Santiago de Cali, 25 de mayo de 2024

Ingeniero
Diego Luis Linares
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Proyecto de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, me permito presentar como Director a: **Diego Fernando Mosquera Valencia** identificado con C.C. 94459188 del Proyecto de Grado denominado "Desarrollo de un modelo predictivo de abandono y segmentación de clientes para COLOMBIA INTERNET ISP : análisis del churn-rate", el cual será realizado por los estudiantes, Evelyn Jimeno Grisales con código 8992275, Oscar Mauricio Garrido Garcia con código 8992475, Deybison Antonio Garcia Lemus con código 8986469.

Atentamente,

Evelyn Jimeno Grisales

C.C. 1144068017 de Cali

Deybison Antonio Garcia Lemus

C.C. 1111763457 de Buenaventura

Oscar Mauricio Garrido García

C.C. 1098773037 de Bucaramanga

Diego Fernando Mosquera Valencia

C.C. 94459188 de Cali

Santiago de Cali, 25 de 05 de 2024

Ingeniero
Diego Luis Linares
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Proyecto Aplicado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el anteproyecto de Trabajo de Grado denominado "Desarrollo de un modelo predictivo de abandono y segmentación de clientes para COLOMBIA INTERNET ISP : análisis del churn-rate", el cual será realizado por los estudiantes Evelyn Jimeno Grisales con código 8992275, Oscar Mauricio Garrido García con código 8992475, Deybison Antonio Garcia Lemus con código 8986469 perteneciente a la Maestría en Ciencia de Datos, bajo la dirección del profesor Diego Fernando Mosquera Valencia.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este Anteproyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Evelyn Jimeno Grisales

C.C. 1144068017 de Cali

Oscar Mauricio Garrido García

C.C. 1098773037 de Bucaramanga

Deybison Antonio Garcia Lemus

C.C. 1111763457 de Buenaventura

Diego Fernando Mosquera Valencia

C.C. 94459188 de Cali

DESARROLLO DE UN MODELO PREDICTIVO DE ABANDONO Y SEGMENTACIÓN DE CLIENTES PARA COLOMBIA INTERNET ISP: ANÁLISIS DEL CHURN RATE

Nombre del o los estudiantes

Evelyn Jimeno Grisales cod. 8992275
Oscar Mauricio Garrido Garcia cod. 8992475
Deybison Antonio Garcia Lemus cod. 8986469

Anteproyecto del Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
Diego Fernando Mosquera Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 25 DE 2024

FICHA RESUMEN
ANTEPROYECTO DE TRABAJO DE GRADO

TÍTULO: DESARROLLO DE UN MODELO PREDICTIVO DE ABANDONO Y SEGMENTACIÓN DE CLIENTES PARA COLOMBIA INTERNET ISP: ANÁLISIS DEL CHURN-RATE

1. ÁREA DE TRABAJO:
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Investigación
3. ESTUDIANTE(S): Evelyn Jimeno Grisales, Oscar Mauricio Garrido Garcia, Deybison Antonio Garcia Lemus
4. CORREO ELECTRÓNICO: evelynjimeno@javerianacali.edu.co, osmagar26@javerianacali.edu.co, degale@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO:
   a) Evelyn Jimeno: Carrera 4 # 38-20, Las delicias, Cali, Valle del Cauca, Tel. 3116740918
   b) Oscar Garrido: Calle 200 # 22b - 645, Mirador de versalles, Floridablanca-Santander, Tel. 3186239513
   c) Deybison Garcia: Av 2B2 #73bis – Norte 98, Brisas de los álamos, Cali, Valle del Cauca, Tel. 3162920259
6. DIRECTOR: Diego Fernando Mosquera Valencia
7. VINCULACIÓN DEL DIRECTOR:
8. CORREO ELECTRÓNICO DEL DIRECTOR: dfmosquera@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): NA
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5): Churn-rate, datos, segmentación, clientes, predicción, telecomunicaciones, servicios.
13. FECHA DE INICIO: 25 de mayo de 2024
14. DURACIÓN ESTIMADA (En meses): 12
15. RESUMEN:

COLOMBIA INTERNET ISP, operando en el Valle del Cauca y Tolima, enfrenta una alta tasa de churn-rate en sus ventas mensuales. A pesar de diversas estrategias para mejorar la calidad del servicio y atención al cliente, la empresa no ha logrado reducir significativamente la tasa de deserción. Este proyecto busca utilizar técnicas de análisis predictivo para identificar patrones y causas, desarrollando estrategias basadas en datos para predecir el abandono de un cliente.

RESUMEN

El presente proyecto se centra en el desarrollo de un modelo predictivo orientado a la estimación del churn-rate en COLOMBIA INTERNET ISP, empresa de servicios de Internet con operaciones en los departamentos del Valle del Cauca y Tolima. La empresa registra una tasa de abandono promedio del 2.1% mensual. A través del uso de técnicas de análisis de datos y algoritmos de aprendizaje automático, se busca identificar patrones de comportamiento asociados a la pérdida de clientes y construir un modelo capaz de predecir dicho fenómeno con base en variables demográficas, comerciales y de uso del servicio.

La metodología aplicada sigue el enfoque CRISP-DM, e incluye etapas de comprensión del negocio, recolección y preparación de datos, modelado, evaluación e implementación. Se implementaron técnicas de imputación, transformación y depuración de datos, así como análisis univariado y de correlación para seleccionar variables relevantes. Posteriormente, se entrenaron y evaluaron distintos modelos de clasificación, tales como regresión logística, árboles de decisión, random forest y XGBoost, utilizando métricas de desempeño como sensibilidad, precisión, F1-score y área bajo la curva ROC (AUC) para seleccionar el modelo con mejor rendimiento.

Finalmente, se realizó una segmentación de clientes utilizando técnicas de clustering, con el propósito de identificar perfiles con distintos niveles de riesgo de abandono.

Este proyecto constituye una aplicación integral de la ciencia de datos al análisis del churn en el sector de telecomunicaciones, desde la preparación y exploración de datos hasta la evaluación comparativa de modelos predictivos y la segmentación de usuarios con base en su comportamiento.

Contenido

INTRODUCCIÓN 10
1. DEFINICIÓN DEL PROBLEMA 12
1.1 PLANTEAMIENTO DEL PROBLEMA 12
1.2 FORMULACIÓN DEL PROBLEMA 12
2 OBJETIVOS DEL PROYECTO 13
2.1 OBJETIVO GENERAL 13
2.2 OBJETIVOS ESPECÍFICOS 13
2.3 RESULTADOS ESPERADOS 13
3 ALCANCE 14
4 JUSTIFICACIÓN 16
5 MARCO DE REFERENCIA 17
5.1 MARCO TEORICO 17
5.2 TASA DE ABANDONO (CHURN-RATE) 17
5.3 VALOR DE VIDA DEL CLIENTE (CUSTOMER LIFETIME VALUE - LTV) 17
5.4 COEFICIENTE DE ASIMETRIA 17
5.5 TÉCNICAS DE ANÁLISIS PREDICTIVO 18
5.6 SEGMENTACIÓN DE CLIENTES 24
5.7 MATRIZ DE CONFUSIÓN 27
6 ANTECEDENTES 29
6.1 PROYECTO DE REFERENCIA 1 29
6.2 PROYECTO DE REFERENCIA 2 30
6.3 PROYECTO DE REFERENCIA 3 32
7 METODOLOGÍA 34
7.1 COMPRENSIÓN DEL NEGOCIO 37
7.2 OBTENCIÓN Y RECOLECCIÓN DE LOS DATOS 40
7.3 PREPARACION DE LOS DATOS 40
7.3.1 DESCRIPCIÓN INICIAL DEL CONJUNTO DE DATOS 41
7.3.2 IMPUTACIÓN DE VALORES FALTANTES 43
7.3.3 VERIFICACIÓN Y DEPURACIÓN FINAL DEL CONJUNTO DE DATOS 44
7.3.4 ANÁLISIS UNIVARIADO DE VARIABLES NUMÉRICAS 45
7.3.5 ANÁLISIS UNIVARIADO DE VARIABLES CATEGÓRICAS 47
7.3.6 ANÁLISIS DE CORRELACIÓN 50
7.4 MODELADO 53
7.4.1 MODELOS SELECCIONADOS PARA LA PREDICCIÓN DEL CHURN 54
7.4.2 MODELADO DE DATOS 54
7.5 EVALUACIÓN 55
7.5.1 EVALUACIÓN DEL MODELO CON MEJOR DESEMPEÑO (XGBOOST) 56

7.5.2 EVALUACIÓN DEL MODELO CON EL SEGUNDO MEJOR RENDIMIENTO (RANDOM FOREST) 58
7.5.3 EVALUACIÓN DEL MODELO CON DESEMPEÑO INTERMEDIO (LOGISTIC REGRESSION) 59
7.5.4 EVALUACIÓN DEL MODELO CON MENOR RENDIMIENTO (DECISION TREE) 61
7.5.5 AJUSTE DE HIPERPARAMETROS PARA EL MEJOR MODELO (XGBOOST) 62
7.5.6 SEGMENTACIÓN DE CLIENTES 65
7.6 IMPLEMENTACIÓN 70
8 CONCLUSIONES Y TRABAJOS FUTUROS 72
8.1 CONCLUSIONES 72
8.2 TRABAJOS FUTUROS 74
9 REFERENCIAS BIBLIOGRÁFICAS 75

