector, la cual ha permeado la mayoría de las empresas del sector telecomunicaciones en Colombia, la cual tiene como intencionalidad principal la retención de clientes.

2.2.2 Enfoque híbrido para la gestión del churn en telecomunicaciones

A partir de los argumentos anteriores, la gestión del churn es una preocupación primordial para los operadores de telecomunicaciones (Bugajev et al., 2022). Diversas técnicas de aprendizaje supervisado se han utilizado para estudiar el churn de clientes (Grzybowski et al., 2021). Sin embargo, la investigación sobre el uso de técnicas de aprendizaje no supervisado para la predicción del churn es limitada, aunque como se mostró en el primer capítulo, ha tenido un crecimiento importante en los últimos cinco años.

Pejić Bach et al. (2021) desarrollaron en su investigación un enfoque estructurado para la gestión del churn, utilizando un método híbrido que combina el análisis de clústeres y árboles de decisión. Este enfoque se basa en un análisis en tres etapas. Primero, se prepara un conjunto de datos de churn para el análisis, que incluye datos demográficos, tiempo de uso de servicios de telecomunicaciones, contratos y facturación, valor monetario y churn (abandono o no del servicio). En la segunda etapa, se utiliza el análisis de clusters k-means para identificar segmentos de mercado. Finalmente, se emplea el algoritmo de árbol de decisión (CHAID) para desarrollar modelos de clasificación que identifiquen los determinantes del churn en los clústeres en el que se ha identificado su crecimiento. En este caso, las técnicas de identificación del churn con base en clústeres es una tendencia en la literatura especializada sobre este fenómeno en el sector telecomunicaciones.

En otro estudio, Bose y Chen (2009) proponen modelos híbridos que combinan técnicas de clustering no supervisado con árboles de decisión para la predicción del churn. El uso de los clústeres permitió una mejora en la predicción del churn en comparación con los casos en los que no se utilizó dicha técnica. En términos prácticos, esta herramienta es clave en la definición de estrategias que eviten el abandono de clientes. En adición, Łapczyński (2014) predijo la terminación de relaciones en servicios de telecomunicaciones utilizando un modelo híbrido C&RT-

logit. La combinación de árboles de decisión con el modelo logístico, en este caso, enriquecieron la interpretación del modelo y llevaron a una mejor precisión del abandono.

2.2.3 Tipologías de Churn

Existen varios tipos de churn que las empresas deben considerar, los cuales son consistentes según el análisis de la literatura (Bugajev et al., 2022):

Churn Voluntario: ocurre cuando un cliente decide por su cuenta abandonar un servicio o producto. Las razones de esto pueden ser: la insatisfacción con el servicio o producto, la atracción por ofertas de la competencia o cambios en las necesidades o preferencias del cliente (Abou el Kassem et al., 2020).

Churn Involuntario: se refiere a la pérdida de clientes debido a circunstancias fuera del control del cliente. Esto puede incluir situaciones como la pérdida de empleo, reubicación a una zona donde el servicio no está disponible, o, incluso, la muerte del cliente (Arshad et al., 2022).

Churn Pasivo: aunque un cliente puede seguir suscrito a un servicio, es posible que no lo esté utilizando activamente o de forma frecuente. Este tipo de churn es más difícil de detectar porque el cliente todavía está generando ingresos, pero no está obteniendo valor del servicio (Xiahou & Harada, 2022).

Churn Temporal: algunos clientes pueden abandonar un servicio por un período determinado, pero tienen la intención de regresar en el futuro. Este tipo de churn es común en industrias estacionales o en servicios que las personas utilizan por ciclos o modas, como las suscripciones a gimnasios o la realización de cursos por Internet (Sudharsan & Ganesh, 2022).

Así pues, identificar los diferentes tipos de churn es esencial para las empresas, a causa de que cada tipología requiere de un enfoque y estrategia de retención diferente (Arshad et al., 2022). Las empresas deben monitorear y analizar constantemente las razones detrás del churn para

implementar estrategias efectivas de retención y mejorar la experiencia del cliente (Xiahou & Harada, 2022), lo cual es correspondiente con el objetivo específico número tres formulado en el capítulo 1.

2.2.4 Análisis del Churn en las empresas

Analizar el churn es esencial para las empresas por cuatro razones: (1) costo de clientes: adquirir clientes nuevos suele ser más costoso que retener a los actuales. Una alta tasa de churn puede indicar que la empresa está invirtiendo recursos significativos en adquirir clientes que no se quedan a largo plazo (Kaya et al., 2018); (2) reputación y referencias: los clientes que abandonan pueden compartir sus experiencias negativas, lo que puede afectar la reputación de la empresa e influenciar a clientes potenciales (Kurmann et al., 2022); (3) pérdida de ingresos: la pérdida de clientes lleva a una disminución directa de los ingresos y, obviamente, afecta la rentabilidad de la empresa (Matuszelański & Kopczewska, 2022); y (4) comprensión del comportamiento del cliente: analizar las razones detrás del churn puede proporcionar conocimiento de valor sobre las necesidades y preferencias de los clientes, lo que puede guiar a la empresa en la mejora de sus productos o servicios de una forma estratégica y con impacto social (Mustač et al., 2022). Esto último es clave en los procesos de innovación de productos y servicios.

2.3 Gestión Estratégica de la Información (GEI)

La GEI se puede definir como el proceso de adquirir, administrar, interpretar y utilizar la información de manera efectiva para lograr los objetivos estratégicos de una organización (Kurmann et al., 2022). La importancia de la GEI radica en su capacidad para proporcionar a las organizaciones una ventaja competitiva en un mercado globalizado (Ivanova, 2019). La gestión apropiada de la información permite a las empresas anticiparse a las tendencias del mercado, tomar decisiones informadas y responder rápidamente a los cambios. Adicionalmente, en el ámbito de la

seguridad, la gestión estratégica de la información es crucial para proteger a las empresas de amenazas internas y externas, garantizando la confidencialidad y seguridad de los datos (Zatonatskiy, 2019). Junto con lo anterior, la relación entre la toma de decisiones y la evaluación del desempeño es esencial en la GEI. Las decisiones estratégicas deben basarse en datos precisos y confiables para garantizar el éxito de las intervenciones, el control y las acciones organizacionales (Bedilia Estrada-Torres et al., 2018).

2.3.1 La información como recurso estratégico en las organizaciones

En el entorno empresarial actual, la información ha emergido como un recurso estratégico que determina el éxito o fracaso de una organización (Mustač et al., 2022). Las organizaciones del sector público y privado están reconociendo la importancia de gestionar y utilizar eficazmente la información para lograr sus objetivos estratégicos (Haeruddin, 2017; Zatonatskiy, 2019). Con el avance de las tecnologías emergentes, la integración de la infraestructura de sistemas de información se ha vuelto fundamental para lograr flexibilidad estratégica e innovación (Kaya et al., 2018). Las organizaciones que, hoy día, pueden integrar eficazmente sus sistemas de información y adaptar sus prácticas de gestión a las demandas cambiantes, se posicionan mejor para lograr una ventaja competitiva en el mercado (Yoshikuni et al., 2023).

Finalmente, la alineación de las estrategias competitivas con las Tecnologías de la Información es esencial para maximizar el valor de la información como recurso (Gupta, 2019). Las organizaciones que logran esta alineación pueden aprovechar la información para impulsar la innovación abierta, mejorar la eficiencia y fortalecer su posición en el mercado (Dairo et al., 2021; Gupta, 2019; Liu et al., 2019). Dicha alineación se ha consolidado como una capacidad estratégica de las organizaciones.

2.3.2 Toma de decisiones basada en datos e información

Tomar decisiones con base en datos e información es un enfoque que ha ganado relevancia en el ámbito académico y empresarial. En este enfoque, las decisiones no se toman simplemente basándose en la intuición o experiencia, sino que se fundamentan en datos concretos y análisis rigurosos, lo que constituye en una característica de las llamadas empresas *Data-Driven*.

De acuerdo con Cerda y García (2021), durante la pandemia de COVID-19 se requirió información precisa para determinar las variables que afectaban la probabilidad de rechazo o indecisión hacia una vacuna contra el virus, lo cual es esencial para diseñar políticas de salud pública efectivas. En el ámbito educativo, las universidades han comenzado a utilizar técnicas de minería de datos para predecir el rendimiento académico de los aspirantes antes de su admisión (Mengash, 2020). Estas predicciones se basan en criterios previos a su ingreso a la Universidad, como las calificaciones de la escuela secundaria y los resultados de las pruebas estandarizadas por el gobierno.

2.4 Metodología CRISP-DM en el contexto de la retención de clientes en empresas de telecomunicaciones

La industria de las telecomunicaciones experimenta un marco de hipercompetencia, en el que la retención de clientes se ha convertido en un factor crítico para el éxito y la sostenibilidad de las empresas (Yoshikuni et al., 2023). En este contexto, la metodología CRISP-DM se ha posicionado como una herramienta valiosa para abordar los desafíos asociados con la retención de clientes y problemas relacionados con la analítica de datos en las organizaciones (Martínez et al., 2021). Esta se ha utilizado en diversos sectores para mejorar la eficacia de los algoritmos de clasificación en la predicción de respuestas de marketing de clientes. Por lo precedente, a través de un análisis

exhaustivo de datos, se pueden identificar factores esenciales que influyen en la lealtad y la retención de los clientes (Apampa, 2016).

Específicamente, en el sector de los servicios de telecomunicaciones, la gestión de la lealtad del cliente es una capacidad estratégica que las empresas deben desarrollar. La aplicación de técnicas de minería de datos, siguiendo la metodología CRISP-DM, permite a las empresas identificar patrones y tendencias en el comportamiento del cliente (Suryana & Rizki Tri Prasetio, 2021). Por ejemplo, se ha propuesto el uso de técnicas de muestreo como Synthetic Minority Over Sampling, combinadas con técnicas de Boosting, para abordar el desequilibrio de datos en la predicción de la rotación de clientes en telecomunicaciones (Suryana & Rizki Tri Prasetio, 2021). Estas técnicas, cuando se aplican correctamente, pueden mejorar significativamente la precisión de las predicciones y ayudar a las empresas a desarrollar estrategias de retención más efectivas.

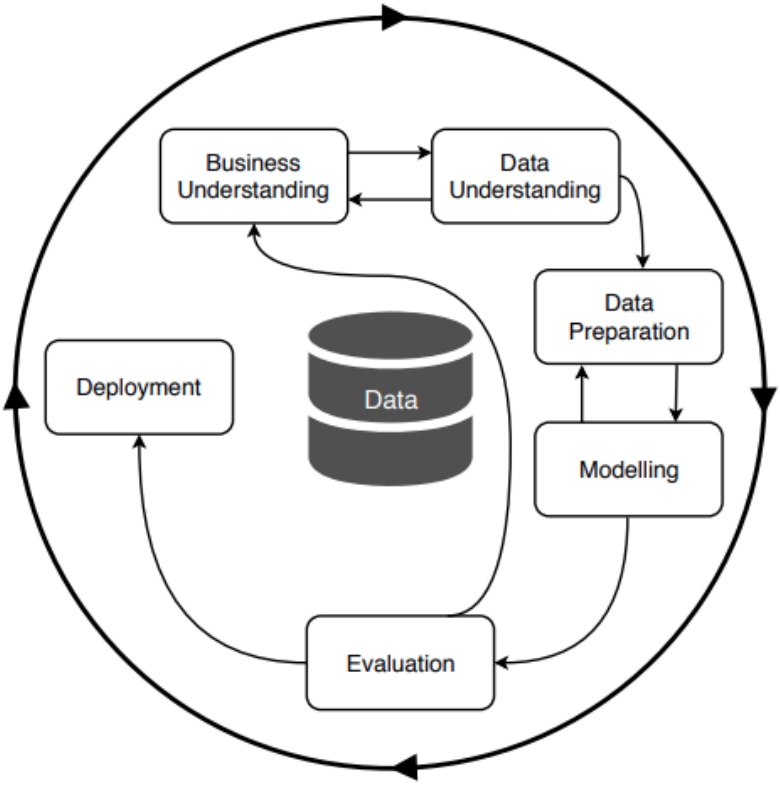
En este orden de ideas, la metodología CRISP-DM ofrece un marco eficaz para abordar los desafíos de la retención de clientes en la industria de las telecomunicaciones (Suryana & Rizki Tri Prasetio, 2021). A través de un enfoque sistemático y basado en datos, las empresas pueden obtener resultados valiosos sobre el comportamiento de sus clientes y desarrollar estrategias efectivas para mejorar la lealtad y reducir la rotación. Por esta razón, es que se utiliza como el referente metodológico en esta tesis.

2.4.1 Fases de la metodología CRISP-DM

La metodología CRISP-DM ha sido reconocida como el estándar general para desarrollar proyectos de minería de datos y generación de conocimiento accionable. A pesar de originarse en la segunda mitad de la década del 90, sigue siendo relevante y ampliamente adoptada en la comunidad de ciencia de datos (Martínez et al., 2021; Suryana & Rizki Tri Prasetio, 2021). CRISP-MD suele implementarse en seis fases de interacción iterativa durante el ciclo de vida de los proyectos de

minería de datos, las cuales son: (1) comprensión del problema o negocio; (2) comprensión de datos; (3) preparación de datos; (4) modelado; (5) evaluación del modelo; y (6) implementación del modelo (ver Figura 1).

Figura 1. Fases del CRISP-MD



Fuente: Martínez et al. (2021).

Cada una de las fases del CRISP-DM se describe, a continuación, según los planteamientos de Martínez et al. (2021) y Wiemer et al. (2019):

Entendimiento del negocio: implica comprender el problema a resolver, determinar los objetivos y requisitos del proyecto y, por último, traducirlos en objetivos técnicos. Esta fase es necesaria para establecer el alcance y las expectativas del proyecto. Suele sustentarse en una pregunta de investigación o de negocio.

Entendimiento de los datos: engloba la recolección inicial de datos para familiarizarse con el problema. Las tareas incluyen la recolección, descripción, exploración y verificación de la calidad de los datos al principio del proyecto.

Preparación de los datos: los datos se adaptan y transforman de acuerdo con la técnica de análisis de datos seleccionada. Por ende, se visualizan los datos y se buscan relaciones entre las variables.

Modelamiento: se selecciona un modelo específico adecuado para el problema en cuestión. Los datos disponibles, los requisitos del problema y el conocimiento del dominio deben ser considerados.

Evaluación del modelo: se evalúa el modelo seleccionado interpretando los resultados, por lo que se utilizan diferentes herramientas para determinar si se cumplieron los criterios de éxito.

Implementación del modelo: una vez que el modelo ha sido construido y validado, el conocimiento adquirido se traduce en acciones (conocimiento accionable). Se planifica la implementación, se realiza el mantenimiento y monitoreo, y se revisa el proyecto.

2.5 Machine learning

El Machine Learning es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos que permiten a las máquinas aprender a partir de datos (Yoshikuni et al., 2023). Estos algoritmos buscan identificar patrones y relaciones en los datos para hacer predicciones o tomar decisiones sin ser explícitamente programados para ello (Dairo et al., 2021). Ahuja et al. (2019) plantea los objetivos siguientes del Machine Learning: (1) descubrimiento de patrones: consiste en identificar patrones y relaciones ocultas en grandes conjuntos de datos; predicción: se utilizan datos históricos para predecir eventos o resultados futuros; optimización: se mejoran procesos o

decisiones basándose en datos; y automatización: se hace posible que las máquinas realicen tareas sin intervención humana directa.

2.5.1 Tipos de Análisis y Modelos de Maching Learning

Los principales tipos de aprendizaje, relacionados con el análisis basado en el Maching Learning, se describen a continuación (Ahuja et al., 2019):

Aprendizaje supervisado: en este enfoque se proporciona a la máquina un conjunto de datos etiquetado, es decir, cada ejemplo de datos viene con la respuesta correcta. El objetivo es que, después de entrenarse con este conjunto de datos, la máquina pueda predecir respuestas correctas para nuevos datos que no haya visto antes.

Aprendizaje no Supervisado: a diferencia del aprendizaje supervisado, en este enfoque no se proporcionan etiquetas. En su lugar, el algoritmo intenta identificar patrones o estructuras en los datos por sí mismo. Un ejemplo común es la agrupación o "clustering", en el que el objetivo es dividir un conjunto de datos en grupos de elementos similares (Ahuja et al., 2019).

Aprendizaje por refuerzo: en este tipo de aprendizaje, un agente toma decisiones interactuando con un entorno. El agente recibe recompensas (positivas o negativas) basadas en las acciones que realiza. Su objetivo es maximizar la recompensa acumulada a lo largo del tiempo.