LISTA DE FIGURAS

Figura 1. Formula arboles de decisión 20
Figura 2. Ejemplo arboles de decisión 20
Figura 3. Formula Random Forest 21
Figura 4. Arquitectura XGBoost 23
Figura 5. Algoritmo K-Means 25
Figura 6. Ejemplo de dendograma 26
Figura 7. Matriz de confusión 27
Figura 8. Curva ROC 29
Figura 9. Fases del proceso de la metodología CRISP-DM 34
Figura 10.Comportamiento churn-rate mensual 2024 – Colombia Internet 38
Figura 11. Duración clientes retirados 39
Figura 12. Motivos terminaciones 40
Figura 13. Distribución variable antigüedad 46
Figura 14. Distribución variable precio del plan 47
Figura 15. Distribución variable total pagado 47
Figura 16. Distribución variable departamento 48
Figura 17. Distribución variable tipo de cliente 48
Figura 18. Distribución variable estrato 49
Figura 19. Distribución variable tipo de plan 49
Figura 20. Distribución variable canal de venta 50
Figura 21. Distribución variable municipio 50
Figura 22. Matriz de correlación variables numéricas 51
Figura 23. Matriz de correlación variables categóricas 52
Figura 24. Matriz de confusión XGBoost 57
Figura 25. Curva ROC XGBoost 57
Figura 26. Matriz de confusión Random Forest 59
Figura 27. Curva ROC Random Forest 59
Figura 28. Matriz de confusión Logistic Regression 60
Figura 29. Curva ROC Logistic Regression 61
Figura 30. Matriz de confusión Decision Tree 62
Figura 31. Curva ROC Decision Tree 62
Figura 32. Curva ROC XGBoost base vs Ajustado 65
Figura 33. Método del codo para determinar número óptimo de clusters 67
Figura 34. Segmentación de clientes por antigüedad y probabilidad de churn 68

Figura 35. Análisis por componentes principales multivariado 69
Figura 36. Flujo de implementación 71

LISTA DE TABLAS

Tabla 1. Diccionario de datos y tipos de variables 41
Tabla 2. Cantidad y proporción de valores nulos por variable 42
Tabla 3. Análisis descriptivo variables numéricas 42
Tabla 4. Análisis descriptivo variables categóricas 43
Tabla 5. Distribución variable objetivo 43
Tabla 6. Cantidad de valores nulos y tipo por cada variable 45
Tabla 7. Prueba Kolmogorov-Smirnov 46
Tabla 8. Prueba Kruskal-Wallis 52
Tabla 9. Variables dataset final 53
Tabla 10. Resultados por modelo 55
Tabla 11. Resultados de métricas para XGBoost ajustado 64
Tabla 12. Resultados de métricas para XGBoost ajustado vs XGBoost base 64
Tabla 13. Resultado de predicciones del modelo en el conjunto de test 66
Tabla 14. Perfilamiento de clientes 70

INTRODUCCIÓN

El mercado de servicios de Internet en Colombia se caracteriza por su alta competitividad. Según datos del MINTIC a febrero de 2023, el país contaba con 3.393 empresas registradas que brindan servicio de Internet fijo. Aunque las compañías más grandes son ampliamente reconocidas, muchas de estas empresas tienen un enfoque local y operan únicamente en un municipio o en zonas específicas.

En este contexto, COLOMBIA INTERNET ISP es un proveedor de servicios de internet (ISP) el cual inició operaciones el 1 de octubre de 2019 en el Valle del Cauca y en el año 2022 en el departamento del Tolima. La compañía ofrece a sus clientes el servicio de Internet fijo por medio de fibra óptica, con planes simétricos desde las 100 Mb hasta las 600 Mb. Como parte de los compromisos sociales, la compañía ofrece el servicio de Internet de forma gratuita a todas las escuelas por donde se cuente con cobertura, actualmente se están beneficiando 500 escuelas públicas con aproximadamente 247.000 estudiantes de estrato 1, 2 y 3, la empresa ha buscado desde el inicio una diferenciación de los operadores tradicionales, compitiendo con calidad de servicio en lugar de cláusulas, por lo tanto, los clientes no cuentan con tiempos de permanencia obligatorios, es decir, se puede cancelar el servicio en cualquier momento, además el primer mes es gratuito, lo anterior hace parte de la promesa de valor.

A pesar de contar con una infraestructura robusta y una base de 104.551 clientes residenciales, la compañía ve con preocupación el índice de abandono de clientes (churn-rate) que se ha venido presentando desde hace varios meses. Esta tasa se ha mantenido un poco por encima del promedio de la industria (1.5 – 2% mensual) pero nunca se ha analizado en detalle, en el momento la cifra es 2.1% mensual, lo cual representa alrededor de 1470 clientes que abandonan mensualmente el servicio. Como resultado, la empresa enfrenta desafíos significativos en la fidelización de sus clientes.

Para la compañía, la reducción del churn-rate es un aspecto clave para garantizar la sostenibilidad estratégica y solidez empresarial. A través de un análisis cuantitativo es posible identificar las variables internas asociadas a la pérdida de clientes, diseñar estrategias más efectivas para aumentar su satisfacción y prolongar su ciclo de vida. Reducir el churn-rate impacta cinco aspectos claves corporativos: ingresos recurrentes, costos de adquisición, reputación de marca, eficiencia operativa y satisfacción en la experiencia de los clientes.

La necesidad de abordar este problema es imperiosa, ya que un alto churn-rate puede indicar problemas en la calidad del servicio, la insatisfacción de los clientes o una baja competitividad en el mercado. Asimismo, comprender y mejorar el valor de vida del cliente (LTV) es esencial para optimizar los ingresos netos generados a partir de la relación con el cliente. El desafío radica en predecir el churn-rate, utilizando técnicas de análisis de datos y machine learning.

Este proyecto sigue la metodología CRISP-DM y desarrolla un modelo predictivo que permite estimar el riesgo de abandono de un cliente según sus datos correspondientes con el servicio adquirido y el comportamiento en el transcurso del tiempo. A través del análisis de sus

características demográficas, datos históricos de cancelaciones, comportamientos de pago y solicitudes de soporte, se identificaron patrones de comportamiento y se segmentaron los clientes según su riesgo de churn-rate. Esto le permitirá a COLOMBIA INTERNET ISP diseñar estrategias de retención de clientes y disminuir el promedio mensual de churn-rate.

Los resultados de este proyecto incluyen un modelo de estimación del churn-rate que genera alertas tempranas, un documento de análisis detallado sobre los comportamientos de los clientes retirados, y dashboards de visualización de indicadores clave de rendimiento (KPIs) para un seguimiento periódico. Con los resultados de este proyecto, COLOMBIA INTERNET ISP podrá contar también con las herramientas que le permitirán tomar decisiones informadas, mejorar la satisfacción del cliente, asegurar su competitividad y crecimiento sostenible en el mercado.

Finalmente, el desarrollo del proyecto permitió validar la utilidad de los enfoques de ciencia de datos para enfrentar retos reales del sector de telecomunicaciones. La combinación de análisis exploratorio, modelado predictivo y segmentación facilitó una comprensión más profunda del fenómeno de abandono de clientes. Las conclusiones obtenidas confirman que el uso de modelos de clasificación y técnicas de clustering aporta herramientas valiosas para la toma de decisiones basadas en evidencia, abriendo la posibilidad de implementar estrategias más efectivas de retención y consolidando una cultura organizacional orientada al análisis de datos.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La tasa de abandono de clientes (churn-rate) representa para COLOMBIA INTERNET ISP un problema significativo, ya que evidencia la pérdida de usuarios sin identificar con precisión los factores que la provocan. Entre 2021 y 2024 se registraron 38,210 terminaciones de servicio, lo que equivale a un churn-rate de un 2.1% mensual y genera una pérdida económica considerable para la empresa. Ante este desafío, se requiere un enfoque basado en datos que permita identificar características, patrones y variables relevantes asociadas al abandono, con el fin de mitigar sus efectos y fortalecer la sostenibilidad del negocio.

Este proyecto busca utilizar técnicas de análisis predictivo para identificar características y clasificar clientes según el churn-rate, entregando insumos a la compañía para que pudiesen desarrollar estrategias basadas en datos para predecir y disminuir el abandono de clientes.

1.2 FORMULACIÓN DEL PROBLEMA

Por medio de las siguientes preguntas, se pretende plantear de forma más precisa el problema y sus posibles dificultades.

¿Cómo puede COLOMBIA INTERNET ISP estimar/predecir la tasa de abandono de sus clientes haciendo uso de la ciencia de datos?, ¿Cuáles son las diferentes tipologías de los clientes que abandonan el servicio?, ¿Cuáles son los principales factores que impactan la tasa de churn-rate en COLOMBIA INTERNET ISP?, ¿Cómo se distribuye la tasa de churn-rate entre diferentes segmentos de clientes?, ¿Cuál es el mejor modelo de Machine Learning que permite ayudar a entender (predecir) el riesgo de fuga?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo para predecir el riesgo de abandono de clientes que permita generar alertas tempranas y facilitar la toma de decisiones estratégicas, utilizando técnicas de análisis predictivo.

2.2 OBJETIVOS ESPECÍFICOS

- Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.

- Explorar y desarrollar diferentes modelos de predicción de churn-rate.

- Evaluar los modelos de aprendizaje desarrollados para definir el mejor modelo de acuerdo con las preguntas de la investigación, considerando la optimización de sus hiperparámetros.

- Tipificar a los clientes de acuerdo con su nivel de churn-rate, empleando las variables características que los describen.

2.3 RESULTADOS ESPERADOS

- **Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.**
    - Informe con análisis basado en la exploración de datos que contiene los comportamientos diferenciales de los clientes retirados.

- **Explorar y desarrollar diferentes modelos de predicción de churn-rate**
    - Modelos predictivos generados en lenguajes como R o Python para la aplicación de algoritmos de aprendizaje automático o machine learning.

- **Evaluar los modelos de aprendizaje desarrollados para definir aquel con mejores resultados.**
    - Documento con los resultados de cada uno de los modelos, identificando aquel con las mejores métricas.