Además de estos tipos de análisis, los siguientes son los modelos de Machine Learning de acuerdo con Ahuja et al. (2019): (1) clasificación: predice la categoría a la que pertenece una entrada. Ejemplos de técnicas de clasificación incluyen la regresión logística, los árboles de decisión, K-vecinos más Cercanos y Naive Bayes (Ahuja et al., 2019); (2) regresión: predice valores continuos basándose en variables independientes.; (3) clusterización: agrupa datos similares en conjuntos. Las técnicas comunes incluyen K-medias y clustering Jerárquico (Ahuja et

al., 2019); y (4) Redes Neuronales: estos son modelos inspirados en el cerebro humano, especialmente útiles para tareas complejas como el reconocimiento de imágenes.

2.5.2 Algoritmos de Machine Learning para el análisis de Churn

Los principales algoritmos de Machine Learning usados en procesos de retención de clientes en telecomunicaciones están relacionados con tareas de clasificación. Estos son:

Árboles de Decisión (Decision Tree): los árboles de decisión son herramientas muy eficaces en el campo del aprendizaje automático y la inteligencia artificial, y son utilizadas para descubrir reglas y relaciones en conjuntos de datos (Yoshikuni et al., 2023). Estas estructuras jerárquicas dividen sistemáticamente la información de la base de datos en función de ciertos criterios, permitiendo la toma de decisiones según características específicas de los datos. La construcción de un árbol de decisión se basa en los valores de las variables predictoras para dividir el conjunto de datos en subconjuntos. Existen diferentes algoritmos para la construcción de árboles de decisión, entre los que se destacan: (1) CART (Classification and Regression Tree): este algoritmo divide el conjunto de datos en dos subconjuntos basándose en los valores de las variables predictoras. Es ampliamente utilizado para tareas de clasificación y regresión (Jaworski et al., 2018); (2) CHAID (Chi Squared Automatic Interaction Detector): a diferencia de CART, CHAID puede dividir el conjunto de datos en más de dos subconjuntos. Es especialmente útil cuando se desea dividir una población en diferentes segmentos basándose en un criterio de decisión específico (Lin & Fan, 2019).

Por último, el proceso de partición de subconjuntos continúa aplicando el mismo algoritmo hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción en uno de estos grupos hacia el valor de la variable de respuesta. El resultado es un árbol cuya raíz

representa el conjunto de datos íntegro, las ramas representan los conjuntos y subconjuntos, y cada conjunto donde se realiza una partición que se denomina nodo.

Random Forest. Los algoritmos de Random Forest son una técnica de aprendizaje automático que pertenece a la categoría de métodos de aprendizaje en conjunto (Yoshikuni et al., 2023). Estos métodos han logrado un rendimiento de vanguardia en diversas aplicaciones de aprendizaje automático al combinar las predicciones de dos o más modelos base. La idea fundamental detrás del aprendizaje en conjunto es el reconocimiento de que los modelos de aprendizaje automático tienen limitaciones y pueden cometer errores. Por lo tanto, el aprendizaje en conjunto busca mejorar el rendimiento de la clasificación aprovechando las fortalezas de múltiples modelos base.

Los métodos de aprendizaje en conjunto se dividen en tres categorías: boosting, bagging y stacking. Entre ellos, el Random Forest es un algoritmo bien conocido que pertenece a la categoría de bagging. La idea principal detrás de Random Forest es entrenar múltiples árboles de decisión y combinar sus predicciones para obtener un rendimiento mejorado y una mejor capacidad de generalización que los árboles individuales (Yoshikuni et al., 2023). El desempeño superior de los métodos de aprendizaje en conjunto, como Random Forest, se debe en parte a su capacidad para limitar los errores de varianza y sesgo asociados con modelos de aprendizaje automático individuales. Por ejemplo, mientras que el bagging reduce la varianza sin aumentar el sesgo, el boosting se centra en reducir el sesgo (Yoshikuni et al., 2023).

En la implementación de clasificadores en conjunto, la precisión y diversidad de los aprendices base son dos factores esenciales que deben considerarse. La mayoría de los algoritmos de conjunto aseguran la diversidad a través del uso de un proceso de muestreo iterativo de datos o alterando la estructura de los aprendices individuales. Además, es deseable que los aprendices base aseguren una precisión superior al adivinar al azar (Mienye & Sun, 2022)

2.6 Factores que inciden en el Churn desde la perspectiva de la información en empresas de Telecomunicaciones

Diversos factores pueden influir en el churn, por lo que es esencial comprenderlos estos para desarrollar estrategias efectivas de retención. Adicionalmente, guían el diseño de los modelos de análisis que sugiere el CRISP-DM. Los factores que tienen mayor incidencia son:

Problemas técnicos: las dificultades técnicas, como fallos en el sistema o problemas con la interfaz de usuario, pueden llevar a una disminución de la satisfacción del cliente y, en última instancia, al churn. Las empresas de telecomunicaciones deben abordar estos problemas técnicos para garantizar una experiencia de usuario fluida y sin problemas.

Influencia Social: La percepción de los usuarios sobre la calidad del servicio y las opiniones de sus pares pueden influir en su decisión de permanecer o abandonar un servicio de telecomunicaciones (Ghode et al., 2020). Las empresas deben ser conscientes de la influencia social y trabajar para mantener una imagen positiva en la mente de los consumidores (Wang et al., 2017).

Expectativas de desempeño: las expectativas de rendimiento, es decir, el grado en que un individuo cree que el uso de un servicio mejorará su experiencia, pueden influir en la decisión de continuar utilizando un servicio de telecomunicaciones (Lange, et al., 2016). Si un servicio no cumple con las expectativas de rendimiento, es más probable que los usuarios abandonen el servicio.

Políticas de precios. los cambios de precios en los servicios, sin una justificación en el mejoramiento de valor, puede llevar al churn. Por ende, las organizaciones deben analizar la relación entre su política de precios y el abandono de los clientes (Alqahtani & Rajkhan, 2020).

CAPÍTULO 3: METODOLOGÍA

3.1 Alcance metodológico

Este proyecto de investigación se enfoca en un análisis descriptivo y predictivo en el contexto de la retención de clientes de una empresa de telecomunicaciones que opera en el suroccidente del departamento de Caldas, en Colombia. La investigación es descriptiva, dado que se busca presentar un panorama detallado de los factores que inciden en la retención de clientes, describiendo las variables y patrones observados en los datos históricos de la empresa (Maskew et al., 2022). Además, es predictiva, a razón de que se emplearán algoritmos de Machine Learning para modelar el comportamiento de retención de los clientes basándose en las variables identificadas. Esto último permitirá la formulación de las estrategias de retención.

3.1.1 Análisis Descriptivo de Factores de Retención

En esta tesis se realizó, en primer lugar, un análisis exploratorio de los datos para identificar variables que puedan estar asociadas con la retención de clientes, tales como: demografía, uso del servicio y comportamiento de pago. Este análisis permitió comprender las características de los clientes que permanecen y también las de los que optan por abandonar los servicios de la empresa de telecomunicaciones.

3.1.2 Modelado Predictivo para Identificación de Factores de Churn

En segundo lugar, se emplearon algoritmos de Machine Learning. En específico, se implementaron Árboles de Decisión (Decision Tree), Random Forest, Regresión Logística y XGBoost, para desarrollar un modelo predictivo que identifica a los clientes con mayor propensión al churn. En el contexto de la retención de clientes, los árboles de decisión son una técnica de aprendizaje supervisado que se utiliza para clasificar los de acuerdo con sus características. Así, un árbol de decisión puede ayudar a segmentar la base de clientes en diferentes grupos según sus patrones de

comportamiento, ingresos, hábitos de pago y características demográficas. Esta segmentación permite a las empresas de servicios en telecomunicaciones diseñar estrategias de retención específicas para cada segmento, maximizando así la eficacia de sus esfuerzos de retención y la rentabilidad de cada tipo de cliente (Dahiya & Bhatia, 2015). El Random Forest es una técnica de clasificación que ha demostrado ser eficiente en la gestión de grandes cantidades de datos y en la producción de modelos con menos errores, siendo ampliamente utilizado en investigaciones similares para predecir y clasificar datos (Basha et al., 2020).

Lo anterior conecta con la utilización de la técnica de La regresión logística es una técnica de modelado estadístico que se utiliza para predecir la probabilidad de un resultado binario (1 / 0, Sí / No, Verdadero / Falso, Éxito/Fracaso) dadas un conjunto de variables independientes. En el contexto de la retención de clientes, la regresión logística ayuda a predecir la probabilidad de que un cliente abandone (churn) o permanezca con la empresa. Esta técnica es especialmente útil cuando se trata de identificar las características y comportamientos de los clientes que pueden influir en su decisión de permanecer o abandonar un servicio (Sharma et al., 2022). Por esta razón, fue el sustento del Machine Learning presentado en este estudio.

3.1.3 Estrategias de Retención de Clientes

Con base en los factores identificados y los resultados del modelo predictivo, se proponen en el capítulo 5 de esta tesis de maestría estrategias específicas para mejorar la retención de clientes en la empresa de telecomunicaciones estudiada, enfocándose en los segmentos de clientes identificados con alto riesgo de Churn.

3.2 Tipo de estudio

Esta investigación se desarrolla desde el paradigma cuantitativo de tipo correlacional, dado que busca comprender y determinar la interacción-incidencia entre las diversas variables y grupos de

variables que intervienen en el fenómeno de la retención de clientes en la empresa de telecomunicaciones estudiada. Asimismo, es de tipo explicativo, porque tiene como objetivo identificar las variables que inciden en que los clientes abandonen el servicio de telecomunicaciones (Banabo & Ndiomu, 2023).

Este estudio, en particular, permitió la identificación y clasificación de las relaciones entre variables, así como determinar en qué condiciones se manifiesta la propensión de los clientes a abandonar el servicio. Esta información es esencial para que el departamento de atención al cliente y fidelización de la empresa de telecomunicaciones pueda activar sus protocolos de retención de manera más eficiente y oportuna. Además, permite desarrollar estrategias adicionales basadas en las indicaciones proporcionadas por el modelo diseñado, con el objetivo de mejorar la lealtad y satisfacción del cliente. Es importante destacar que la experiencia del cliente y la calidad del servicio son factores cruciales en la retención de clientes en el sector de telecomunicaciones (Banabo & Ndiomu, 2023).

3.3 Diseño de investigación

Esta investigación se desarrolló con base en la metodología CRISP-DM, la cual es específica para proyectos de minería de datos, explicada de manera general en el capítulo dos de esta tesis (Referente teórico). A partir de esta se desarrollaron tres fases: análisis exploratorio de los datos; generación del modelo de Machine Learning clasificatorio; y formulación de acciones o estrategias empresariales para mejorar la retención de los clientes de la empresa de servicios de telecomunicaciones.

3.3.1 Fase 1: Análisis exploratorio de los datos

Entendimiento del negocio. Se inició con la comprensión de las políticas y protocolos que sigue la empresa de telecomunicaciones para mitigar el fenómeno de la pérdida de clientes o "churn". Es

esencial analizar los datos históricos relacionados con la retención y abandono de clientes. Una primera fuente de información fue el Sistema de Gestión de Clientes (SGC) de la empresa, en el que se consolidan datos de cada cliente, procedentes de las áreas de ventas, atención al cliente, facturación y soporte técnico. Sin embargo, este sistema, en muchos casos, solo proporciona análisis descriptivos del fenómeno. Una segunda fuente fue el sistema de información de la Comisión de Regulación de Comunicaciones (CRC), que consolida información de diversas empresas de telecomunicaciones. Aunque este sistema ofrece una visión más amplia del sector, su enfoque es principalmente descriptivo. Finalmente, está el "Programa de Fidelización y Retención de Clientes" de la empresa de telecomunicaciones, que contiene las políticas, orientaciones y estrategias que se implementan para enfrentar y minimizar el "churn". Es evidente que el departamento encargado de la retención de clientes tiene un conocimiento profundo del fenómeno y necesita herramientas que emitan alertas tempranas para identificar y actuar sobre los clientes en riesgo de abandonar el servicio.

Entendimiento de los datos. En esta fase se realizó la la recolección de datos a partir de los sistemas de información y otras fuentes de datos de la empresa de telecomunicaciones, comprendiendo su naturaleza y significado, y analizando su calidad para reconocer su validez. Para esto, se llevaron a cabo de manera ordenada las siguientes actividades: (1) recepción de los archivos de datos en formato de texto (.csv). Los archivos corresponden a: Historial de uso del servicio, registros de facturación, interacciones de atención al cliente, y registros de promociones y ofertas aceptadas por el cliente; (2): conversión de los archivos de texto (.csv) a archivos en Excel; y (3) almacenamiento de los archivos en Excel (.xls) en las tablas de las bases de datos stage BD_RETENCION_STG. Se importan sin ninguna transformación. Este proceso técnicamente se conoce como ELT (extracción, carga y transformación de la data); (4) descripción de los datos Al revisar la totalidad de los datos recibidos como fuente, se determina la categoría de cada una de las

variables, ejecutando validaciones de calidad sobre ellas y aplicando algunas reglas de ajuste para adaptarlas a los significados apropiados. Los datos residen en SQL Server en la base de datos BD_RETENCION_MRL, mediante conteos, valores únicos, valores comunes, agrupamiento de valores, etc.

Para las variables continuas, se construyeron medidas de tendencia central (media, mediana, moda, desviación estándar y distribución por cuartiles). Los análisis exploratorios se enfocaron en análisis univariado, bivariado y multivariado.

3.3.2 Fase 2: Implementación de modelos clasificatorios que permitan explicar el churn

Durante esta etapa del proceso investigativo se busca que los datos se ajusten a los requerimientos técnicos del modelado de preparación para los algoritmos de decisión tree random forest y regresión logística, con acciones como agrupar variables existentes, seleccionar las mejores variables para el modelo (feature importance) y derivar nuevas variables en función de las existentes.

Modelado. En esta etapa del proceso se prepara y configuran los algoritmos de decisión tree random forest y regresión logística que van a emplearse para la clasificación de la retención de clientes. El algoritmo toma las variables previamente seleccionadas como de mayor importancia para la explicación del fenómeno (feature importance). Posterior a ello, se construyen las variables X (variable a predecir: desertor o no desertor), y la variable Y (conjunto de variables que explican el fenómeno). Establecidas plenamente X e Y, se seleccionan el conjunto de datos para entrenamiento, equivalente al 70% de la data, y el conjunto de datos test, equivalente al 30% de la data.

Evaluación. La base fundamental de la evaluación de resultados se dirige por la matriz de confusión, la cual muestra la cantidad de falsos positivos, falsos negativos, verdaderos positivos y

verdaderos negativos de la predicción. Posteriormente, se chequean los resultados ofrecidos por la sensibilidad, exactitud, precisión, especificidad y el área bajo la curva y se hace una comparación de estas métricas entre los 3 modelos.

3.3.3 Fase 3: Formulación de acciones o estrategias empresariales para mejorar la retención de los clientes

Dentro de la metodología de esta investigación, una vez obtenidos y analizados los resultados, se procedió con la formulación de estrategias empresariales basadas en los hallazgos. Esta fase es esencial para traducir los conocimientos derivados de la analítica analíticos en acciones concretas que la empresa de telecomunicaciones pueda implementar. Se procedió de la forma siguiente:

- Identificación de Factores Críticos: con base los resultados, se identificaron las variables y factores con una correlación significativa con el abandono de clientes (Xiahou et al., 2022). Estos factores proporcionaron una comprensión clara de las áreas y variables de intervención prioritarias.
- Segmentación de Clientes: a partir de los patrones identificados, se segmentaron los clientes en diferentes grupos según su propensión al abandono o Churn (De Caigny et al., 2021).
- Formulación de Estrategias: con la información obtenida, se establecieron estrategias dirigidas a abordar las principales causas de abandono identificadas. Estas estrategias se diseñaron para ser implementadas por la empresa de telecomunicaciones con el objetivo de mejorar la retención de clientes (Wu et al., 2021).
- Propuesta de Acciones Personalizadas: basándose en la segmentación, se propusieron acciones específicas para cada grupo de clientes, que incluyen comunicaciones personalizadas, ofertas especiales, o programas de lealtad (De Caigny et al., 2021).

Esta fase metodológica aseguró que los resultados de la investigación no solo proporcionen conocimientos accionables valiosos, sino que también guíen a la empresa en la implementación de

acciones concretas para mejorar la retención de clientes y, en efecto, su desempeño tanto operacional como estratégico.

CAPÍTULO 4: RESULTADOS

4.1 Fase 1: Análisis exploratorio de los datos

4.1.1 Entendimiento del Negocio

En el sector de las telecomunicaciones, la retención de clientes es esencial para garantizar la sostenibilidad y el crecimiento rentable. Las empresas que no logran retener a sus clientes enfrentan costos elevados de adquisición y una disminución en la lealtad de la marca. Las empresas de telecomunicaciones se han enfocado en ofrecer soluciones de comunicación de alta calidad y servicios múltiples, con el objetivo principal de satisfacer y retener a sus clientes. Para abordar el desafío de la retención, la empresa estudiada, a partir del trabajo desarrollado en esta tesis, ha decidido explorar técnicas avanzadas de análisis de datos.

En específico, el departamento de estrategia y retención, a partir del direccionamiento del *Top Management,* ha emprendido un proyecto con el propósito de identificar y clasificar las principales razones detrás del retiro de los clientes. Así pues, utilizando técnicas de análisis exploratorio de datos, el equipo descubrió patrones y tendencias que indican las posibles causas de churn de los clientes de la empresa de telecomunicaciones que opera en el suroccidente de Caldas. Adicionalmente, en la organización se evalúan de forma constante modelos de machine learning, como decision tree, random forest y regresión logística, para predecir y clasificar los posibles retiros de clientes. Estos modelos permiten anticipar posibles deserciones y diseñar estrategias proactivas de retención.

En consecuencia, a partir de entrevistas, sesiones y análisis de datos, , se identificaron varios puntos clave para la estrategia: (1) los datos para las predicciones provienen de diversas fuentes,

incluyendo bases de datos de CRM, retroalimentación de clientes y registros de interacciones de servicio al cliente; (2) se han utilizado herramientas como Python y SQL para la extracción de datos, el análisis exploratorio y la implementación de los modelos de clasificación; y (3) se llevó a cabo la contrastación entre los algoritmos decision tree, random forest y regresión logística para determinar cuál ofrece la mejor precisión y adaptabilidad al contexto de la empresa.

Determinación de los Objetivos Institucionales. Para garantizar que este estudio esté alineado con las metas y las necesidades de la empresa de telecomunicaciones, fue esencial definir los objetivos organizacionales. Estos objetivos se establecieron tanto a nivel general como específico, con el propósito de contextualizar la investigación en un marco práctico y orientado a la estrategia competitiva organizacional. De esta manera, se busca trascender el ámbito puramente académico y enfocar el estudio hacia resultados tangibles y aplicables para la empresa. Por lo tanto, los objetivos empresariales en los que se enfoca este estudio son:

- Reducir la tasa de retiro (churn) de clientes identificando y actuando sobre las variables que más impactan esta métrica.
- Comprender las razones subyacentes detrás del retiro de clientes para diseñar y aplicar estrategias efectivas que permitan minimizarlo en el futuro.

Evaluación de la Situación. Tras un análisis exhaustivo del entorno en el que se llevó a cabo la investigación sobre el churn en la empresa de telecomunicaciones, se identificaron varios elementos clave que influirían en el desarrollo y enfoque del estudio:

- Fuente de Datos: El conjunto de datos para la investigación proviene exclusivamente de los registros internos de la empresa de telecomunicaciones. Esto significa que cualquier limitación o sesgo presente en estos datos afectará directamente los resultados de la investigación.

- Herramientas Utilizadas: Se optó por herramientas de código abierto, específicamente Python, Jupyter y Google Colab. Estas herramientas fueron esenciales durante las fases de Conocimiento de los Datos, Preparación de los Datos, Modelado y Evaluación.

- Idioma de la Documentación: Gran parte de la documentación científica y las librerías de código abierto relevantes para la investigación estaban en inglés. Esto podría presentar desafíos en términos de interpretación y aplicación.

- Alcance de la Investigación: El estudio se limitó a la fase de Evaluación. Aunque se identificaron posibles pasos para la implementación, estos se considerarán para investigaciones futuras.

- Tiempo Estipulado: Se estableció un marco temporal de 12 meses para completar la investigación. Este límite de tiempo influiría en la profundidad y alcance del estudio.

Determinar Objetivos de la Minería de Datos. Los objetivos de la minería de datos se establecieron para guiar el enfoque técnico y metodológico del proyecto. Estos objetivos también se alinearon con el objetivo general y los objetivos específicos de este estudio de maestría: (1) Descripción de Clientes con Churn: uno de los principales objetivos es describir y entender las características distintivas de los clientes que optan por abandonar los servicios de la empresa. Esto ayudará a identificar patrones y factores comunes entre estos clientes; (2) Clasificación de Clientes según Churn: con base en los datos y características identificadas, el objetivo es clasificar a los clientes según su probabilidad de abandonar los servicios. Esta clasificación permitirá a la empresa tomar medidas preventivas y adaptar sus estrategias de retención.

4.1.2 Entendimiento de los datos

Para emprender el desafío de la retención de clientes en la empresa de telecomunicaciones, es esencial comprender en profundidad la información disponible. La base de datos, construida y gestionada utilizando SQL Server, ofrece una visión detallada de los servicios proporcionados a

los clientes y su comportamiento a lo largo del tiempo. SQL Server ha permitido estructurar, consultar y manipular los datos de manera eficiente. Su capacidad para manejar grandes volúmenes de datos y realizar consultas complejas ha sido fundamental para preparar y analizar la información relevante para este estudio. A continuación, se presenta la lista de variables y la descripción detallada de las mismas:

- PERIODO: proporciona una referencia temporal específica (año y mes) para cada registro. Es esencial para analizar tendencias y patrones a lo largo del tiempo, y para identificar posibles estacionalidades en el comportamiento de retiro de los clientes.
- ID_SERVICIO: la identificación única por servicio permite su rastreo y análisis de forma individual, lo que es crucial para un análisis detallado.
- PRODUCTO: al categorizar los servicios en Banda Ancha, Televisión y Telefonía, se identifican qué productos son más propensos al churn y adaptar estrategias específicas para cada uno.
- VELOCIDAD: específicamente para el servicio de Banda Ancha, la velocidad de navegación puede influir en la satisfacción del cliente y, por ende, en su decisión de permanecer o retirarse.
- MUNICIPIO: al conocer la ubicación geográfica del servicio, se especifican áreas con mayores tasas de churn y es posible correlacionar con factores geográficos o de infraestructura.
- ESTRATO: esta clasificación socioeconómica es vital para entender las capacidades y expectativas de pago de los clientes, y adaptar los servicios y tarifas según las necesidades de cada estrato.
- TECNOLOGIA: la infraestructura tecnológica puede influir en la calidad del servicio, lo que a su vez puede afectar las tasas de retención.

- TARIFA: al analizar las tarifas en relación con el churn, se pueden obtener insights sobre la percepción de valor del cliente.
- RETIRO: esta es una columna crucial que indica si un cliente ha abandonado o no, y es la variable objetivo para los modelos de predicción.
- PQR: los procesos de atención al cliente y las quejas pueden ser indicadores tempranos de insatisfacción y posibles churns.
- AGRUP_ANTIGUEDAD y ANTIGÜEDAD: la lealtad y la duración de la relación con el cliente pueden influir en la probabilidad de churn.
- PLAN: al entender qué planes tienen mayores tasas de churn, se pueden adaptar o mejorar esos planes específicos.
- AÑO y MES: estas columnas adicionales facilitan el análisis temporal y la identificación de tendencias estacionales.

Esta es una visión general de los datos, pero no quiere decir que todos sean incluidos en los modelos. Simplemente es un inventario de los datos a los que se puede acceder en la organización, en términos de su relación con el problema.

Con esta comprensión detallada de los datos y con el soporte de SQL Server, se abordó entonces el desafío de predecir y prevenir el churn de clientes en la empresa de telecomunicaciones que opera en el suroccidente de caldas. El siguiente paso consistió en realizar un análisis exploratorio de datos para identificar patrones, correlaciones e insights que guíen la modelización y las estrategias de retención. A continuación, se presenta un resumen de las variables cualitativas en la Tabla 1.

Tabla 1. Resumen de Variables Cualitativas

Variable	Count	Unique	Top	Freq
PRODUCTO	18571	3	BA	8246
MUNICIPIO	18571	3	MANIZALES	15378
TECNOLOGIA	18571	2	HFC	17947
AGRUP_ANTIGUEDAD	18571	5	>36 MESES	8623
PLAN	18571	25	ILIMITADO	5232

Fuente: empresa de Telecomunicaciones.

Dentro de la base de datos de la empresa de telecomunicaciones, que cuenta con un total de 18,571 registros, se destaca que la Banda Ancha (BA) es el producto más solicitado por los clientes, con 8,246 registros. La mayoría de estos servicios se concentran en el municipio de Manizales, integrando 15,378 de los registros totales. En cuanto a la infraestructura tecnológica, la tecnología HFC (Hybrid Fiber-Coaxial) es la dominante, abarcando 17,947 de los registros, lo que sugiere una fuerte inclinación o ventaja técnica hacia esta tecnología. Además, una significativa proporción de clientes, específicamente 8,623, ha estado suscrita a los servicios de la empresa por más de 36 meses, indicando una alta retención. Finalmente, el plan "Ilimitado" es el preferido entre los usuarios, con 5,232 registros, señalando una tendencia hacia servicios sin restricciones. Siguiendo con las variables cuantitativas, estas se presentan en la Tabla 2.

Tabla 2. Resumen de variables cuantitativas

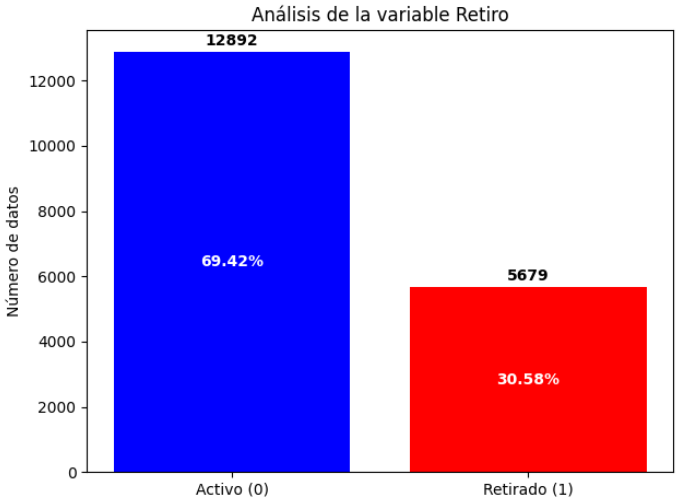
Variable	count	mean	std	min	25%	50%	75%	max
VELOCIDAD	18571.0	4,84E+07	75,11634	0.0	0.0	0.0	60.0	300.0
ESTRATO	18571.0	3,06E+06	1,11607	1.0	2.0	3.0	3.0	6.0
TARIFA	18571.0	4,76E+10	29.645,13710	-1.0	16807	46017	69900.0	230000.0
PQR	18571.0	2,14E+04	0,14464	0.0	0.0	0.0	0.0	1.0
ANTIGUEDAD	18571.0	5,59E+07	60,09572	2.0	22.0	34.0	69.0	405.0

Fuente: empresa de Telecomunicaciones.

Dentro de la base de datos de 18,571 registros, la velocidad promedio de conexión es de aproximadamente 48.4 Mbps, aunque la mayoría de los usuarios no reportan velocidad, con un 75% de ellos registrando 60 Mbps y algunos alcanzando hasta 300 Mbps. En cuanto al estrato, la media es 3, con un rango que va desde el estrato 1 hasta el 6. La tarifa promedio que los clientes pagan es de aproximadamente $47.600, con una tarifa mínima reportada de $-1 y una máxima de $230,000. Es notable que haya valores negativos, lo que podría indicar errores o devoluciones. En referencia a las PQRs, la mayoría de los clientes, representando el 75%, no ha presentado quejas o reclamos, pero hay registros que sí lo han hecho. Finalmente, la antigüedad promedio de los clientes es de 55.9 meses, con un cliente que ha estado suscrito durante 405 meses, lo que indica una relación a largo plazo con la empresa.

La descripción cuantitativa de la variable retiro, según porcentajes, se presenta en la Figura 2.

Figura 2. Comportamiento de la variable retiro

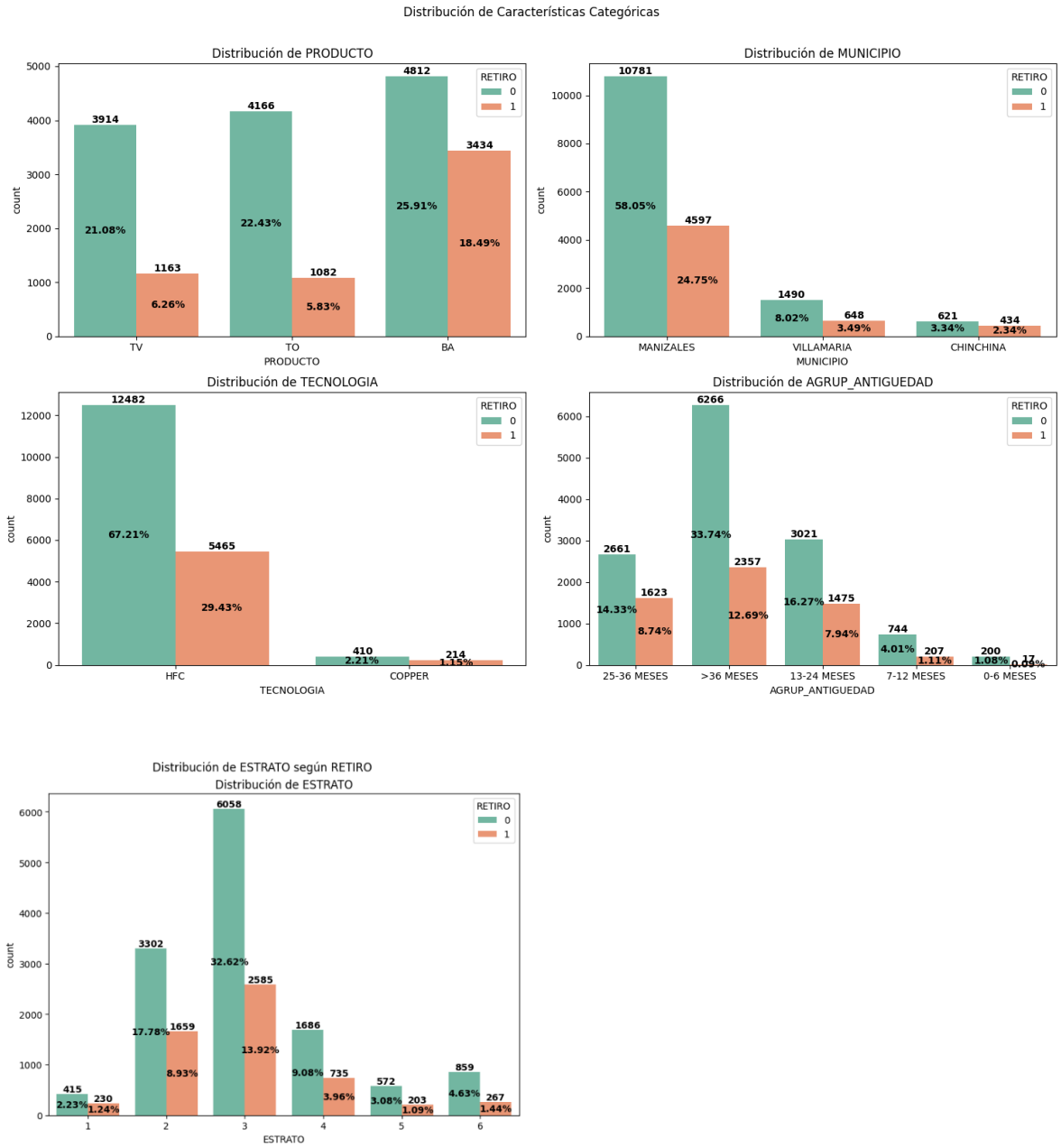


Fuente: empresa de Telecomunicaciones.

La variable de respuesta "Retiro" muestra que una mayoría significativa de los datos, específicamente el 69.42%, corresponde a clientes que aún están activos (etiquetados como 0). Por otro lado, el 30.58% de los datos representa a clientes que se han retirado (etiquetados como 1). Esta distribución sugiere que, aunque la empresa ha logrado retener a una buena proporción de sus clientes, todavía existe un porcentaje considerable que ha decidido retirarse, lo que podría indicar áreas de mejora en términos de retención de clientes.