- **Tipificar a los clientes de acuerdo con su nivel de riesgo de fuga, empleando las variables características que los describen.**
    - Documento con el modelo de clasificación de clientes, que se realizó mediante técnicas de clustering y análisis de datos multivariado para identificar tipologías de clientes.

3 ALCANCE

El proyecto se centró en desarrollar y aplicar técnicas de análisis predictivo para estimar la probabilidad de abandono de clientes en COLOMBIA INTERNET ISP. La investigación se dirigió a usuarios residenciales de los departamentos del Valle del Cauca y Tolima, donde la empresa presta principalmente el servicio. A través del análisis de los datos disponibles, se identificaron patrones y causas asociadas al churn-rate, lo que permitió entrenar un modelo capaz de anticipar abandonos futuros. Para ello, se utilizaron bases de datos que comprendieron el período de octubre de 2019 a octubre de 2024.

El alcance del proyecto se desglosa para cada objetivo específico de esta manera:

- **Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.**
  Exploración de las bases de datos, limpieza y generación de nuevas variables a partir de los datos.

- **Explorar y desarrollar diferentes modelos de predicción de churn-rate.**
  Se exploraron al menos 4 modelos de predicción para determinar el mejor.

- **Evaluar los modelos de aprendizaje desarrollados para definir el mejor modelo de acuerdo con las preguntas de la investigación, considerando la optimización de sus hiperparámetros.**
  Establecer la comparativa de modelos supervisados, optimizando sus hiperparámetros y seleccionando el de mejor desempeño según métricas como F1-Score entre otras.

- **Tipificar a los clientes de acuerdo con su nivel de riesgo de fuga, empleando las variables características que los describen.**
  Analizar el total de clientes que se han retirado analizando las bases de datos de 38210 abandonos y lograr la identificación de patrones de abandono para hacer la segmentación por grupos.

Se diseñó un dashboard de visualización de la información que facilita la toma de decisiones relacionadas con el churn-rate. Este incluye datos sobre la ubicación geográfica, el tipo de clientes, la cantidad de solicitudes de soporte y los motivos de cancelación, lo que permite realizar un seguimiento de la evolución de la tasa de abandono e identificar tendencias relevantes.

El estudio no abordó el impacto financiero a largo plazo ni evaluó los efectos de la aplicación del modelo en el tiempo. Sin embargo, se reconoce que futuras investigaciones podrían extender el análisis a otras áreas de la compañía, incorporando variables adicionales y evaluando el impacto de nuevas estrategias de fidelización.

En el desarrollo de este proyecto se identificaron posibles amenazas que pudieron afectar su progreso y resultados. Entre ellas se destacaron las siguientes:

- La accesibilidad a los datos históricos necesarios para el análisis puede ser limitada. Esto incluye datos de cancelaciones, comportamientos de pago, y solicitudes de soporte. Cualquier restricción en el acceso o retraso en la entrega de esta información habría podido afectar el cronograma y la calidad del análisis.

- La precisión, integridad y consistencia de los datos disponibles fueron factores determinantes para la confiabilidad de los resultados. La presencia de registros incompletos, erróneos o desactualizados podía comprometer la efectividad del modelo predictivo y los procesos de segmentación.

- La disponibilidad de recursos tecnológicos adecuados, tanto en software como en hardware, fue esencial para el procesamiento de grandes volúmenes de datos. Limitaciones en este aspecto podían restringir el desarrollo óptimo de los modelos de aprendizaje supervisado y el análisis requerido.

No obstante, se formalizó un acuerdo de confidencialidad entre COLOMBIA INTERNET ISP y los integrantes del proyecto, garantizando el acceso controlado a la información sensible y el cumplimiento de los principios éticos en el manejo de los datos.

Bajo estas condiciones, el proyecto aportó al fortalecimiento de la capacidad analítica de COLOMBIA INTERNET ISP, proporcionando herramientas para anticipar el abandono de clientes y mejorar su comprensión del fenómeno del churn-rate, en consonancia con sus objetivos de sostenibilidad y competitividad en el mercado.

4 JUSTIFICACIÓN

El proyecto se enfocó en el desarrollo de un modelo predictivo de churn-rate y en una estrategia de segmentación de clientes para COLOMBIA INTERNET ISP, empresa que ha enfrentado una alta tasa de abandono mensual del 2.1%. Reducir este indicador es crucial para mejorar la fidelización, optimizar los recursos y fortalecer la competitividad en el mercado. En el último año, la compañía registró la pérdida de 19.113 clientes, lo que evidencia la magnitud del problema y la necesidad de implementar soluciones basadas en el análisis de datos.

Ante este panorama, el desarrollo de este proyecto resultó útil porque abordó una de las principales preocupaciones de la empresa: la alta tasa de abandono de clientes. La construcción del modelo predictivo y la segmentación de usuarios proporcionaron insumos valiosos para anticipar el riesgo de cancelación y caracterizar distintos perfiles de clientes. Estos resultados fortalecieron la capacidad analítica de COLOMBIA INTERNET ISP y aportaron una base técnica para la toma de decisiones sustentadas en datos, en un entorno altamente competitivo.

Por consiguiente, el proyecto fue viable gracias a la disponibilidad de datos relevantes, el acceso a herramientas tecnológicas avanzadas y el conocimiento teórico-práctico del equipo de trabajo. La combinación de estos elementos permitió la implementación de un modelo predictivo orientado a anticipar el abandono de clientes. Adicionalmente, la capacidad de estimar comportamientos asociados al churn-rate facilitó la identificación de clientes con mayor probabilidad de cancelación, aportando información clave para el análisis estratégico de este fenómeno.

Asimismo, el proyecto tuvo un impacto positivo al mejorar la comprensión, por parte de COLOMBIA INTERNET ISP, de los patrones de comportamiento de sus clientes y de las posibles causas de abandono. Las técnicas de machine learning empleadas permitieron detectar relaciones significativas entre variables y caracterizar con mayor precisión los perfiles de riesgo. La implementación de modelos predictivos proporcionó herramientas útiles para fortalecer el análisis del churn-rate desde una perspectiva basada en datos.

Este proyecto no sólo representó una respuesta técnica a un desafío crítico, sino que constituyó una oportunidad para COLOMBIA INTERNET ISP de avanzar en la incorporación de enfoques analíticos para la toma de decisiones. La capacidad de estructurar, analizar y aprovechar los datos disponibles evidenció cómo la ciencia de datos puede contribuir a enfrentar retos complejos y a generar oportunidades de mejora operativa en un entorno altamente competitivo.

5 MARCO DE REFERENCIA

5.1 MARCO TEORICO

El presente marco teórico se centra en los conceptos fundamentales necesarios para abordar el problema de una alta tasa de churn-rate, proporcionando una base para comprender los desafíos, oportunidades asociadas con la reducción del churn-rate y la mejora del valor de vida del cliente. Además, establece el contexto necesario para desarrollar estrategias basadas en datos que aseguren el crecimiento sostenible de COLOMBIA INTERNET ISP.

5.2 TASA DE ABANDONO (CHURN-RATE)

La tasa de abandono o churn-rate, es una métrica que mide el porcentaje de clientes que dejan de utilizar los servicios de una empresa en un período específico. Generalmente, lo evalúan principalmente empresas que ofrecen servicios por suscripción, ya que esas organizaciones necesitan construir una relación duradera y contractual con sus clientes. En este contexto, es una tasa esencial para identificar problemas de satisfacción del cliente, calidad del servicio o alta competencia en el mercado. Para empresas de servicios como los proveedores de internet (ISP), el churn-rate es esencial para evaluar la estabilidad financiera y el crecimiento sostenible. Reducirlo ayuda a mantener la base de clientes, reducir costos asociados con la adquisición de nuevos suscriptores y mejorar la rentabilidad. [1]

Este valor se calcula dividiendo la cantidad de clientes perdidos en un periodo de tiempo entre la cantidad de clientes totales al inicio del mismo periodo y multiplicarlo por 100.

$$\text{Churn-rate} = \frac{\text{Cantidad de clientes perdidos en un periodo de tiempo}}{\text{Cantidad de clientes totales al inicio del mismo periodo}} \times 100$$

5.3 VALOR DE VIDA DEL CLIENTE (CUSTOMER LIFETIME VALUE - LTV)

El valor de vida del cliente (LTV) se define como el ingreso neto total que una empresa espera generar del mismo durante su relación con la empresa. Se calcula considerando los ingresos, costos y la duración de la relación. Un análisis detallado permite a las empresas tomar decisiones informadas sobre inversiones en retención, estrategias de marketing y mejoras en el servicio. Medir y predecir el LTV durante la fase activa de la relación con el cliente es un desafío, pero es crucial para maximizar el retorno de inversión y asegurar la sostenibilidad a largo plazo. [2]

5.4 COEFICIENTE DE ASIMETRIA

El coeficiente de asimetría de Pearson es una medida que estima el grado de simetría de una distribución de datos, comparando la posición relativa de la media y la mediana en relación con la dispersión. Se utiliza para identificar si los datos están sesgados hacia la derecha (asimetría positiva) o hacia la izquierda (asimetría negativa). [3]

La fórmula clásica es:

$$A_s = \frac{3(\bar{x} - Me)}{S}$$

Donde:

- $\bar{x}$ es la **media aritmética**
- Me es el valor central de la distribución
- S es la **desviación estándar**

Este indicador toma el valor **0** en distribuciones perfectamente simétricas. Valores **positivos** indican una asimetría hacia la derecha (la media está por encima de la mediana), mientras que valores **negativos** indican asimetría hacia la izquierda (la media está por debajo de la mediana).

5.5 TÉCNICAS DE ANÁLISIS PREDICTIVO

El análisis predictivo utiliza técnicas estadísticas y algoritmos de machine learning para analizar datos históricos y predecir futuros eventos. En el contexto del churn-rate, el análisis predictivo puede identificar patrones y factores que contribuyen al abandono de clientes, permitiendo abordar el problema desde distintos enfoques y garantizar un buen desempeño en las predicciones. Entre estas técnicas y algoritmos se encuentran la regresión logística, árboles de decisión, random forest, support vector machine y redes neuronales. Estos modelos analizan variables como el comportamiento de pago, interacciones con el soporte al cliente, y características demográficas para predecir la probabilidad de que un cliente abandone el servicio. [1]

A continuación, se presenta un acercamiento a las técnicas anteriormente mencionadas [4]:

- **Regresión logística**

  Es una técnica de aprendizaje supervisado utilizada para problemas de clasificación binaria. Es útil para predecir la probabilidad de un evento binario, como la de que un cliente abandone el servicio. La simplicidad y la interpretabilidad del modelo lo hacen una opción popular en análisis predictivo.

  - **Ventajas**:

- La regresión logística es más fácil de implementar, interpretar y muy eficiente de entrenar.
- No solo proporciona una medida de cuán apropiado es un predictor (tamaño del coeficiente), sino también su dirección de asociación (positiva o negativa).
- Puede interpretar los coeficientes del modelo como indicadores de la importancia de las características.
- Buena precisión para muchos conjuntos de datos simples y funciona bien cuando el conjunto de datos es linealmente separable.

- **Desventajas**:
  - La principal limitación de la regresión logística es la suposición de linealidad entre la variable dependiente y las variables independientes.
  - Los problemas no lineales no se pueden resolver con regresión logística porque tiene una superficie de decisión lineal. Los datos linealmente separables rara vez se encuentran en escenarios del mundo real.
  - La regresión logística requiere multicolinealidad media o nula entre variables independientes.

- **Árboles de decisión**

Un árbol de decisiones puede considerarse un mapa de un proceso de razonamiento. Utiliza una estructura similar a la de un árbol para describir un conjunto de datos y las soluciones se pueden visualizar siguiendo diferentes rutas a través del árbol. [5]

Es un conjunto jerárquico de reglas que explican la forma en que un gran conjunto de datos se puede dividir en particiones de datos más pequeñas. Cada vez que se produce una división, los componentes de las particiones resultantes se vuelven cada vez más similares entre sí con respecto al objetivo.

- **Ventajas**:
  - Son simples de entender y de interpretar.
  - No requiere una preparación de los datos demasiado exigente (aunque la implementación de Scikit-Learn no soporta valores nulos).
  - Se puede trabajar tanto con variables cuantitativas como cualitativas

- **Desventajas**:
  - Los árboles de decisión tienden al sobre entrenamiento, especialmente cuando el número de características predictivas es alto.
  - Son inestables: cualquier pequeño cambio en los datos de entrada puede suponer un árbol de decisión completamente diferente.
  - No se puede garantizar que el árbol generado sea el óptimo.
  - Si hay clases dominantes es fácil que los árboles se generen sesgados, por lo que se recomienda balancear el conjunto de datos antes de entrenar el modelo.

El aprendizaje de árboles de decisión suele ser el más adecuado para problemas con las siguientes características:

- Los patrones se describen mediante un conjunto fijo de atributos xj; j = 1, 2, …, n, y cada atributo xj toma una pequeña cantidad de posibles valores disjuntos (categóricos o numéricos) vlxj; l = 1, 2, …, dj.

- La variable de salida y es una función de valor booleano (problemas de clasificación binaria) definida sobre el conjunto S de patrones {s(i)} ≡ {x(i)} ; i = 1, 2, …, N. Es decir, y toma los valores yq ; q = 1, 2. Si asumimos y1 ≡ 0 e y2 ≡ 1, entonces y : S → [0, 1].

- Los datos de entrenamiento se describen mediante el conjunto de datos D de N patrones con los resultados observados correspondientes:

$$\mathscr{D} = \left\{ s^{(i)}, y^{(i)} \right\} = \left\{ \mathbf{x}^{(i)} y^{(i)} \right\}; i = 1, 2, \ldots, N$$

Figura 1. Formula arboles de decisión

Ejemplo: El árbol de decisiones que se muestra en la siguiente figura, se creó a partir de un conjunto de datos meteorológicos diminutos y totalmente ficticios que supuestamente se refieren a las condiciones adecuadas para jugar al tenis. La muestra se muestra en la Tabla 8.1. Las variables de entrada son: x1 = Pronóstico, x2 = Temperatura, x3 = Humedad y x4 = Viento; y la variable objetivo y = PlayTennis. La tarea consiste en predecir el valor de PlayTennis para una mañana de sábado arbitraria, basándose en los valores de sus atributos.

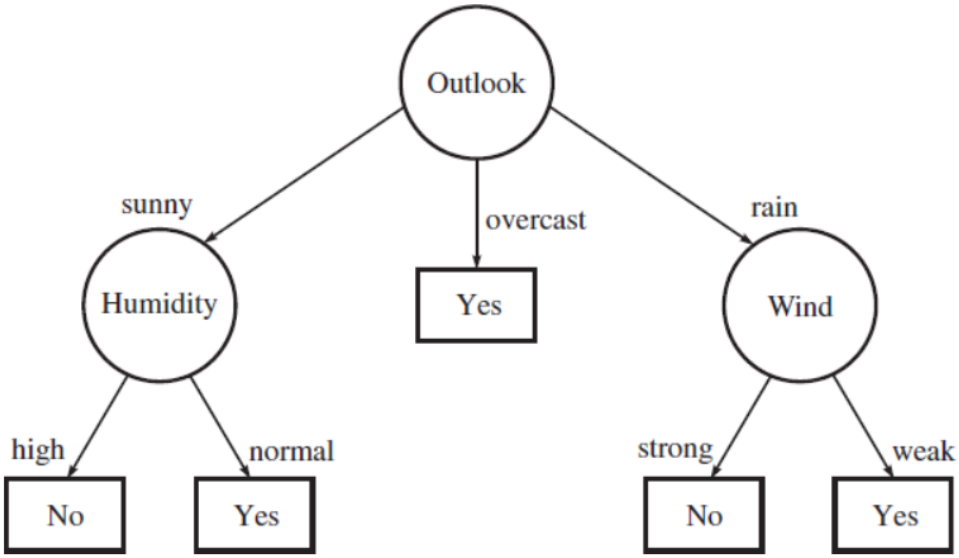

Figura 2. Ejemplo arboles de decisión

- **Random forest (Bosques aleatorios)**

  Es un modelo de aprendizaje automático que combina múltiples clasificadores en forma de árboles de decisión. Cada árbol en el bosque se construye de manera independiente a partir de un conjunto de datos muestreado mediante bootstrap y utilizando un subconjunto aleatorio de características en cada división de nodo. Esto introduce variabilidad en los árboles individuales, reduciendo la correlación entre ellos y mejorando el desempeño del modelo. [6]

  - **Ventajas:**
    - Combina múltiples árboles de decisión para mejorar la precisión general del modelo.
    - Utiliza el promedio de varios árboles, reduciendo la posibilidad de sobreajuste en comparación con un solo árbol de decisión.
    - Es capaz de manejar grandes conjuntos de datos con muchas características y detectar interacciones complejas entre ellas.

  - **Desventajas:**
    - Debido a la combinación de muchos árboles, es complicado entender cómo el modelo llega a sus decisiones.
    - Necesita más tiempo y recursos para entrenar y realizar predicciones debido al uso de múltiples árboles.
    - Al ser un conjunto de modelos, el tiempo de predicción puede ser más lento en comparación con modelos más simples.

  Una de las propiedades clave del modelo es que su error de generalización converge casi seguramente a un límite conforme aumenta el número de árboles en el bosque. Este error depende de:

  - La fortaleza de los árboles individuales: su capacidad para realizar predicciones precisas.
  - La correlación entre los árboles: una menor correlación mejora el desempeño global.

  El uso de una selección aleatoria de características para dividir los nodos de cada árbol reduce la correlación entre ellos, logrando tasas de error competitivas con métodos como Adaboost (Freund y Schapire), pero con mayor robustez frente al ruido. [6]

  Para una entrada x, la predicción de un Random Forest se calcula como el promedio (regresión) o el voto mayoritario (clasificación) de las predicciones individuales de los árboles [6]:

  $$f(x) = 1/\mathrm{T} \sum\_(t = 1)\hat{}T \blacksquare h(x, \theta\_T)$$

  Figura 3. Formula Random Forest

  Donde:

- T: Número total de árboles en el bosque.
- $h(x, \theta\_T)$: Predicción del árbol t para la entrada x, basado en el vector aleatorio θ_T, que incluye datos y características seleccionadas aleatoriamente.

- **Support vector machine**

  Técnica de aprendizaje no supervisado que agrupa a los clientes en segmentos homogéneos basados en sus características. Esta técnica es útil para identificar patrones y segmentos de clientes con comportamientos similares.

  - **Ventajas:**
    - Funciona bien con datos que tienen muchas características, especialmente cuando las muestras no son linealmente separables.
    - Encuentra el hiperplano óptimo que maximiza la separación entre clases, lo que es útil en problemas de clasificación binaria.
    - Permite la transformación de los datos a espacios de mayor dimensión mediante funciones de kernel, lo que facilita la separación de datos no lineales.

  - **Desventajas:**
    - Requiere una cuidadosa selección de parámetros, como el valor de C y el tipo de kernel, lo que puede afectar su rendimiento.
    - El tiempo de entrenamiento puede ser muy largo cuando se trabaja con grandes conjuntos de datos.
    - Los resultados no son tan interpretables como los de otros modelos y ajustar los parámetros para obtener el mejor rendimiento puede ser complicado.