En la Figura 3 se presenta la distribución de los datos de los clientes de acuerdo con las características categóricas.

Figura 3. Distribución de características categóricas



Fuente: empresa de Telecomunicaciones.

De acuerdo con la Figura 3, el análisis de la gráfica revela patrones notables en la distribución de la variable retiro en una empresa de telecomunicaciones. En términos de producto, el 'BA' registra la tasa de retiro más elevada con un 25.91%, seguido de 'TO' y 'TV'. Al observar la

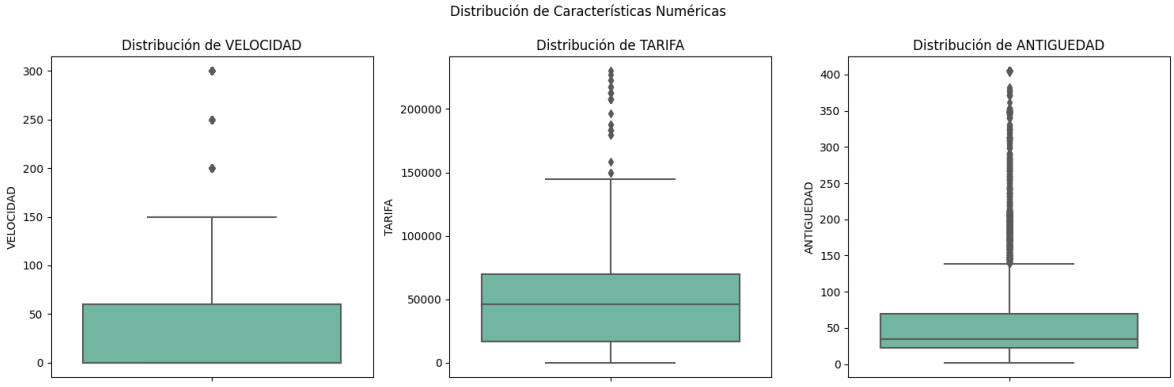
tecnología, 'HFC' presenta una tasa de retiro del 29.43%, notablemente alta considerando su amplia base de clientes.

En cuanto a la distribución por municipios, 'MANIZALES' destaca con el mayor número de clientes, pero también con una tasa de retiro del 24.75%. Respecto a la antigüedad, es preocupante observar que los clientes con más de 36 meses presentan la mayor tasa de retiro, llegando al 33.74%.

En lo que respecta al estrato, se observa una distribución variada en la tasa de retiro. El estrato 2 presenta el mayor número de clientes y, a su vez, una significativa tasa de retiro del 40.92%. Por otro lado, el estrato 3, aunque cuenta con un número menor de clientes en comparación con el estrato 2, muestra una tasa de retiro del 33.65%. Los demás estratos tienen un número reducido de clientes y tasas de retiro comparativamente más bajas, pero aun así es importante considerar estas cifras en cualquier estrategia de retención. Estos datos indican que, si bien el estrato 2 es una fuente principal de clientes, también representa un segmento en la que las tasas de retiro son altas, lo que puede requerir intervenciones específicas para mejorar la retención.

En la Figura 4 se presenta la distribución de los datos según las características numéricas.

Figura 4. Distribución de características numéricas

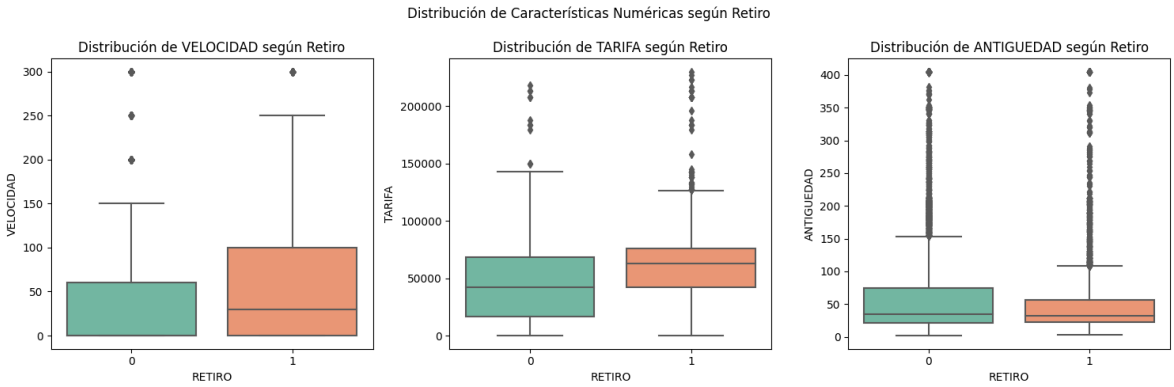


Fuente: empresa de Telecomunicaciones.

La Figura 4 muestra la distribución de las variables "VELOCIDAD", "TARIFA" y "ANTIGÜEDAD", que se relacionan con los servicios ofrecidos a los clientes. En "VELOCIDAD", la mayoría de los clientes se benefician de velocidades más bajas, aunque hay algunos que optan por velocidades significativamente más altas, lo que puede indicar paquetes premium o necesidades específicas. En cuanto a "TARIFA", si bien la tarifa promedio ronda los 100,000, existen clientes que pagan significativamente más o menos, reflejando posiblemente diferentes paquetes o promociones. Respecto a la variable "ANTIGÜEDAD" sugiere que gran parte de la base de clientes es de adquisiciones más recientes, pero existen clientes que han permanecido con la empresa durante un tiempo prolongado. Esta diferencia en la antigüedad podría señalar una buena adquisición de nuevos clientes, pero también destaca la importancia de estrategias para mantener a los clientes más antiguos satisfechos y retenerlos.

Según retiro, en la Figura 5 se presentan las características cuantitativas para el segmento de retiro.

Figura 5. Distribución de características numéricas según retiro



Fuente: empresa de Telecomunicaciones.

En "VELOCIDAD", se observa que aquellos clientes que decidieron retirarse contaban con velocidades menores en comparación con los que permanecieron. Respecto a "TARIFA", los

clientes que se retiraron pagaron montos ligeramente menores que los que decidieron quedarse, aunque la variabilidad es amplia en ambos grupos, con valores atípicos tanto en la parte superior como inferior. En cuanto a "ANTIGÜEDAD", es notable que la mayoría de los clientes que se retiraron tienen una antigüedad menor, indicando que es más probable que los clientes recientes abandonen la empresa, mientras que aquellos con mayor tiempo son más propensos a permanecer. Estos hallazgos sugieren la necesidad de revisar las ofertas para clientes nuevos y considerar estrategias para mejorar la retención temprana.

4.2 Fase 2: Generación de modelo clasificatorio que permita explicar el churn en una empresa de Telecomunicaciones

4.2.1 Preparación de los datos

La preparación de los datos es una etapa crucial en cualquier proyecto de análisis de datos o Machine Learning. Es en esta fase se asegura que los datos estén en el formato adecuado, sean consistentes y estén listos para su posterior análisis o modelado. En el contexto de esta investigación sobre el churn en una empresa de telecomunicaciones del suroccidente de Caldas, se cuenta con una base de datos compuesta por 18,571 registros. Cada registro representa un servicio específico y está descrito por un conjunto de 15 variables o características. Estas variables abarcan tanto aspectos cuantitativos como cualitativos del servicio. Algunos puntos clave a destacar sobre la estructura de la base de datos son:

- Variables Cuantitativas: la mayoría de las variables de la base de datos es de tipo numérico (int64). Estas incluyen "PERIODO", "ID_SERVICIO", "VELOCIDAD", "ESTRATO", "TARIFA", "RETIRO", "PQR", "ANTIGUEDAD", "AÑO" y "MES". Estas variables proporcionan información cuantitativa sobre aspectos como la velocidad del servicio, tarifas, antigüedad del cliente, entre otros.

- Variables cualitativas: existen cinco variables de tipo objeto, que son "PRODUCTO", "MUNICIPIO", "TECNOLOGIA", "AGRUP_ANTIGUEDAD" y "PLAN". Estas variables ofrecen información descriptiva sobre el tipo de producto, la ubicación geográfica del servicio, la tecnología utilizada, entre otros aspectos.
- Identificadores únicos: la variable "ID_SERVICIO" parece ser un identificador único para cada servicio, lo que es esencial para garantizar que cada registro de la base de datos sea distinto y pueda ser identificado de manera individual.
- Variable objetivo: la variable "RETIRO" es de particular interés, dado que indica si un servicio está activo (0) o retirado (1). Esta será la variable objetivo o dependiente en el análisis posterior, especialmente cuando se trate de modelar y predecir el churn.
- Integridad de los Datos: es relevante destacar que no existen valores nulos en ninguna de las variables, lo que sugiere una buena integridad y completitud de los datos.

Con esta estructura de datos, el siguiente paso ejecutado en la preparación de los datos involucra la exploración detallada de cada variable, la identificación y tratamiento de posibles valores atípicos, la transformación de variables si es necesario y la selección de las características más relevantes para el análisis posterior. Un resumen detallado de esta información se puede observar en la Tabla 3.

Tabla 3. Variables relevantes para el análisis

```
RangeIndex: 18571 entries, 0 to 18570
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   PERIODO           18571 non-null  int64
 1   ID_SERVICIO       18571 non-null  int64
 2   PRODUCTO          18571 non-null  object
 3   VELOCIDAD         18571 non-null  int64
 4   MUNICIPIO         18571 non-null  object
 5   ESTRATO           18571 non-null  int64
 6   TECNOLOGIA        18571 non-null  object
 7   TARIFA            18571 non-null  int64
 8   RETIRO            18571 non-null  int64
 9   PQR               18571 non-null  int64
 10  AGRUP_ANTIGUEDAD  18571 non-null  object
 11  PLAN              18571 non-null  object
 12  ANTIGUEDAD        18571 non-null  int64
 13  AÑO               18571 non-null  int64
 14  MES               18571 non-null  int64
dtypes: int64(10), object(5)
memory usage: 2.1+ MB
```

Selección de Características Relevantes. Una etapa esencial en la preparación de los datos es la selección de variables. Esta fase implica identificar y conservar solo aquellas variables que se consideran relevantes para el análisis y modelado, mientras se descartan las que no aportan información significativa o que podrían introducir ruido en el modelo. Esta fase se realiza según los objetivos planteados en el apartado anterior. En efecto, se eliminaron las variables siguientes:

- PERIODO: aunque esta variable proporciona una marca temporal de cuándo se registró el servicio, no es relevante para determinar el retiro de un cliente. Además, en la base se cuenta con las variables "AÑO" y "MES" que la identifican.

- ID_SERVICIO: siendo un identificador único para cada servicio, esta variable no presenta una variabilidad que pueda ser útil para predecir el churn. Es esencial para garantizar la unicidad de cada registro, pero no aporta información predictiva.

- MES y AÑO: estas variables desglosan el "PERIODO" en componentes separados. Si bien pueden ser útiles para análisis temporales, si se considera que el churn no tiene una tendencia

temporal específica o si ya se han extraído características temporales relevantes, estas variables pueden ser redundantes.

Tras la eliminación de estas variables, el conjunto de datos se ha reducido en términos de dimensionalidad, lo que puede facilitar el análisis y modelado. Es de anotar que la selección de características es un proceso iterativo. A medida que avanzamos en el análisis, es posible que se identifiquen otras variables que se tienen que eliminar o, por el contrario, que se define la importancia de alguna que se haya descartado previamente. La línea de código utilizada para este propósito fue:

```
df = df.drop(['PERIODO', 'ID_SERVICIO', 'MES', 'AÑO'], axis=1)
```

Con esta acción, el conjunto de datos ahora se enfoca en las características que se consideran directamente relacionadas con el retiro de clientes, optimizando así el proceso de modelado y análisis posterior. La Tabla 4 presenta las variables consolidades.

Tabla 4. Variables esenciales para modelamiento

```
RangeIndex: 18571 entries, 0 to 18570
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   PRODUCTO          18571 non-null  object
 1   VELOCIDAD         18571 non-null  int64
 2   MUNICIPIO         18571 non-null  object
 3   ESTRATO           18571 non-null  int64
 4   TECNOLOGIA        18571 non-null  object
 5   TARIFA            18571 non-null  int64
 6   RETIRO            18571 non-null  int64
 7   PQR               18571 non-null  int64
 8   AGRUP_ANTIGUEDAD  18571 non-null  object
 9   PLAN              18571 non-null  object
 10  ANTIGUEDAD        18571 non-null  int64
dtypes: int64(6), object(5)
memory usage: 1.6+ MB
```

Codificación de Variables Cualitativas. Las variables cualitativas, también conocidas como categóricas, representan categorías o etiquetas que no tienen un orden o relación numérica inherente. Para que los algoritmos de machine learning puedan procesar y entender estas variables, es necesario convertirlas en un formato numérico. Una técnica común para lograr esto es la

codificación "one-hot". Esta implica convertir una variable categórica que tiene n posibles categorías en n variables binarias. Cada una de estas variables binarias representa una categoría y toma el valor de 1 si la categoría está presente y 0 si no lo está. En el contexto de nuestro análisis sobre el churn en la empresa de telecomunicaciones, se aplicó la codificación "one-hot" a las siguientes variables categóricas: producto, municipio, tecnología, antigüedad y plan. El código utilizado para realizar esta transformación fue:

```
df = pd.get_dummies(df, columns=['PRODUCTO', 'MUNICIPIO', 'TECNOLOGIA',
                                 'AGRUP_ANTIGUEDAD', 'PLAN'])
```

A partir de su transformación, cada categoría única dentro de estas variables se convierte en una nueva columna en el conjunto de datos. Por ejemplo, si la variable "PRODUCTO" tiene tres categorías: BA, TV y TO, después de aplicar la codificación "one-hot", se generan tres columnas: "PRODUCTO_BA", "PRODUCTO_TV" y "PRODUCTO_TO". Cada una es una variable binaria que indica la presencia o ausencia de esa categoría específica. Esta codificación facilita la inclusión de información categórica en modelos de machine learning y permite que los algoritmos procesen estas variables de manera efectiva. En la Tabla 5 se puede observar la transformación de las variables categóricas en nuevas variables discretas que contienen las categorías.

Tabla 5. Conversión de variables categóricas en discretas

```
RangeIndex: 18571 entries, 0 to 18570
Data columns (total 44 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   VELOCIDAD                       18571 non-null  int64
 1   ESTRATO                         18571 non-null  int64
 2   TARIFA                          18571 non-null  int64
 3   RETIRO                          18571 non-null  int64
 4   PQR                             18571 non-null  int64
 5   ANTIGUEDAD                      18571 non-null  int64
 6   PRODUCTO_BA                     18571 non-null  uint8
 7   PRODUCTO_TO                     18571 non-null  uint8
 8   PRODUCTO_TV                     18571 non-null  uint8
 9   MUNICIPIO_CHINCHINA             18571 non-null  uint8
 10  MUNICIPIO_MANIZALES             18571 non-null  uint8
 11  MUNICIPIO_VILLAMARIA            18571 non-null  uint8
 12  TECNOLOGIA_COPPER               18571 non-null  uint8
 13  TECNOLOGIA_HFC                  18571 non-null  uint8
 14  AGRUP_ANTIGUEDAD_0-6 MESES      18571 non-null  uint8
 15  AGRUP_ANTIGUEDAD_13-24 MESES    18571 non-null  uint8
 16  AGRUP_ANTIGUEDAD_25-36 MESES    18571 non-null  uint8
 17  AGRUP_ANTIGUEDAD_7-12 MESES     18571 non-null  uint8
```

4.3 Modelamiento

El modelamiento es una parte esencial de los modelos de Machine Learning, pues hace posible la creación de los datos de entrenamiento y validación, cubriendo las técnicas de balanceo de datos y, finalmente, la generación de los modelos de machine Learning.

4.3.1 Creación de la Variable X e Y

En el modelamiento fue necesario crear la variable X, compuesta de todas aquellas otras que se incluirán en el modelo. Además, se crea Y, conformada por la variable objetivo del proyecto que es el retiro. Su conformación se da de la forma siguiente:

```
X=df[['VELOCIDAD', 'ESTRATO', 'TARIFA', 'PQR', 'ANTIGUEDAD',
'PRODUCTO_BA', 'PRODUCTO_TO', 'PRODUCTO_TV',
'MUNICIPIO_CHINCHINA', 'MUNICIPIO_MANIZALES',
'MUNICIPIO_VILLAMARIA', 'TECNOLOGIA_COPPER',
'TECNOLOGIA_HFC', 'AGRUP_ANTIGUEDAD_0-6 MESES',
'AGRUP_ANTIGUEDAD_13-24 MESES', 'AGRUP_ANTIGUEDAD_25-36MESES',
'AGRUP_ANTIGUEDAD_7-12 MESES', 'AGRUP_ANTIGUEDAD_>36 MESES',
'PLAN_100Mb', 'PLAN_100Min', 'PLAN_10Mb', 'PLAN_120Mb', 'PLAN_150Mb',
'PLAN_15Mb', 'PLAN_1Mb', 'PLAN_200Mb', 'PLAN_20Mb', 'PLAN_250Mb',
'PLAN_25Mb', 'PLAN_2Mb', 'PLAN_300Mb', 'PLAN_300Min', 'PLAN_30Mb',
'PLAN_3Mb', 'PLAN_4Mb', 'PLAN_500Min', 'PLAN_50Mb', 'PLAN_5Mb',
'PLAN_600Min', 'PLAN_60Mb', 'PLAN_AVANZADO', 'PLAN_BASICO',
```

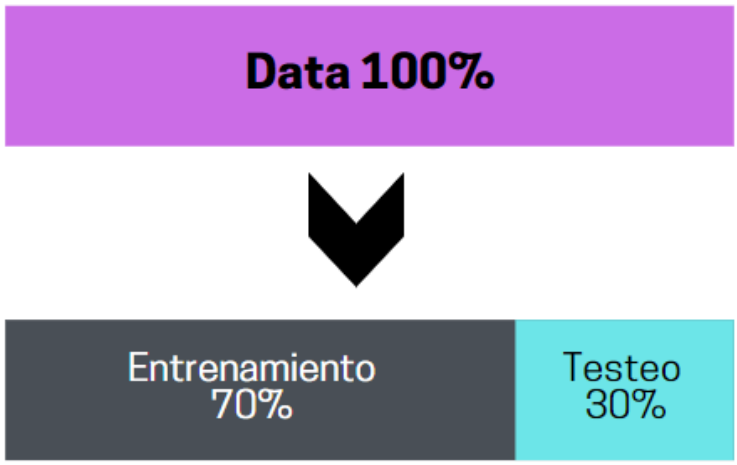
```
'PLAN_ILIMITADO']]

y=df['RETIRO']
```

4.3.2 Creación de Test y Train

El conjunto de datos original se divide en un conjunto de prueba (test) y en un conjunto de entrenamiento (train). En esta investigación se realizò en una proporción de 70% entrenamiento (Train) y 30% de prueba (test). Esto tuvo el fin de entrenar los modelos con una cantidad de datos suficiente y, posterior a ello, validar los resultados obtenidos frente al conjunto de datos de prueba (ver Figura 6).

Figura 6. División de datos para Machine Learning



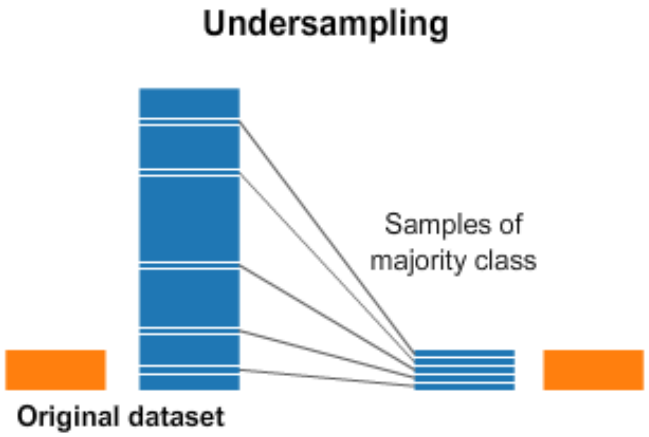
4.3.3 Balanceo de datos

En la Figura 7 se observa que la variable Churn se encuentra medianamente desbalanceada, siendo la categoría NO la predominante con un 69,9% de los datos y la categoría SI con el 30,1% restante.

Figura 7. Variable churn

Desbalanceo Variable Churn. Mediante la técnica random undersampler; la cual elimina de manera aleatoria una cantidad de muestras pertenecientes a la clase mayoritaria hasta igualar ambas clases, se realizó el balanceo de la variable churn. El enfoque de la técnica se presenta en la Figura 8.

Figura 8. Undersampling



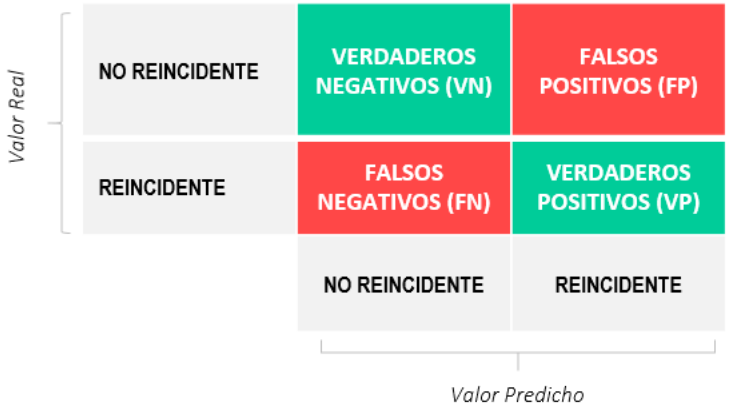
Fuente: www.kaggle.com/code/nikunjmalpan

Modelos implementados. Se implementaron los modelos de clasificación Regresión logística, Decision Tree, Random Forest, Regresión logística y Gradient Boosted Machines GBM. Estos modelos se implementaron con la totalidad de variables previamente descritas.

4.4 Evaluación

En esta investigación se utilizaron diferentes métricas para evaluar los modelos. Una de ellas la matriz de confusión. En la Figura 9 se presenta la estructura de una matriz de confusión

Figura 9. Matriz de confusión



Como lo establecen Igual y Seguí (2020), los significados del contenido de la matriz de confusión se observan a continuación:

- **Verdaderos positivos:** cuando el clasificador predice una muestra como positiva y realmente es positiva.

- **Falsos positivos:** cuando el clasificador predice una muestra como positiva, pero de hecho es negativa.

- **Negativos verdaderos:** cuando el clasificador predice una muestra como negativa y realmente es negativo.

- **Falsos negativos:** cuando el clasificador predice una muestra como negativa, pero de hecho es positivo.

De la matriz de confusión derivan las métricas exactitud, sensibilidad, precisión y especificidad (ver Tabla 6).

Tabla 6. Métricas de valuaciòn del modelo

Métrica	**Formula**
Exactitud (Accuracy) =	(VP+VN)/(VP+FP+FN+VN)
Precisión (Precision) =	(VP)/(VP+FP)
Sensibilidad (Recall) =	(VP)/(VP+FN)
Especificidad (Especificity) =	(VN)/(VN+FP)

Si bien la **¡Error! No se encuentra el origen de la referencia.** Tabla 6 define las fórmulas para obtener el resultado, se explica a continuación la definición de estas métricas aplicadas a la variable objetivo: el retiro de clientes de la empresa de telecomunicaciones (ver Tabla 7).

Tabla 7. Definición de las métricas según variables de estudio

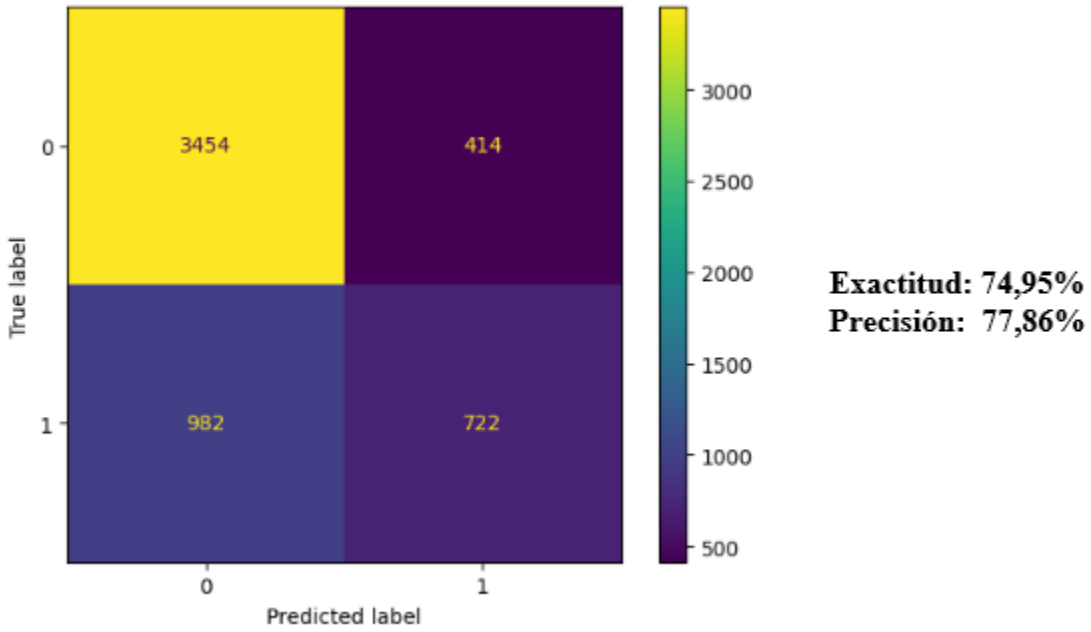
Métrica	**Interpretación**
Exactitud (Accuracy)	Identificar los retiros y no retiros **en general**.
Precisión (precision)	Detectar correctamente los retiros **sin tener que clasificar erróneamente a personas que no se van a retirar**
Sensibilidad (Recall)	Detectar correctamente los **retiros entre los verdaderos retiros**.
Especificidad (Especificity)	Identificar los casos de las **personas que no se retiraron entre todos los que no se retiraron de la empresa de telecomunicaciones**

4.4.1 Evaluación de los modelos

A razón de que el objetivo del modelo es distinguir las características de los clientes que realizan retiro (churn) y, además, considerando que la base de entrenamiento está medianamente balanceada, se ha elegido el accuracy (precisión) como métrica de optimización. Esta elección se debe a su facilidad de interpretación y a su capacidad para medir tanto la precisión en la predicción de los churners como en la predicción de los no churners.

Decision Tree. En la Figura 10 se presentan los resultados de la implementación del Decision Tree.

Figura 10. Resultados Decision Tree

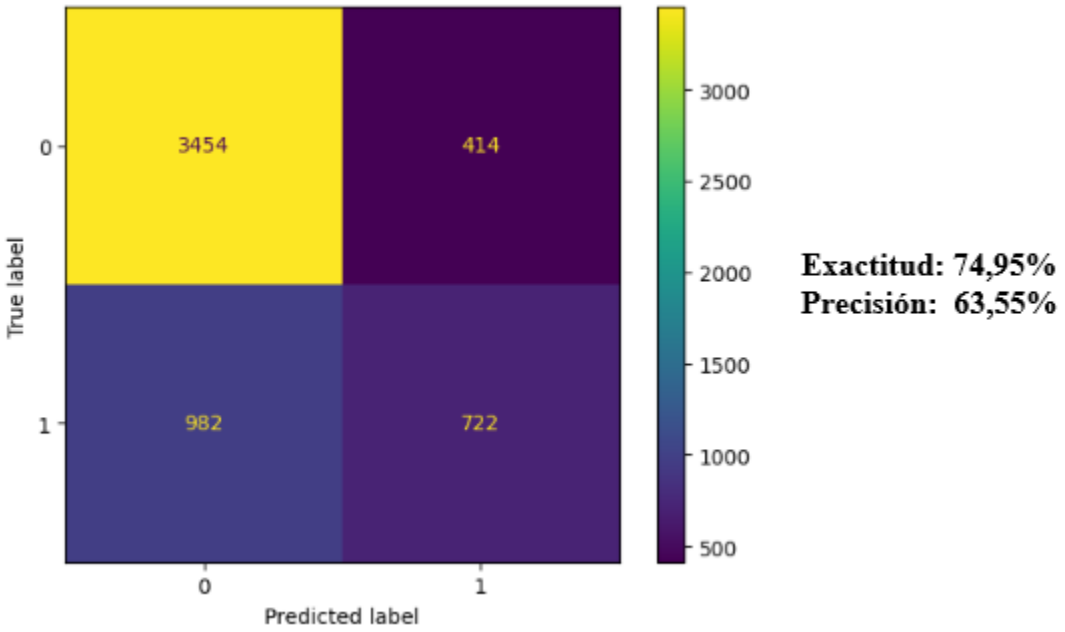


La exactitud (Accuracy) alcanzó un valor del 74.95%, lo que indica que el modelo clasifica correctamente. Esto significa que, aproximadamente, categoriza correctamente el 74.95% de las instancias en el conjunto de prueba. Este nivel refleja la efectividad general del modelo en prever si los clientes harán churn o no churn. Además, la precisión del modelo, respecto a la clasificación de 'no churners' (clientes que se quedan), alcanzó un 77.86%. Esto significa el porcentaje de las

predicciones de retención que son correctas, lo que es fundamental para garantizar que estas sean precisas y se minimicen las falsas alarmas.

Random Forest. En la Figura 11 se presentan los resultados para el modelo Random Forest.

Figura 11. Resultados Random Forest



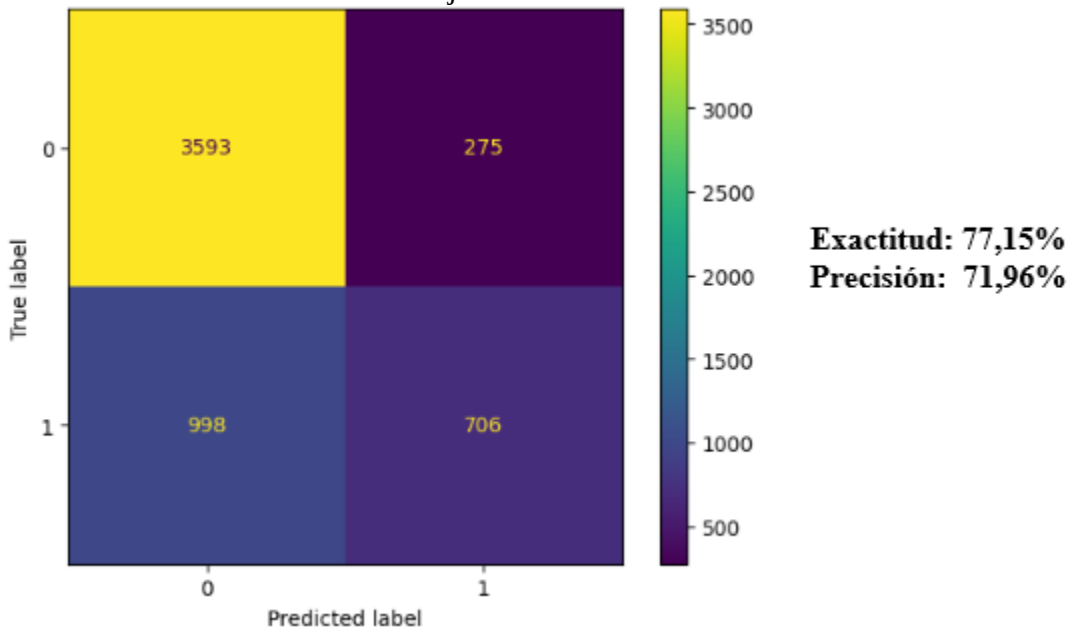
La comparación entre el modelo de árbol de decisión y el modelo de Random Forest, en el contexto del churn de servicios de telecomunicaciones, revela diferencias significativas en cuanto a precisión. Ambos modelos lograron una exactitud (Accuracy) del 74.95%, lo que indica una capacidad similar para clasificar correctamente las instancias en el conjunto de prueba. Sin embargo, en términos de precisión, el modelo de árbol de decisión -*decision tree*- supera al modelo de Random Forest. El modelo de árbol de decisión logró una precisión del 77.86%, mientras que el modelo de Random Forest obtuvo una precisión del 63.55%.

Estos resultados destacan la importancia de considerar la precisión en la retención de clientes en la empresa de telecomunicaciones analizadas. A pesar de que ambos modelos tienen un rendimiento sólido, el modelo de árbol de decisión demuestra una mayor precisión en la

clasificación de 'no churners', lo que reduce la probabilidad de falsas alarmas en la gestión de la retención de clientes.