- **Extreme Gradient Boosting (XGBOOST)**

  La tecnología XGBoost es una herramienta escalable de optimización de árboles en machine learning que ha ganado amplia popularidad en disciplinas relacionadas con el análisis de datos. La técnica XGBoost se desarrolló como una máquina de boosting basada en gradientes única en su tipo, utilizada especialmente en problemas de regresión y clasificación mediante árboles. El concepto de "boosting" es el núcleo de XGBoost, combinando la predicción de modelos débiles con métodos de entrenamiento aditivos para formar un modelo sólido. Este enfoque no solo ayuda a prevenir el sobreajuste, sino que también mejora la precisión matemática del modelo.
  La arquitectura de XGBoost se representa en la Figura 4, donde las funciones objetivo se simplifican permitiendo que los términos de predicción y regularización se combinen, al tiempo que se preserva la mayor velocidad de procesamiento posible. La función general para realizar predicciones en el paso *p* se define como sigue en la Ecuación (1):

$$f_i^{(p)} = \sum_{k=1}^{p} f_k(x_i) = f_i^{(p-1)} + f_p(x_i) \qquad (1)$$

donde fp(xi) representa el modelo en el paso p, fi(p) corresponde a la predicción en el paso p, fi(p−1) a la predicción en el paso previo (p−1) y xi denota las características de entrada.
Para controlar el sobreajuste de manera efectiva sin comprometer la velocidad del cálculo matemático, XGBoost emplea la fórmula analítica que se muestra en la Ecuación (2) para estimar la "bondad" del modelo respecto a la función original:

$$Objective^{(p)} = \sum_{k=1}^{n} l(\bar{y}_i,\ y_i) + \sum_{k=1}^{p} \sigma(f_i) \qquad (2)$$

donde l es la función de pérdida, n representa el número de observaciones utilizadas, y σ es el término de regularización, descrito en la Ecuación (3):

$$\sigma(f) = \mathrm{Y}T + 0.5\lambda\omega^2 \qquad (3)$$

En esta ecuación, ω representa los valores del vector en las hojas, Y corresponde a la pérdida mínima requerida para dividir el nodo hoja aún más, y λ denota los parámetros de regularización. [7]

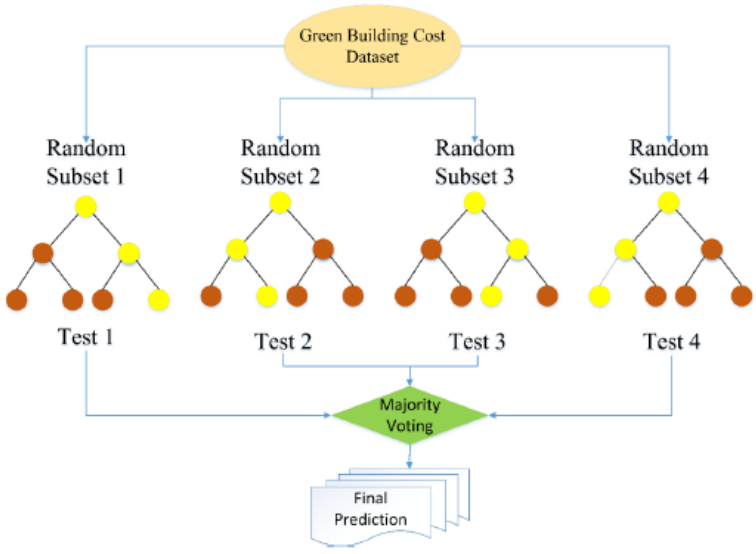

Figura 4. Arquitectura XGBoost

- **Ventajas:**
  - Ofrece un alto nivel de precisión
  - Maneja de forma eficiente los valores faltantes, sin necesidad de imputación previa.
  - Incorpora regularización (L1 y L2), lo que ayuda a reducir el riesgo de sobreajuste.

- Está optimizado para trabajar rápidamente, utilizando técnicas de paralelización y aprovechando mejor los recursos computacionales.
- Permite identificar la importancia de las variables, lo que facilita, en parte, la interpretación del modelo.

- **Desventajas:**
  - Requiere una adecuada configuración de parámetros como la profundidad de los árboles, la tasa de aprendizaje o el número de iteraciones, lo que puede resultar complejo sin experiencia.
  - Sus modelos no son tan fácilmente interpretables como otros métodos más simples, como la regresión lineal o los árboles de decisión individuales.
  - A pesar de contar con mecanismos de regularización, puede sobreajustarse si no se controla correctamente su configuración.

5.6 SEGMENTACIÓN DE CLIENTES

La segmentación de clientes consiste en agrupar a los clientes en segmentos homogéneos basados en características demográficas, comportamientos de uso y riesgo de churn-rate. Las técnicas de clustering, como K-means y análisis de conglomerados jerárquicos, son comunes para este propósito. Una segmentación efectiva permite a las empresas diseñar estrategias específicas y personalizadas para cada grupo. [8]

A continuación, se presenta un acercamiento a las técnicas de clustering anteriormente mencionadas:

- **K-Means**

  El algoritmo K-means es una una técnica de aprendizaje no supervisado utilizada para dividir un conjunto de datos en K grupos o clusters, basándose en las características intrínsecas de los datos. Su propósito es dividir los datos en K grupos predefinidos, asignando cada elemento al grupo cuya media es la más cercana.

  El proceso inicia con la selección de K centros iniciales (pseudocentros). Luego, cada punto del conjunto de datos se asigna al pseudocentro más cercano, formando grupos. A continuación, se actualiza la ubicación de cada pseudocentro para que se sitúe en el centro de todos los puntos de su grupo. Este procedimiento se repite iterativamente hasta que los grupos dejen de cambiar.

  En la figura 5 se observa el proceso de funcionamiento del algoritmo K-means.

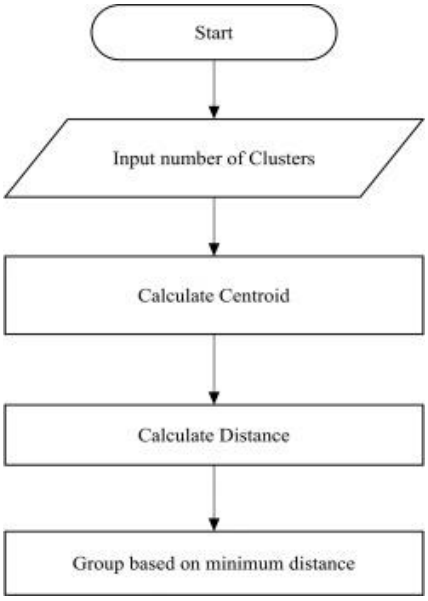

Figura 5. Algoritmo K-Means

La elección del número óptimo de clusters (K) es crucial y puede determinarse mediante métodos como el "codo" o el análisis de la silueta, que evalúan la coherencia interna y la separación entre clusters. Una buena agrupación se caracteriza por una menor distancia intra-grupo y una mayor distancia inter-grupos. [9]

Es importante tener en cuenta que el algoritmo K-means puede converger a mínimos locales y su rendimiento depende de la inicialización de los centroides. Por lo tanto, es común ejecutar el algoritmo varias veces con diferentes inicializaciones y seleccionar la partición que presente el menor error cuadrático dentro de los clusters.

- **Análisis de conglomerados jerárquicos**

El análisis de conglomerados jerárquicos es una técnica de agrupamiento que organiza los datos en una estructura jerárquica basada en su similitud, representada comúnmente mediante un dendrograma (Fi el cual muestra la relación entre los datos a diferentes niveles de agrupación). Este método no requiere especificar previamente el número de clusters y es útil para descubrir estructuras subyacentes en los datos. [10]

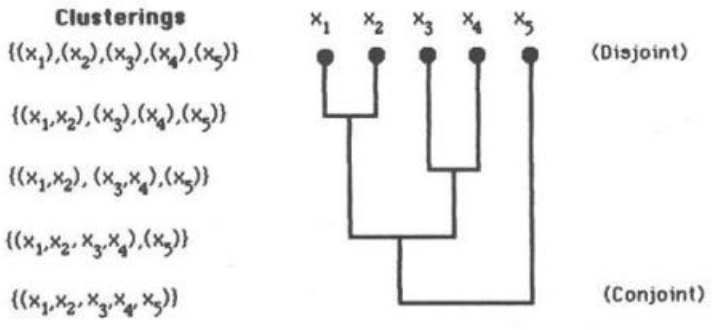

Figura 6. Ejemplo de dendograma

Para agrupar los datos, se necesita una medida de similitud o distancia entre los objetos. Algunas opciones comunes incluyen:

- Distancia Euclidiana:

$$d(A, B) = \sqrt{\sum(A_i - B_i)^2}$$

- Distancia de Manhattan:

$$d(A, B) = \sum|A_i - B_i|$$

- Coeficiente de correlación de Pearson (para datos normalizados).

El algoritmo sigue dos enfoques principales:

- **Métodos aglomerativos (Bottom-Up):** Comienza con cada observación en su propio cluster y va fusionando los más cercanos hasta formar un único cluster. [10]

  Proceso de cálculo en métodos aglomerativos:

  a) Cálculo de la matriz de distancias: Se determina la distancia entre cada par de observaciones utilizando una métrica adecuada, como la distancia euclidiana.
  b) Fusión de clusters: En cada iteración, se fusionan los dos clusters más cercanos según un criterio de enlace (linkage). Los criterios de enlace más comunes incluyen:

  - Enlace sencillo (single-linkage): Distancia mínima entre puntos de diferentes clusters.
  - Enlace completo (complete-linkage): Distancia máxima entre puntos de diferentes clusters.
  - Enlace promedio (average-linkage): Promedio de las distancias entre todos los pares de puntos de diferentes clusters.

c) Actualización de la matriz de distancias: Tras cada fusión, se actualiza la matriz de distancias para reflejar las distancias entre el nuevo cluster formado y los existentes.
d) Repetición del proceso: Los pasos de fusión y actualización se repiten hasta que todos los datos estén en un único cluster o se alcance un criterio de parada definido.

El resultado de este proceso se representa mediante un dendrograma, que ilustra las fusiones realizadas y permite visualizar la estructura jerárquica de los datos.

- **Métodos divisivos (Top-Down):** Comienza con un solo cluster y divide los datos en grupos cada vez más pequeños. [10]

5.7 MATRIZ DE CONFUSIÓN

La matriz de confusión es una herramienta fundamental para evaluar el desempeño de un modelo de clasificación. Se trata de una tabla que compara las predicciones del modelo con los valores reales del conjunto de prueba, permitiendo identificar los aciertos y errores cometidos. Para problemas binarios, esta matriz se organiza como una tabla de 2×2 que incluye: verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). A partir de ella se derivan métricas clave como la precisión, el recall, la exactitud y el F1-score.

En la figura 7 se muestra la composición de una matriz de confusión y como interpretarla.

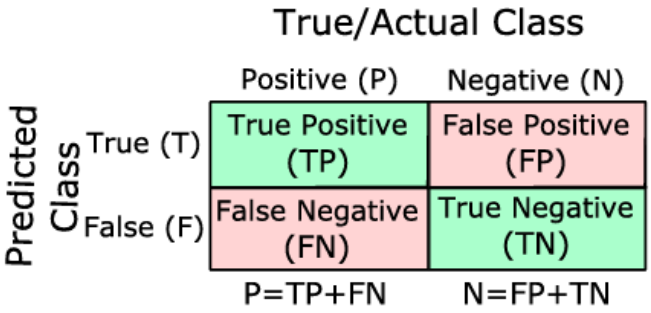

Figura 7. Matriz de confusión

Cada predicción realizada por un modelo de clasificación puede clasificarse en una de las siguientes cuatro categorías, dependiendo de su coincidencia con la etiqueta real:

- **Verdadero Positivo (TP):** La instancia fue clasificada como positiva y efectivamente lo era.
- **Verdadero Negativo (TN):** El modelo predijo una clase negativa y el valor real también era negativo.

- **Falso Positivo (FP):** El modelo predijo una clase positiva cuando en realidad era negativa; también se conoce como "falsa alarma".
- **Falso Negativo (FN):** El modelo predijo una clase negativa cuando en realidad era positiva; se considera un error por omisión.

A continuación, se presentan y describen las principales métricas de evaluación, explicando cómo se calculan a partir de los valores obtenidos en la matriz de confusión.

- **Exactitud (Accuracy):** Es la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de muestras evaluadas. Aunque es ampliamente utilizada, puede resultar engañosa en conjuntos de datos desbalanceados. Se calcula como:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Sensibilidad (Recall o Tasa de Verdaderos Positivos - TPR):** Mide la proporción de casos positivos que fueron correctamente identificados por el modelo.

$$Recall = \frac{TP}{TP + FN}$$

- **Especificidad (Tasa de Verdaderos Negativos - TNR):** Indica la proporción de casos negativos correctamente clasificados.

$$Specificity = \frac{TN}{TN + FP}$$

- **F1-Score:** Es una métrica que combina la precisión y el recall en un único valor mediante su media armónica. Es especialmente útil en contextos con clases desbalanceadas, ya que penaliza fuertemente los extremos (alta precisión y bajo recall, o viceversa). Se calcula de la siguiente manera:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

Un valor alto de F1-score indica que el modelo mantiene un buen equilibrio entre la identificación correcta de los positivos y la reducción de falsos positivos.

- **Curva ROC (Receiver Operating Characteristic):** Es una representación gráfica que evalúa el rendimiento de un modelo binario al variar el umbral de clasificación. Muestra la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Una curva ROC más cercana al vértice superior izquierdo indica mejor desempeño. Es

especialmente útil para comparar clasificadores y analizar su capacidad discriminativa sin depender de un umbral específico.[13]

En la figura 8 se observa un ejemplo y composición técnica de la curva ROC.

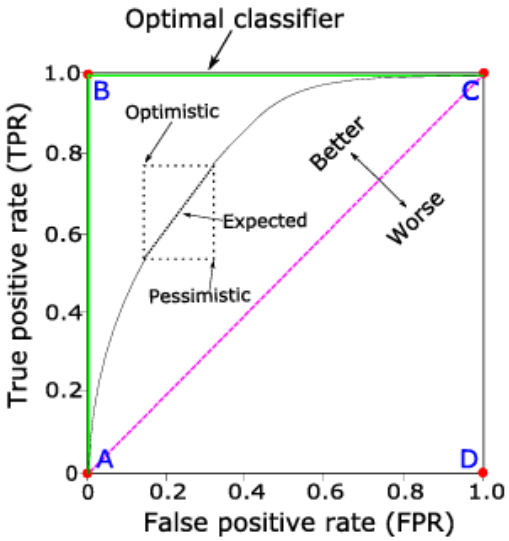

Figura 8. Curva ROC

6 ANTECEDENTES

6.1 PROYECTO DE REFERENCIA 1

Titulo: Modelo predictivo de churn-rate de clientes para el negocio de Telecomunicaciones
Autor: Andrés Felipe Echeverri Giraldo
Universidad: Universidad de Antioquia
Año: 2019

Resumen

El proyecto desarrollado por Andrés Felipe Echeverri Giraldo se centró en abordar el problema del churn (abandono) en el sector de telecomunicaciones. Dado que es más costoso atraer nuevos clientes que retener a los existentes, se elaboró un modelo predictivo utilizando técnicas de Machine Learning. El objetivo principal era predecir y prevenir el abandono de clientes, permitiendo de esta manera mejorar la fidelización. El modelo desarrollado, utilizando algoritmos como Árboles de Decisión, Random Forest,

XGBoost, logró predecir hasta un 66% del churn mensual, a través de una base de datos centralizada y la implementación de algoritmos supervisados desarrollados en Python.

Durante el desarrollo del proyecto de grado, se implementaron tres técnicas principales de Machine Learning para predecir el churn: Árboles de Decisión, Random Forest y XGBoost. Los Árboles de Decisión lograron un recall del 23%, precisión del 18% y un AUC de 0.61, mostrando un desempeño limitado. Random Forest mejoró la precisión a 54%, pero redujo el recall a 7%, con un AUC de 0.70, lo que no cumplió con los objetivos del proyecto. XGBoost, en cambio, demostró ser la técnica más efectiva con un AUC inicial de 0.96, mejorado a 0.95 tras ajustes, y un recall optimizado del 66%, detectando correctamente dos tercios de los casos de abandono.

El recall mide la proporción de verdaderos positivos correctamente identificados entre todos los casos positivos, crucial para capturar la mayor cantidad de abandonos posibles. La precisión indica la proporción de verdaderos positivos entre todos los casos positivos predichos, importante para evaluar la exactitud del modelo. El AUC (Área Bajo la Curva ROC) mide la capacidad del modelo para distinguir entre clases positivas y negativas, siendo un indicador global de rendimiento. Estos resultados destacan que XGBoost fue la técnica más adecuada, superando significativamente a las otras en términos de recall y AUC, haciendo más efectivo el modelo predictivo de churn.

Este proyecto aporta una base sólida en la implementación de modelos predictivos de churn-rate mediante técnicas de Machine Learning, lo cual es relevante para los objetivos del proyecto actual en COLOMBIA INTERNET ISP. La metodología utilizada, incluyendo la centralización de datos, la selección de algoritmos y la evaluación de modelos, proporciona un marco de referencia útil para desarrollar una estrategia similar.

No obstante, mientras que este proyecto se centra en el sector de telecomunicaciones en general, el proyecto actual está específicamente orientado hacia COLOMBIA INTERNET ISP. Además, el enfoque del proyecto actual no solo incluye la predicción del churn-rate, sino también la segmentación de clientes según su riesgo de abandono y la generación de visualizaciones de BI para identificar tendencias y patrones. [11]

6.2 PROYECTO DE REFERENCIA 2

Título: Modelo Análisis Predictivo para el Cálculo de Tasa de Deserción en una Empresa Aseguradora

Autor: Sergio Felipe Sierra Morales
Universidad: Fundación Universitaria Konrad Lorenz
Año: 2022

Resumen

El proyecto desarrollado por Sergio Felipe Sierra Morales se enfoca en diseñar un modelo de probabilidad para calcular la tasa de deserción en una empresa aseguradora, priorizando la reducción de falsos negativos (error tipo II) para minimizar la fuga de recursos. La empresa en cuestión opera en el sector asegurador colombiano, vendiendo pólizas y administrando riesgos laborales. El modelo se basa en algoritmos de análisis predictivo y utiliza la metodología CRISP-DM para abordar el problema. Durante el proyecto, se realizaron múltiples fases que incluyen la comprensión del negocio, entendimiento de los datos, preparación de los datos, modelado y evaluación. Se implementaron algoritmos como XGBoost, Naive Bayes y Redes Neuronales, comparando sus métricas y seleccionando el modelo más adecuado para predecir la deserción de clientes.

Este proyecto proporciona una metodología detallada para la implementación de modelos predictivos de churn-rate utilizando técnicas de Machine Learning, lo cual es muy relevante para los objetivos del proyecto actual en COLOMBIA INTERNET ISP. La aplicación de la metodología CRISP-DM y el uso de diversos algoritmos, así como la evaluación de sus métricas, ofrece un marco de referencia sólido para desarrollar un modelo predictivo similar.