Random Forest Ajustado. En la Figura 12 se presentan los resultados del modelo de Random Forest *Ajustado.*

Figura 12. Resultados Random Forest Ajustado

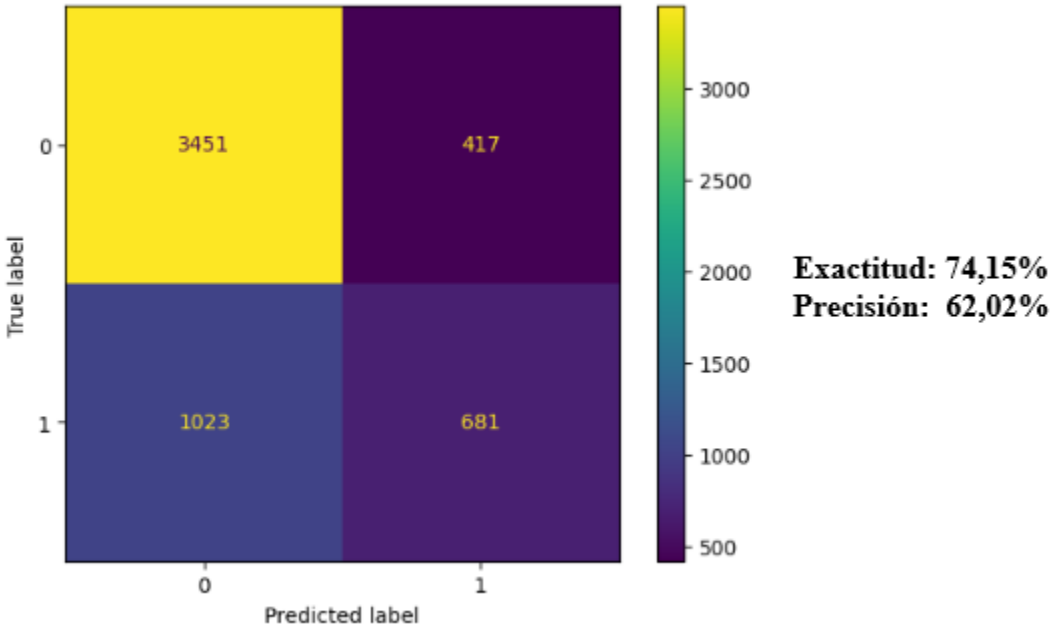


El modelo de Random Forest, después de la realización de ajustes específicos, demostró un rendimiento mejorado en la predicción de la retención de clientes en el contexto del churn de servicios de telecomunicaciones. La precisión global (Accuracy) del modelo de Random Forest, después del ajuste alcanza, el 77.15%. Esto representa una mejora significativa en comparación con el modelo de árbol de decisión original. Esto significa que el modelo clasifica correcta y aproximadamente al 77.15% de las instancias en el conjunto de prueba. Además, la precisión del modelo, medida en 71.96%, indica una capacidad superior para clasificar correctamente a los clientes como 'churners' o 'no churners'. Esta precisión mejorada es fundamental para garantizar que las predicciones de retención sean precisas y confiables, reduciendo así los errores en la gestión de la retención de clientes. Así pues, estos resultados resaltan la eficacia de un modelo de Random

Forest ajustado en la retención de clientes en la industria de las telecomunicaciones y enfatizan la importancia de considerar estrategias de ajuste para mejorar el rendimiento del modelo.

Regresión logística. En la Figura 13 se presentan los resultados para el modelo de regresión logística.

Figura 13. Resultados modelo de regresión logística

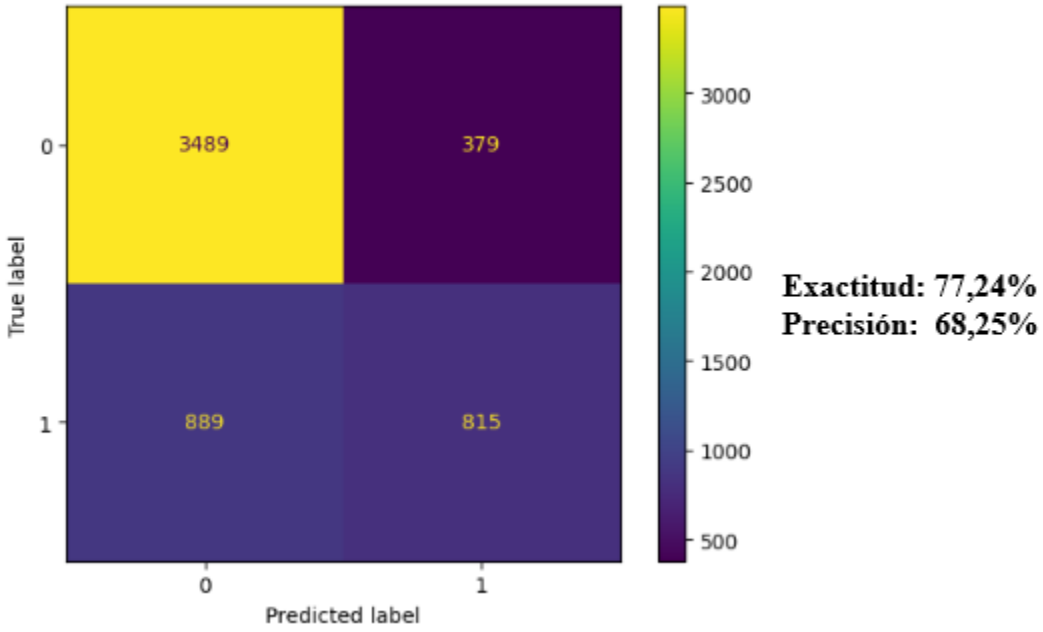


La comparación entre el modelo de árbol de decisión, el modelo de Random Forest -después del ajuste- y el modelo de Regresión Logística en el contexto del churn de servicios de telecomunicaciones revela diferencias en términos de precisión. El modelo de Regresión Logística logra una precisión global (Accuracy) del 74.15%, lo que refleja su capacidad para clasificar correctamente las instancias en el conjunto de prueba. Este nivel de precisión global es similar al del modelo de árbol de decisión. Sin embargo, en términos de precisión, el modelo de Regresión Logística (62.02%) muestra un rendimiento ligeramente inferior al modelo de árbol de decisión (77.86%) y al modelo de Random Forest después del ajuste (71.96%). La precisión del modelo de

Regresión Logística es moderada, lo que sugiere una capacidad adecuada para clasificar correctamente a los clientes como 'churners' o 'no churners.

XGBoost. En la Figura 14 se presentan los resultados del modelo XGBoost.

Figura 14. Modelo XGBoost

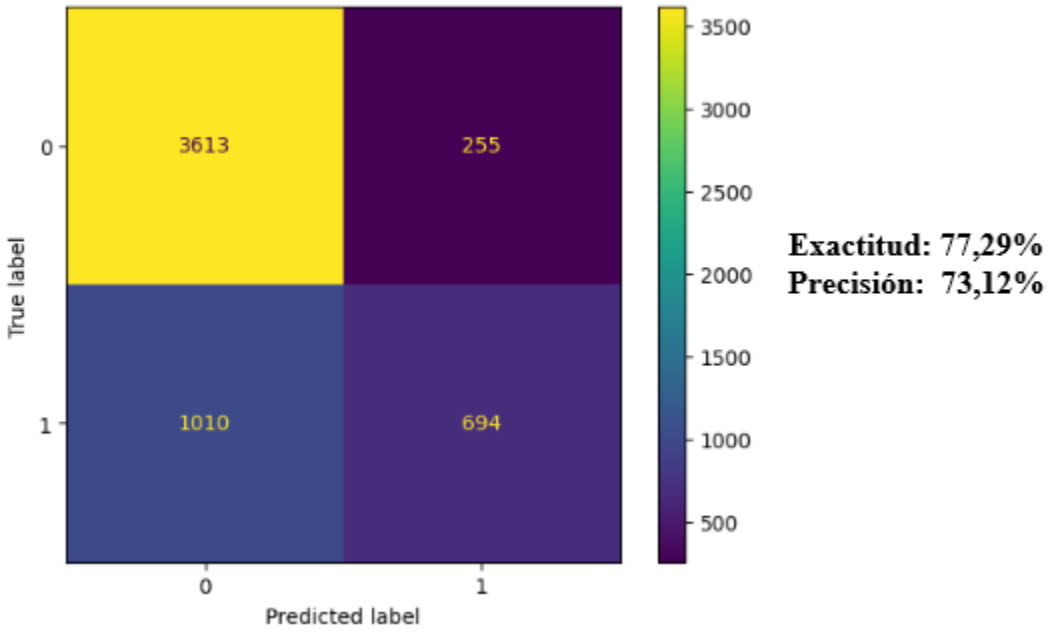


La comparación entre el modelo de árbol de decisión, el modelo de Regresión Logística y el modelo XGBoost, en el contexto del churn de servicios de telecomunicaciones, resalta diferencias significativas en términos de precisión. El modelo XGBoost logra una precisión global (Accuracy) del 77.24%, lo que representa una mejora notable en comparación con el modelo de árbol de decisión original. Esto significa que el modelo XGBoost clasifica correctamente aproximadamente el 77.24% de las instancias en el conjunto de prueba, lo que indica un rendimiento mejorado. En términos de precisión, el modelo XGBoost (68.25%) supera tanto el modelo de árbol de decisión (77.86%) como también el modelo de Regresión Logística (62.02%). La precisión mejorada del modelo XGBoost sugiere su capacidad para clasificar correctamente a los clientes como 'churners' o 'no churners' con una precisión moderada. Estos resultados enfatizan

la utilidad del modelo XGBoost en la retención de clientes en la industria de las telecomunicaciones y subrayan la importancia de considerar algoritmos de aprendizaje automático avanzados para mejorar la gestión de la retención de clientes.

XGBoost Ajustado. En la Figura 15 se presentan los resultados del XGBoost ajustado.

Figura 15. Resultados XGBoost Ajustado

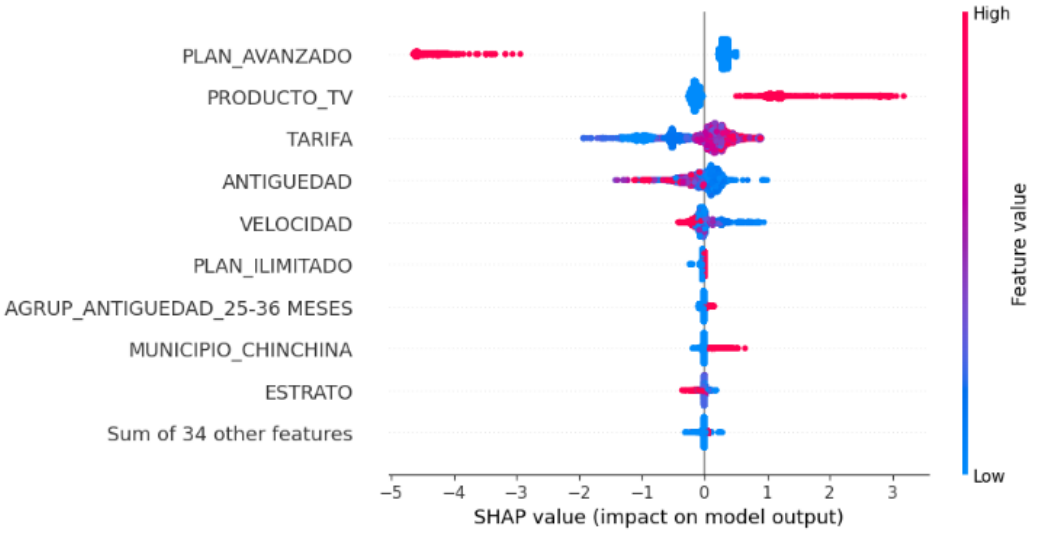


La comparación entre el modelo de árbol de decisión y el modelo XGBoost, después de la realización de ajustes específicos, destaca una mejora significativa en el rendimiento del modelo XGBoost en el contexto del churn de servicios de telecomunicaciones. El modelo XGBoost, después del ajuste, logra una precisión global (Accuracy) del 77.29%, lo que representa una mejora adicional en comparación con el modelo de árbol de decisión original. Esto significa que el modelo XGBoost clasifica correctamente aproximadamente el 77.29% de las instancias en el conjunto de prueba, lo que indica un rendimiento superior en términos de clasificación. Además, la precisión del modelo XGBoost después del ajuste, medida en 73.12%, refleja una capacidad mejorada para clasificar correctamente a los clientes como 'churners' o 'no churners'. Esta precisión mejorada es

fundamental para garantizar que las predicciones de retención sean precisas y confiables, lo que contribuye con la eficacia de la gestión de la retención de clientes en la industria de las telecomunicaciones. A razón de que la Exactitud (Accuracy) es el criterio principal de evaluación, el modelo XGBoost, después del juste, supera a los demás modelos al clasificar con mayor precisión tanto a los "churners" como a los "no churners". Esto lo convierte en la elección preferida según tu criterio principal.

Grafica de enjambre. Esta gráfica se utiliza para mostrar cómo cada característica contribuye con las predicciones del modelo y cómo se distribuyen esos valores. A la derecha apoya el churn, a la izquierda apoya el no churn. El color indica el valor de la variable, lo azul es igual a 0 y lo rojo es igual a 1. La figura 16 presenta un Gráfico de Enjambre de Valores de Predicción para Churn. En el eje Y se representan las categorías de productos de telecomunicaciones, mientras que en el eje X se muestran los valores de predicción de churn. Cada punto en el gráfico representa una predicción para un cliente. Se observa entonces una concentración de predicciones de churn en la categoría de 'Productos de TV', lo que sugiere una mayor probabilidad de churn en este grupo. Esto tiene implicaciones importantes para la formulación de estrategias de retención de clientes.

Figura 16. Gráfica de enjambre



4.4.2 Evaluación de resultados

Se evaluaron y compararon los resultados obtenidos en todos los modelos implementados. Se revisaron las métricas de Sensibilidad ("Recall") y Exactitud ("Accuracy") que, para este caso, se consideran las más relevantes. Los resultados se observan en la

				Métricas	
Modelo	**Exactitud (Accuracy)**	**Exactitud (Accuracy Ajustado)**	**Precisión (Precision)**	**Precisión (Precision Ajustado)**	**Sensibilidad (Recall)**
Decision_Tree	74,95%		77,86%		42,37%
Random_Forest	74,95%	77,15%	63.55%	71,96%	42,37%
LogisticRegression	74,15%		62,02%		39,96%
XGBoost	77,24%	77,29%	68,25%	73,12%	47,82%

.

Tabla 8. Comparación de resultados según modelos

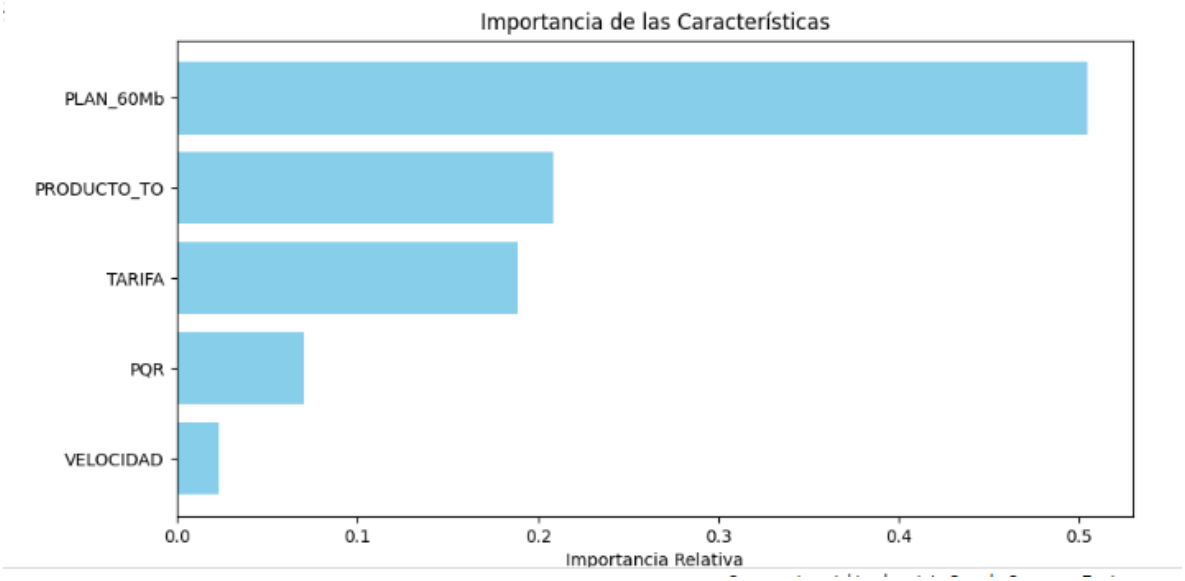
				Métricas	
Modelo	**Exactitud (Accuracy)**	**Exactitud (Accuracy Ajustado)**	**Precisión (Precision)**	**Precisión (Precision Ajustado)**	**Sensibilidad (Recall)**
Decision_Tree	74,95%		77,86%		42,37%
Random_Forest	74,95%	77,15%	63.55%	71,96%	42,37%
LogisticRegression	74,15%		62,02%		39,96%
XGBoost	77,24%	77,29%	68,25%	73,12%	47,82%

De acuerdo con la Tabla 8, en todas las métricas evaluadas, el modelo XGBoost fue el de mejor desempeño, superando en exactitud, precisión, sensibilidad y área bajo la curva, a todos lo demás modelos.