Se enfocó en mejorar el Recall – TNR. Entre los algoritmos empleados, se destaca XGBoost, que sin balancear las clases, logró un 95% de acierto para clientes desertores y un AUC de 0.72. Otro modelo con XGBoost alcanzó una precisión del 81%, AUC de 0.70 y TNR de 0.37. Aunque es efectivo en identificar clientes desertores, también genera muchas alertas falsas, incrementando costos y afectando la operatividad de la aseguradora.

El algoritmo Naive Bayes mostró un Recall alto (0.99) pero mala identificación de la clase positiva y un AUC de 0.70, por lo que se descartó. Las redes neuronales, a pesar de un buen Recall (0.80), no alcanzaron un AUC satisfactorio (<70%) y no mostraron el equilibrio deseado en las métricas comparadas con XGBoost.

El proyecto concluye que la variable "nivel de riesgo" está directamente relacionada con la deserción y su aumento incrementa la activación de pólizas, generando sentimientos negativos que motivan a buscar otro proveedor. La selección de algoritmos se basa en

métricas de calidad y deben acompañarse de múltiples experimentos para acercarse a la solución propuesta. XGBoost se destaca por su flexibilidad y efectividad en la clasificación de clientes, considerando también factores económicos y operacionales.

Finalmente, se sugiere reestructurar la fuente de información para obtener resultados más precisos y mejorar las técnicas de retención, abordando problemas reales y cuellos de botella en la gestión del servicio al cliente.

Mientras que el proyecto de Sierra se centra en el sector asegurador, el proyecto actual está orientado hacia el sector de telecomunicaciones, específicamente para COLOMBIA INTERNET ISP. [12]

6.3 PROYECTO DE REFERENCIA 3

Título: Modelo de Abandono de Clientes en una Empresa de Créditos en Línea
Autores: David Fajardo, Andrés Motta
Universidad: Universidad de Los Andes
Año: 2019

Resumen

Este proyecto desarrollado por David Fajardo y Andrés Motta se centra en la creación de un modelo de predicción de abandono (churn-rate) para una empresa de créditos en línea, Zinobe. El objetivo principal es identificar a los clientes que están en riesgo de abandonar el servicio, utilizando datos de diferentes canales de servicio al cliente y técnicas de Machine Learning. Para ello, se recopilaron datos de correos electrónicos, llamadas, y otras interacciones con los clientes, y se implementaron algoritmos de procesamiento de lenguaje natural (NLP) para analizar el contenido de los correos electrónicos y desarrollar perfiles sentimentales de los clientes. Los modelos utilizados incluyen Support Vector Machine, Regresión Logística, Random Forest, Bagging, Gradient Boosting y Redes Neuronales, evaluando su desempeño y seleccionando el más adecuado basado en métricas como ROC, F1-Score, Precision y Recall.

En el proyecto de modelo predictivo de abandono de clientes (churn-rate) y segmentación, se implementaron algoritmos de machine learning como Regresión Logística, Random Forest, Gradient Boosting y Redes Neuronales, obteniendo mejores resultados con Random Forest, que mostró un ROC (AUC) de 0.786, F1-Score de 0.684, Accuracy de 0.762, Precision de 0.574 y Recall de 0.847, destacándose por identificar correctamente la mayoría de los casos de churn-rate. En términos económicos, Random Forest demostró ser el modelo más

eficiente bajo diferentes escenarios de costos de mantenimiento, logrando beneficios adicionales mensuales significativos, especialmente con un costo de mantenimiento de $1,000 por cliente, donde se alcanzaron beneficios entre $192,160,160 y $1,281,067,736. Se concluyó que las campañas de mantenimiento basadas en análisis de datos de servicio al cliente son más efectivas y menos costosas que las tradicionales de incentivos y promociones.

Este proyecto proporciona una metodología exhaustiva para la implementación de modelos predictivos de churn-rate utilizando diversas técnicas de Machine Learning, lo cual es directamente aplicable al proyecto actual en COLOMBIA INTERNET ISP. La inclusión de perfiles sentimentales basados en análisis de texto y la combinación de estos con perfiles demográficos y crediticios ofrece un enfoque robusto que puede mejorar significativamente la precisión del modelo predictivo de churn-rate. Además, la evaluación económica del impacto de estos modelos en términos de ingresos adicionales y reducción de costos proporciona una perspectiva valiosa sobre la implementación práctica y el retorno de inversión de las estrategias de retención de clientes.

Sin embargo, este proyecto se centra en una empresa de créditos en línea, mientras que el proyecto actual está orientado hacia el sector de telecomunicaciones, específicamente para COLOMBIA INTERNET ISP. [13]

7 METODOLOGÍA

El presente proyecto es de tipo aplicado, combinando análisis teóricos con la implementación práctica de técnicas de ciencia de datos y modelos predictivos de machine learning. Para esto, se aplicó la metodología CRISP-DM, un modelo de procesos estándar para la minería de datos que proporciona una estructura común para planificar y ejecutar el proyecto.

La metodología CRISP-DM se desglosa en seis fases principales (*Figura 9*), que se alinean con los objetivos específicos de la investigación:

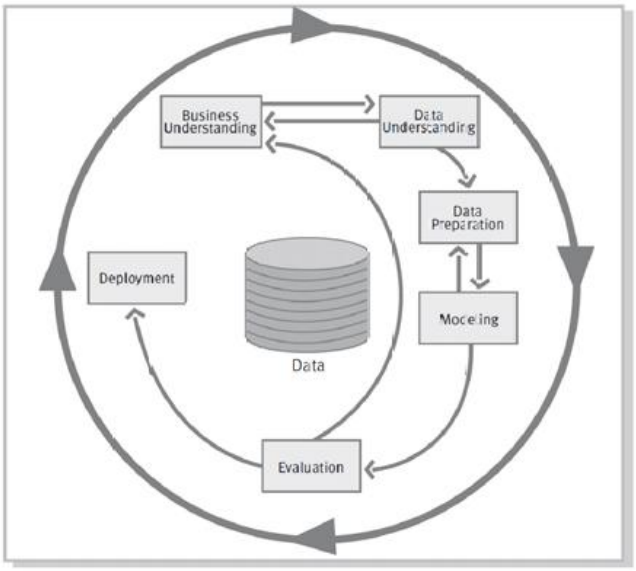

Figura 9. Fases del proceso de la metodología CRISP-DM

A. Comprensión del Negocio

El primer paso del proceso es entender el contexto del negocio y sus requerimientos, en este caso, para abordar la problemática del abandono de clientes de COLOMBIA INTERNET ISP. Se enfocó en la importancia de reducir la tasa de churn-rate para mejorar la retención de clientes.

- **Objetivo:**

- Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.

- **Actividades:**

- Definir los objetivos clave del negocio, como reducir la tasa de churn-rate y mejorar la retención.
- Recolectar los datos históricos de cancelaciones, pagos y solicitudes de soporte.

- Identificar indicadores clave de desempeño (KPIs) relacionados con el churn-rate.

B. Obtención y recolección de los Datos

En esta fase, se recolectan, analizan y entienden los datos disponibles. Esto implica el análisis exploratorio para detectar patrones o tendencias iniciales.

- **Objetivo:**

- Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.

- **Actividades:**

- Revisar la disponibilidad y estructura de los datos.
- Recopilar los datos y analizar el estado de las variables.
- Detectar problemas de calidad de los datos, como valores faltantes o inconsistencias.
- Realizar análisis descriptivos para identificar tendencias y anomalías relevantes.

C. Preparación de los Datos

La fase de preparación de los datos incluye la limpieza, transformación y selección de las variables clave que se utilizarán en el modelado. Este paso es crucial para asegurar que los modelos predictivos funcionen correctamente.

- **Objetivo:**

- Analizar el historial de cancelaciones mensuales, comportamientos de pago y solicitudes de soporte para el entendimiento, limpieza, transformación y posterior uso de los datos.

- **Actividades:**

- Limpieza de datos para corregir errores y tratar valores atípicos.
- Transformación de variables relevantes, como la creación de nuevas características basadas en comportamientos de los clientes.
- Normalización y preparación de datos para su uso en algoritmos de machine learning.

D. Modelado

En esta fase, se exploran y desarrollan diferentes algoritmos de machine learning para para predecir el riesgo de abandono de clientes. Se aplicarán varias técnicas para determinar qué modelo es el más adecuado.

- **Objetivos:**

- Explorar y desarrollar diferentes modelos de predicción de churn-rate.
- Evaluar los modelos de aprendizaje desarrollados para definir el mejor modelo de acuerdo con las preguntas de la investigación, considerando la optimización de sus hiperparámetros.

- **Actividades:**

- Selección de modelos como regresión logística, árboles de decisión, random forest, Support vector machine y Extreme Gradient Boosting.
- Entrenamiento de los modelos utilizando técnicas de validación cruzada.
- Comparación de los resultados obtenidos a través de métricas como precisión, recall, F1 score, AUC-ROC, entre otras.

E. Evaluación

Se evalúan los modelos desarrollados en función de las métricas establecidas, determinando cuál es el más adecuado para la predicción del churn-rate en función de su desempeño.

- **Objetivo:**

- Evaluar los modelos de aprendizaje desarrollados para definir el mejor modelo de acuerdo con las preguntas de la investigación, considerando la optimización de sus hiperparámetros.

- **Actividades:**

- Aplicación de pruebas con datos de validación y testeo de modelos.
- Comparación de modelos según métricas de desempeño.
- Ajuste de hiperparámetros para mejorar la precisión del modelo seleccionado.

F. Implementación

Esta fase final consiste en desplegar el modelo seleccionado en un entorno donde pueda ser utilizado de manera práctica para generar alertas tempranas de abandono de clientes.