4.4.3 Importancia de las características o variables por modelo

Decisión Tree.

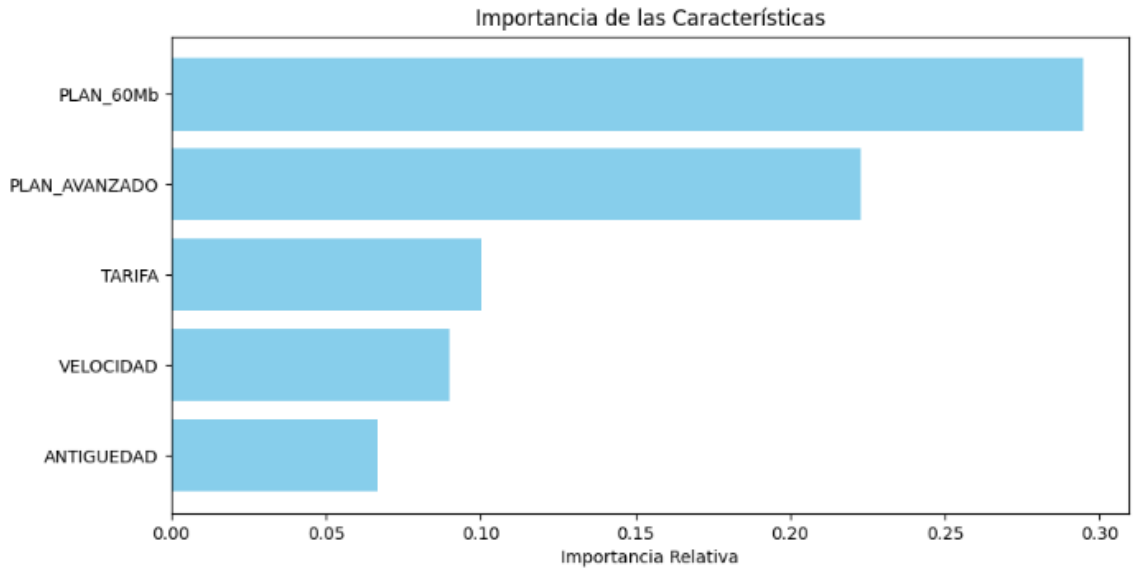
Figura 17. Importancia de las características según Decision Tree



La Figura 17 muestra la importancia relativa de diversas características en un modelo de churn para una empresa de telecomunicaciones basado en el algoritmo de clasificación de Decision Tree. Se destaca la variable "PLAN_60Mb" como el factor más influyente, sugiriendo que la elección de este plan tiene un impacto considerable en la decisión del cliente de permanecer o abandonar; las variables "PRODUCTO_TO" y "TARIFA" le siguen en importancia, indicando que el tipo de producto y las tarifas también juegan un papel crucial en el comportamiento de churn. Mientras tanto, "PQR" y "VELOCIDAD" tienen menor peso en la predicción, pero aún deben ser considerados en las estrategias de retención de clientes.

Random Forest.

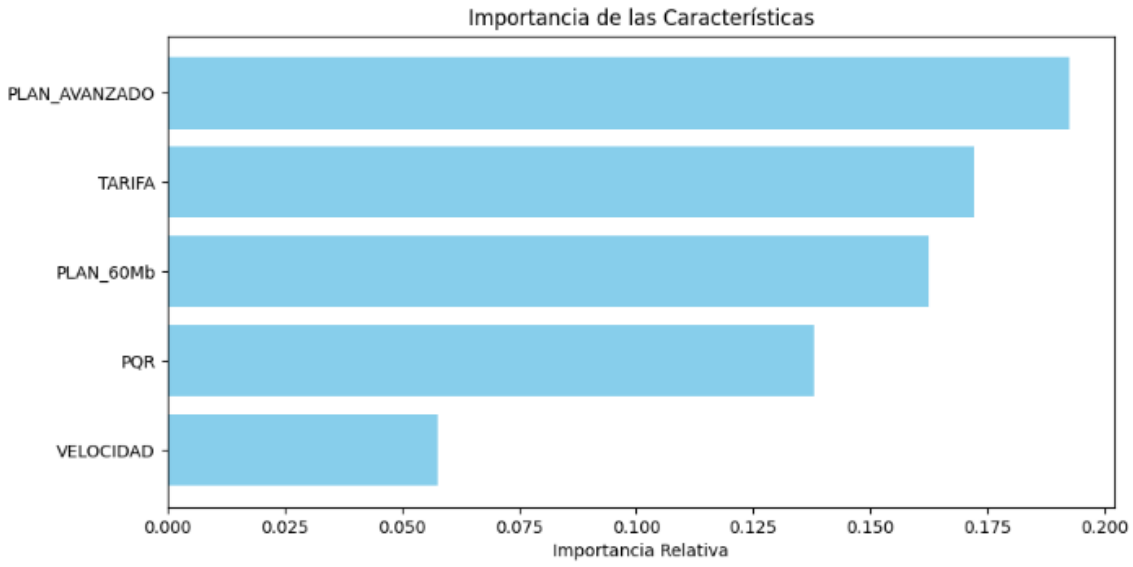
Figura 18. Importancia de las características según Random Forest



La Figura 18 resalta la importancia relativa de las características para el algoritmo de Random Forest; nuevamente la variable "PLAN_60Mb" emerge como el atributo más determinante, indicando que este plan tiene una fuerte influencia en el comportamiento de retención o abandono del cliente. Le sigue "PLAN_AVANZADO", que también juega un papel significativo en el churn. "TARIFA" y "VELOCIDAD" tienen un impacto mediano en la decisión del cliente, mientras que "ANTIGUEDAD" parece ser el factor menos influyente en este contexto, pero aun así debería tenerse en cuenta en las estrategias de retención.

Random Forest Ajustado.

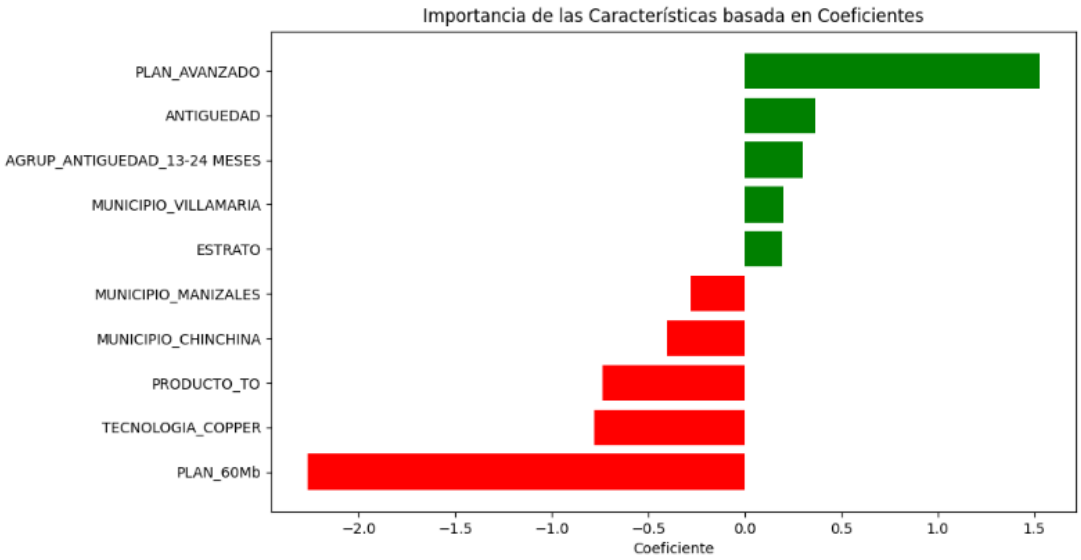
Figura 19. Importancia de las características según Random Forest Ajustado



En Figura *19* se tiene una nueva representación para el algoritmo de Random Forest Ajustado, la variable "PLAN_AVANZADO" se sitúa como el atributo más preponderante, sugiriendo que dicho plan tiene un impacto significativo en el comportamiento de retención o abandono. "TARIFA" ocupa el segundo lugar, subrayando su relevancia en la decisión del cliente. Aunque "PLAN_60Mb" había sido el más influyente en los gráficos anteriores (Figura 17 y Figura 18), aquí se presenta en tercer lugar, manteniendo aún una influencia notable. "PQR" y "VELOCIDAD" se sitúan en los lugares posteriores, con la velocidad siendo la menos determinante en esta ocasión. Estos cambios pueden indicar una evolución en las prioridades o percepciones de los clientes con respecto a los servicios y ofertas de la empresa.

Regresión Logística

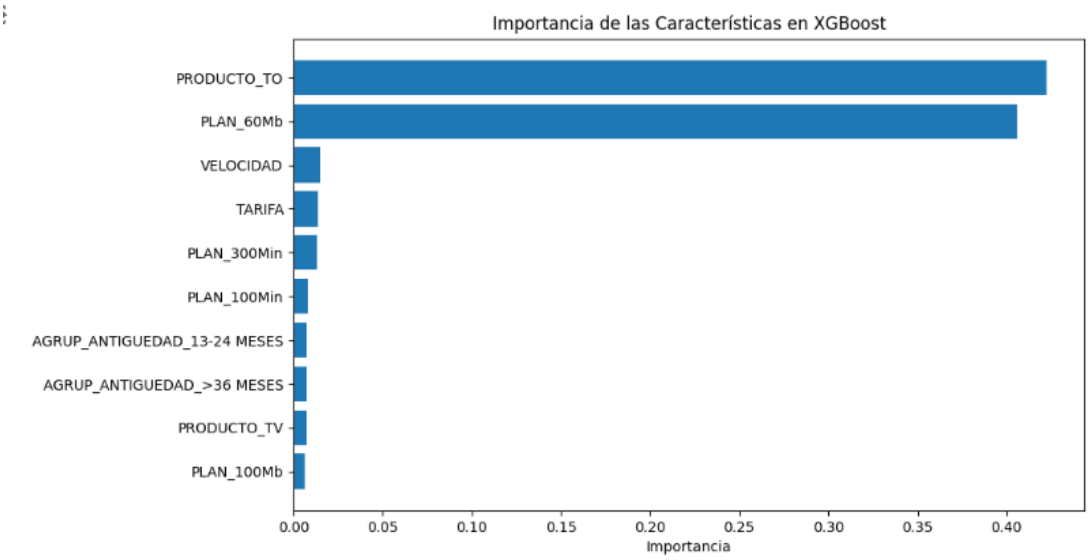
Figura 20. Resultados de variables regresión logística



La Figura *20* refleja la importancia de las características basada en coeficientes de un modelo de regresión logística. Las barras verdes, como "PLAN_AVANZADO" y "ANTIGUEDAD", indican que estas características aumentan la probabilidad del evento objetivo (churn), mientras que las barras rojas, como "SITIO", "MUNICIPIO_VILLAMARÍA" y "PLAN_60Mb", sugieren que reducen esa probabilidad. En esencia, clientes en el "PLAN_AVANZADO" o con mayor antigüedad tienen una mayor tendencia a churn, mientras que ciertos sitios, municipios o planes están asociados con una retención más alta.

XGBoost

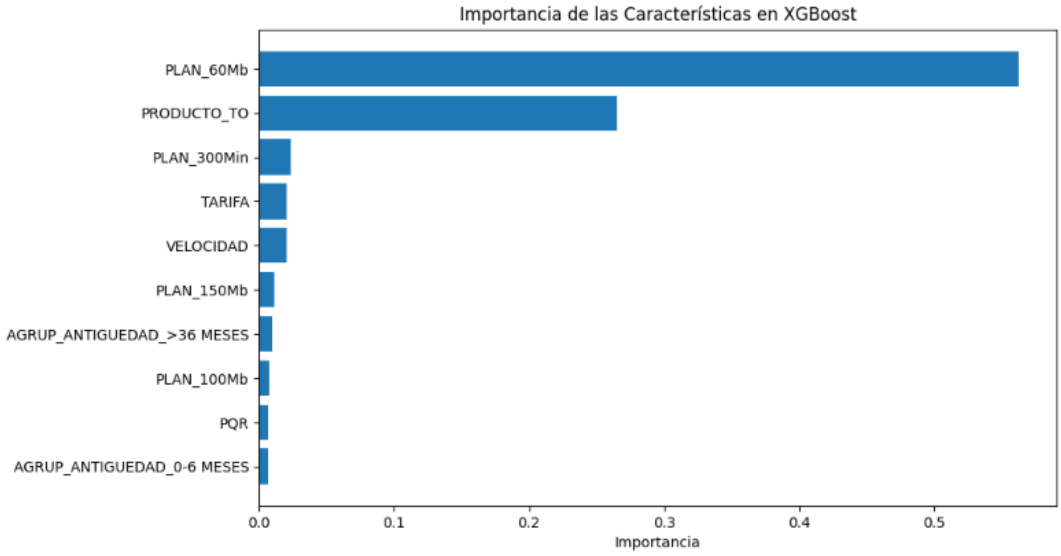
Figura 21. Resultados por características XGBoost



La Figura 21 muestra la importancia de las características para un modelo de clasificación XGBoost para predecir churn en telecomunicaciones. "PRODUCTO_TO" destaca como el atributo más influyente, indicando que tiene una relación fuerte con la tendencia de churn. Le siguen "PLAN_60Mb" y "VELOCIDAD" con importancias medianas, mientras que las otras características, como "TARIFA", "PLAN_100Mb" y "EQUIPAMIENTO 12-24 MESES", tienen impactos menores en el modelo. Es evidente que el tipo de producto y el plan de 60Mb son consideraciones primordiales para entender la retención o abandono de clientes.

XGBoost Ajustado

Figura 22. Resultados de las características según la técnica XGBoost



Finalmente, La Figura 22 ilustra la importancia de las características en un modelo XGBoost Ajustado relacionado con telecomunicaciones. "PLAN_60Mb" sobresale como la característica más significativa, sugiriendo un alto impacto en la predicción del churn. Le siguen "PRODUCTO_TO" y "TARIFA" con importancias notables. Otras características como "VELOCIDAD", "PLAN_150Mb" y "EQUIPAMIENTO 9-24 MESES" presentan relevancias medianas. Las características con menor importancia incluyen "PQR" y "EQUIPAMIENTO 0-4 MESES". Esto implica que el tipo de plan, específicamente "PLAN_60Mb", y el producto desempeñan roles cruciales en la retención o pérdida de clientes.

Tras la evaluación exhaustiva de varios modelos, se observa que las características asociadas con los tipos de plan, en particular "PLAN_60Mb", y la categoría "PRODUCTO_TO" son consistentemente identificadas como las de mayor importancia en la predicción del churn en el sector de telecomunicaciones. Las técnicas basadas en árboles, como Decision Trees, modelos de Random Forest, regresión Logistiva y XGBoost, han subrayado la relevancia de estos atributos en distintos grados, mientras que otros factores como "TARIFA", "VELOCIDAD", y la antigüedad del equipamiento, aunque también influyentes, poseen un peso menor en el análisis global. Estos

hallazgos sugieren una necesidad de focalizar las estrategias de retención en los aspectos relacionados con los planes y productos ofrecidos, optimizando así la retención de clientes y, en última instancia, maximizando la rentabilidad de la empresa.

CAPÍTULO 5: DISCUSIÒN

Este estudio tuvo como objetivo establecer los factores que inciden en el "churn" (abandono de clientes) de una empresa que presta servicios de telecomunicaciones que opera en los municipios de Manizales, Villamaría y Chinchiná (Departamento de Caldas). Para alcanzarlo, se utilizaron técnicas de Machine Learning en el programa Python. En particular, se utilizó el modelo X-Gboost, que es una técnica de frontera de este campo de la analítica de datos, lo que aporta a la novedad de este estudio. Adicionalmente, para contextualizar este estudio a nivel organizacional, se utilizó como referente metodológico el CRISP-DM.

5.1 Implicaciones de investigación

En primer lugar, La inclusión de XGBoost en esta tesis representa una valiosa adición, dado que este modelo ofrece un rendimiento excepcional, maneja datos desequilibrados, facilita la interpretación de características, brinda opciones de regularización efectivas, se beneficia de la paralelización y es ampliamente utilizado en la comunidad de aprendizaje automático.

En segundo lugar, el modelo de XGBoost ajustado sobresalió en la predicción del churn en servicios de telecomunicaciones. Las características clave que influyeron en la retención de clientes fueron el plan de 60MB, el Producto TO (Televisión por Cable) y el Plan de 300 Minutos. Estos hallazgos subrayan la importancia de adaptar estrategias de retención para satisfacer las preferencias y necesidades de los clientes, lo que puede conducir a una mayor satisfacción y a una reducción en las tasas de retiro. Estos resultados tienen implicaciones significativas para la toma de decisiones en la industria de las telecomunicaciones.

5.2 Implicaciones prácticas

En este apartado se da respuesta al objetivo número tres de esta investigación, el cual consiste en Formular acciones o estrategias empresariales para mejorar la retención de los clientes de la empresa de servicios de telecomunicaciones en las ciudades de Manizales, Chinchiná y Villamaría en el departamento de Caldas (Colombia). Así pues, de acuerdo con los resultados de este estudio, en la Tabla 9 se presentan las acciones o estrategias a implementar para prevenir el Churn en la empresa de telecomunicaciones.

Tabla 9. Estrategias para prevenir el Churn

Estrategia	Responsables	Indicador
Implementar un programa de lealtad con descuentos y beneficios exclusivos.	Departamento de Marketing y Ventas.	Incremento en el % de clientes que se inscriben y participan activamente en el programa.
Realizar encuestas periódicas para identificar áreas de mejora.	Departamento de Atención al Cliente.	Reducción en el % de quejas o reclamos relacionados con áreas identificadas en las encuestas.
Ofrecer paquetes personalizados según las necesidades del cliente.	Departamento de Marketing y Ventas.	Incremento en la satisfacción del cliente y reducción en la tasa de churn.
Capacitar al personal de atención al cliente en habilidades blandas.	Departamento de Gestión Humana	Mejora en las calificaciones y feedback de los clientes tras interactuar con el servicio al cliente.
Establecer un canal de comunicación directo para solucionar problemas.	Departamento de Atención al Cliente y servicio técnico.	Reducción en el tiempo medio de resolución de problemas reportados por los clientes.
Realizar campañas de concientización sobre el valor añadido de los servicios.	Departamento de Marketing y ventas.	Incremento en el reconocimiento de marca y en la percepción positiva de los servicios.
Identificar las propuestas de valor de la competencia (operadores del suroccidente de Caldas) para ajustar la propuesta de valor de la compañía	Departamento de Marketing y ventas.	Número de análisis del mercado externo por año

Instaurar una cultura adhocrática: cultura enfocada en la innovación en servicio a través del trabajo en equipo y en el análisis de las necesidades del consumidor.	Departamento de Gestión Humana	Número de empleados que comparten la cultura

En la Tabla 9 se observa la relevancia que tiene el departamento de marketing y ventas de la empresa de telecomunicaciones para prevenir el churn. Adicionalmente, es un departamento clave en guiar la implementación del CRISP-MD, de tal modo que se puedan implementar las técnicas de analítica de datos que soporten el modelo de negocio de la organización. Las acciones presentadas serán más eficaces si son diseñadas con base en las técnicas de analítica de datos apropiadas.

5.3 Limitaciones y futuras investigaciones

Este estudio posee limitaciones que tienen que ser contempladas para la comprensión de los datos. En primer lugar, el estudio fue realizado en una empresa de telecomunicaciones que opera en el suroccidente del departamento de Caldas, específicamente, en los municipios de Manizales, Villamaría y Chinchiná. En este sentido, los resultados tienen que generalizarse con cuidado a otras compañías de telecomunicaciones, pues este estudio solo se realizó con datos de dicha compañía. En segundo lugar, los resultados no pueden abordarse desde una perspectiva de causalidad, a razón de que los datos de las variables independientes (X) y dependientes (Y) fueron obtenidos en el mismo momento temporal. En tercer lugar, las variables independientes seleccionadas fueron seleccionadas con base en la experiencia de la autora de esta tesis, lo que puede generar sesgos en el modelo. Pese a esto, como se identificó en la sección de resultados, los modelos demuestran consistencia.

Respecto a futuras investigaciones, los resultados obtenidos en este estudio podrían confrontarse con el texto obtenido de entrevistas de retiro que diligencian los clientes. Así, a partir

de un proceso de analítica de texto, podrían profundizarse en el fenómeno del Churn desde un punto de vista del cliente y no tanto desde la perspectiva de la empresa. Adicionalmente, es necesario replicar este estudio, con las mismas variables, en otras unidades de operación de la empresa de telecomunicaciones para examinar si los resultados del Churn son consistentes a nivel regional o nacional.

REFERENCIAS

Abou el Kassem, E., Hussein, S. A., Abdelrahman, A. M., & Alsheref, F. K. (2020). Customer churn prediction model and identifying features to increase customer retention based on user generated content. International Journal of Advanced Computer Science and Applications, 11(5). http://dx.doi.org/10.14569/IJACSA.2020.0110567

Ahuja, R., Chug, A., Gupta, S., Ahuja, P., & Kohli, S. (2020). Classification and clustering algorithms of machine learning with their applications. Nature-inspired computation in data mining and machine learning, 855, 225-248. https://doi.org/10.1007/978-3-030-28553-1_11

Alqahtani, A. Y., & Rajkhan, A. A. (2020). E-learning critical success factors during the covid-19 pandemic: A comprehensive analysis of e-learning managerial perspectives. Education sciences, 10(9), 216. https://doi.org/10.3390/educsci10090216

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 94, 290-301. https://doi.org/10.1016/j.jbusres.2018.03.003

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. Journal of International Technology and Information Management, 25(4), 6. https://doi.org/10.58729/1941-6679.1296

Arshad, S., Iqbal, K., Naz, S., Yasmin, S., & Rehman, Z. (2022). Hybrid System for Customer Churn Prediction and Retention Analysis via Supervised Learning. Computers, Materials & Continua, 72(3). https://doi.org/10.32604/cmc.2022.025442

Banabo E., Ndiomu K. (2023), Experience Marketing andCustomer Retention in the Nigerian Telecommunications Industry. International Journal of Entrepreneurship and Business Innovation 6(1), 54-67. https://doi.org/ 10.52589/IJEBIDBFJFL88

Basha, A. M., Rajaiah, M., Penchalaiah, P., Kamal, C. R., & Rao, B. N. (2020). Machine learning-structural equation modeling algorithm: The moderating role of loyalty on customer retention towards online shopping. International Journal of Emerging Trends in Engineering Research, 8, 1578-1585. https://doi.org/ 10.30534/ijeter/2020/17852020.

Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. Journal of Organizational Computing and Electronic Commerce, 19(2), 133-151. https://doi.org/10.1080/10919390902821291

Brmez, S., & Znidaršic, M. (2019). A Case of Churn Prediction in Telecommunications Industry. Ipsi Bgd Transactions on Internet Research, 15, SI.

Bugajev, A., Kriauzienė, R., Vasilecas, O., & Chadyšas, V. (2022). The Impact of Churn Labelling Rules on Churn Prediction in Telecommunications. Informatica, 33(2), 247-277. https://doi.org/10.15388/22-INFOR484

Cerda, A. A., & García, L. Y. (2021). Hesitation and refusal factors in individuals' decision-making processes regarding a coronavirus disease 2019 vaccination. Frontiers in public health, 9, 626852. https://doi.org/10.3389/fpubh.2021.626852

Chu, K. M. (2018). Mediating influences of attitude on internal and external factors influencing consumers' intention to purchase organic foods in China. Sustainability, 10(12), 4690. https://doi.org/10.3390/su10124690

Coeurderoy, R., Guilmot, N., & Vas, A. (2014). Explaining factors affecting technological change adoption: A survival analysis of an information system implementation. Management Decision, 52(6), 1082-1100. https://doi.org/10.1108/MD-10-2013-0540

Dahiya, K., & Bhatia, S. (2015, September). Customer churn analysis in telecom industry. In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions) (pp. 1-6). IEEE. https://doi.org/10.1109/ICRITO.2015.7359318

Dairo, M., Adekola, J., Apostolopoulos, C., & Tsaramirsis, G. (2021). Benchmarking strategic alignment of business and IT strategies: opportunities, risks, challenges and solutions. International Journal of Information Technology, 13, 2191-2197.https://doi.org/10.1007/s41870-021-00815-7

De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. Industrial Marketing Management, 99, 28-39. https://doi.org/10.1016/j.indmarman.2021.10.001

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. Neural Computing and Applications, 34(14), 11751-11768. https://doi.org/10.1007/s00521-022-07067-x

Dumitrache, A., Melian, D., & Stancu, S. (2020). Churn Prepaid Customer Profile in Romanian Postmodernism Telecommunications. Postmodern Openings/Deschideri Postmoderne, 11.

Estrada-Torres, B., del-Río-Ortega, A., Resinas, M., & Ruiz-Cortés, A. (2018, May). On the relationships between decision management and performance measurement. In International Conference on Advanced Information Systems Engineering (pp. 311-326). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91563-0_19

Ghode, D., Yadav, V., Jain, R., & Soni, G. (2020). Adoption of blockchain in supply chain: an analysis of influencing factors. Journal of Enterprise Information Management, 33(3), 437-456. https://doi.org/10.1108/JEIM-07-2019-0186

Grzybowski, L., Liang, J., & Zulehner, C. (2021). Bundling and consumer churn in telecommunications markets. Review of Network Economics, 20(1), 35-54. https://doi.org/10.1515/rne-2021-0032

Gupta, A., & Kumar, A. (2019, February). Information Security Using the Ensemble Approach of Steganography and Cryptography. In Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India. http://dx.doi.org/10.2139/ssrn.3350895

Haeruddin, M. (2017). Should I stay or should I go? Human Resource Information System implementation in Indonesian public organizations. European Research Studies Journal, 20(3A), 989-999. http://eprints.unm.ac.id/14128/1/2017-xx-3-a-68.pdf

Haridasan, V., Muthukumaran, K., & Hariharanath, K. (2023). Arithmetic Optimization with Deep Learning Enabled Churn Prediction Model for Telecommunication Industries. Intelligent Automation & Soft Computing, 35(3), https://doi.org/10.32604/iasc.2023.030628

Jaworski, M., Duda, P., & Rutkowski, L. (2017). New splitting criteria for decision trees in stationary data streams. IEEE transactions on neural networks and learning systems, 29(6), 2516-2529. http://dx.doi.org/10.1109/TNNLS.2017.2698204

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., & Bozkaya, B. (2018). Behavioral attributes and financial churn prediction. EPJ Data Science, 7(1), 41. https://doi.org/10.1140/epjds/s13688-018-0165-5

Kurmann, A., Lalé, E., & Ta, L. (2022). Measuring Small Business Dynamics and Employment with Private-Sector Real-Time Data. http://dx.doi.org/10.2139/ssrn.4204333

Lange, M., Mendling, J., & Recker, J. (2016). An empirical analysis of the factors and measures of Enterprise Architecture Management success. European Journal of Information Systems, 25(5), 411-431.https://doi.org/10.1057/ejis.2014.39

Łapczyński, M. (2014). Hybrid C&RT-Logit Models In Churn Analysis. Folia Oeconomica Stetinensia, 14(2), 37-52. https://doi.org/10.1515/foli-2015-0006

Lin, C. L., & Fan, C. L. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. Journal of Asian Architecture and Building Engineering, 18(6), 539-553. https://doi.org/10.1080/13467581.2019.1696203

Liu, P., Khan, Q., Mahmood, T., & Hassan, N. (2019). T-spherical fuzzy power Muirhead mean operator based on novel operational laws and their application in multi-attribute group decision making. Ieee Access, 7, 22613-22632. doi: 10.1109/ACCESS.2019.2896107.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048-3061. https://doi.org/10.1109/TKDE.2019.2962680

Maskew, M., Sharpey-Schafer, K., De Voux, L., Crompton, T., Bor, J., Rennick, M., ... &

Matuszelański, K., & Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 165-198. https://doi.org/10.3390/jtaer17010009

Melian, D. M., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Customer Churn Prediction in Telecommunication Industry. A Data Analysis Techniques Approach. Postmodern Openings, 13(1 Sup1), 78-104.

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. Ieee Access, 8, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905.

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287.

Mishra, A., & Reddy, U. S. (2017, December). A novel approach for churn prediction using deep learning. In 2017 IEEE international conference on computational intelligence and computing research (ICCIC) (pp. 1-4). IEEE.

Mustač, K., Bačić, K., Skorin-Kapov, L., & Sužnjević, M. (2022). Predicting player churn of a Free-to-Play mobile video game using supervised machine learning. Applied Sciences, 12(6), 2795. https://doi.org/10.3390/app12062795

Pejić Bach, M., Pivar, J., & Jaković, B. (2021). Churn management in telecommunications: Hybrid approach using cluster analysis and decision trees. Journal of Risk and Financial Management, 14(11), 544.https://doi.org/10.3390/jrfm14110544

Pisa, P. (2022). Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Scientific Reports, 12(1), 12715. https://doi.org/10.1038/s41598-022-16062-0

Pustokhina, I. V., Pustokhin, D. A., Aswathy, R. H., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706. https://doi.org/10.1016/j.ipm.2021.102706

Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-kenawy, E. S. M. (2023). Deep churn prediction method for telecommunication industry. Sustainability, 15(5), 4543. https://doi.org/10.3390/su15054543

Saleh, S., & Saha, S. (2023). Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. SN Applied Sciences, 5(7), 173.

Sharma, A., Gupta, D., Nayak, N., Singh, D., & Verma, A. (2022, April). Prediction of Customer Retention Rate Employing Machine Learning Techniques. In 2022 1st International Conference on Informatics (ICI) (pp. 103-107). IEEE. https://doi.org/10.1109/ICI53355.2022.9786903

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. International Journal of Information Management, 48, 238-253. https://doi.org/10.1007/s42452-023-05389-6

Sudharsan, R., & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science, 34(1), 1855-1876. https://doi.org/10.1080/09540091.2022.2083584

Sun, H., Yin, B., Amsah, N. F. B. B., & O'brien, B. A. (2018). Differential effects of internal and external factors in early bilingual vocabulary learning: The case of Singapore. Applied Psycholinguistics, 39(2), 383-411. https://doi.org/10.1017/S014271641700039X

Suryana, N., & Prasetio, R. T. (2020). Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting. IJCIT (Indonesian J. Comput. Inf. Technol), 6(1), 31–37.

Wang, X., Zhao, K., & Street, N. (2017). Analyzing and predicting user participations in online health communities: a social support perspective. Journal of medical Internet research, 19(4), e6834. https://doi.org/ 10.2196/jmir.6834

Wiemer, H., Drowatzky, L., & Ihlenfeldt, S. (2019). Data mining methodology for engineering applications (DMME)—A holistic extension to the CRISP-DM model. Applied Sciences, 9(12), 2407. https://doi.org/10.3390/app9122407

Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. IEEE Access, 9, 62118-62136. https://doi.org/10.1109/ACCESS.2021.3073776

Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475. https://doi.org/10.3390/jtaer17020024

Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475. https://doi.org/10.3390/jtaer17020024

Yoshikuni, A. C., Dwivedi, R., Dultra-de-Lima, R. G., Parisi, C., & Oyadomari, J. C. T. (2023). Role of Emerging Technologies in Accounting Information Systems for Achieving Strategic Flexibility through Decision-Making Performance: An Exploratory Study Based on North American and South American Firms. Global Journal of Flexible Systems Management, 1-20. https://doi.org/10.1007/s40171-022-00334-9

Zatonatskiy, D. (2019). Innovation Methods and Models of Personnel Security Management: Opportunities and Imperatives of Use at Ukrainian Enterprises. https://doi.org/10.21272/mmi.2019.1-24

Zdravevski, E., Lameski, P., Apanowicz, C., & Ślęzak, D. (2020). From Big Data to business analytics: The case study of churn prediction. Applied Soft Computing, 90, 106164. https://doi.org/10.1016/j.asoc.2020.106164