- **Objetivo:**

- Tipificar a los clientes de acuerdo con su nivel de riesgo de fuga, empleando las variables características que los describen.

- **Actividades:**

- Desplegar el modelo en un entorno funcional para realizar predicciones de manera continua.
- Establecer un flujo que procese nuevas observaciones y genere predicciones.
- Evaluar el desempeño del modelo en el entorno de implementación y realizar ajustes si es necesario.

Ventajas de CRISP-DM

- Es **flexible y adaptable** a distintos tipos de proyectos de minería de datos.
- Proporciona una **estructura clara** y ordenada en seis fases.

- Es **independiente de herramientas** o lenguajes específicos.
- Facilita la **documentación y reutilización** del conocimiento del proyecto.
- Tiene amplio **uso en la industria**, lo que mejora la colaboración entre equipos.

Desventajas de CRISP-DM

- Puede ser demasiado general y requerir adaptación para proyectos específicos.
- No incluye metodologías ágiles o de desarrollo iterativo continuo por defecto.
- No contempla actualizaciones del modelo en producción (MLOps).
- Algunas fases, como la comprensión del negocio, pueden volverse subjetivas.
- Puede resultar lento en entornos dinámicos o con necesidades de entrega rápida.

7.1 COMPRENSIÓN DEL NEGOCIO

Colombia Internet es un proveedor de servicios de Internet con inicio de operaciones en 2019 y presencia en los departamentos de Valle del Cauca y Tolima, al inicio de este proyecto (octubre de 2024) la compañía cuenta con una base de clientes activos es de 104.551 clientes, sin embargo, también se tiene un registro de 32.810 clientes que han abandonado el servicio, lo cual representa el 31% del total de clientes actuales.

El churn-rate en la industria de las telecomunicaciones es uno de los más altos en comparación con otros sectores económicos, ubicándose entre 1.5 y 2% mensual, actualmente el churn-rate de Colombia Internet se ubica en 2.1% mensual para lo corrido entre octubre de 2023 a octubre de 2024, aunque el valor se encuentra apenas por encima del promedio, es necesario entender este comportamiento, el impacto generado en la experiencia de los clientes y en las finanzas de la empresa a mediano plazo.

En la figura 10 se observa la relación entre el comportamiento del ingreso de nuevos clientes y las cancelaciones de servicio, en el período comprendido entre octubre de 2023 y octubre de 2024. Además, se incluye una línea de tendencia que muestra el promedio mensual de la tasa de cancelación (churn rate).
En la figura 10 se observa la relación entre el comportamiento del ingreso de nuevos clientes y las cancelaciones de servicio, en el período comprendido entre octubre de 2023 y octubre de 2024. Además, se incluye una línea de tendencia que muestra el promedio mensual de la tasa de cancelación (churn rate).

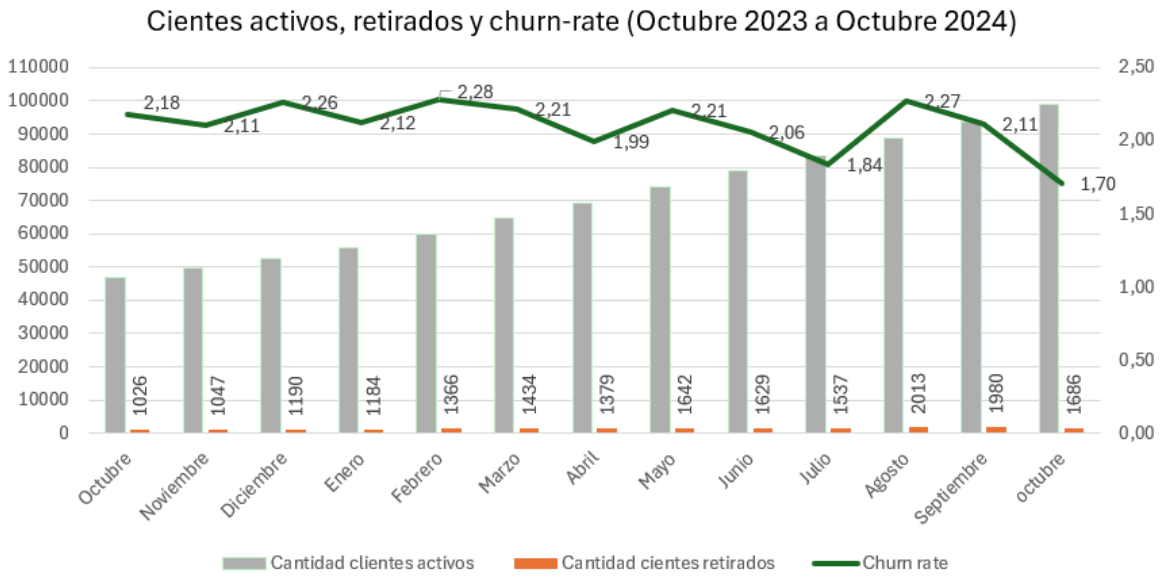

Figura 10.Comportamiento churn-rate mensual 2024 – Colombia Internet

Aunque la tendencia ha mostrado una disminución a lo largo de los meses, en términos de cantidades, el promedio de abandonos mensuales representa 1470 clientes. El valor promedio del ticket mensual se encuentra en $66.521, con lo anterior podemos concluir que se está dejando de percibir ingresos directos de alrededor de $97.787.870 COP mensuales.

Por lo tanto y si tenemos en cuenta el de adquisición de clientes (CAC), que actualmente es de $1.200.000 COP, lo cual incluye costos de marketing, ventas y gastos operativos de instalación, un cliente debería mantener el servicio al menos 18 meses para que la empresa recupere este valor, sin embargo, y como podemos evidenciar en la siguiente gráfica, se realizó el análisis de la duración de los 32810 clientes retirados, evidenciando que esto no ocurre en la mayoría de los casos, concluyendo que solo en el 14.9% de los abandonos se logró recuperar este gasto.

En la figura 11 se muestra la duración de los clientes con el servicio, considerando un rango de hasta 6 meses. Se identifica que el 41 % de los retiros (13.441 casos) ocurrió dentro de los primeros 6 meses.

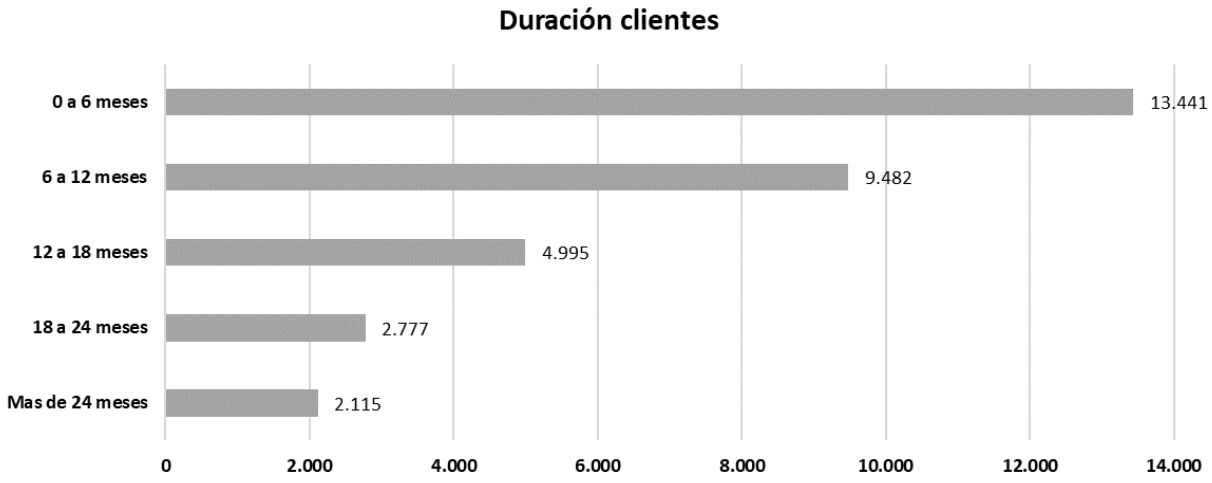

Figura 11. Duración clientes retirados

Con respecto a los motivos de terminación, en primer lugar, se justifica por un cambio de ciudad donde no se cuenta con cobertura para un traslado de servicio, en segundo lugar, se indican dificultades económicas para realizar los pagos y como tercer motivo se encuentran traslados de viviendas a sitios en la misma ciudad, pero donde no hay cobertura de redes.

Podemos observar que se presentan 2127 valores nulos que deberán ser imputados con la técnica apropiada.

Esta variable puede ayudar a comprender la dinámica de los abandonos de servicio, aunque es una variable que no se puede usar en el modelo, ya que esta información se obtiene una vez el cliente se ha retirado.

La figura 12 es una tabla que muestra los motivos por los cuales los clientes han terminado su servicio, desglosados en los dos departamentos: Valle del Cauca y Tolima. Se presenta el número total de casos por cada motivo en ambos departamentos, así como un total general.
La figura 12 es una tabla que muestra los motivos por los cuales los clientes han terminado su servicio, desglosados en los dos departamentos: Valle del Cauca y Tolima. Se presenta el número total de casos por cada motivo en ambos departamentos, así como un total general.

Motivos Terminaciónes

	Departamento / Record Count		
MOTIVO TERMINACIÓN	VALLE DEL ...	TOLIMA	Total
CAMBIO DE CIUDAD	6.212	3.242	9.454
MOTIVOS ECONOMICOS	3.610	1.224	4.834
TRASLADO/SIN COBERTURA	2.889	1.235	4.124
SOLICITUD DE PROPIETARIO	2.101	1.104	3.205
TRASLADO NUEVO PREDIO	1.890	821	2.711
null	2.027	100	2.127
MEJOR OFERTA TV Y OTROS SE...	1.495	616	2.111
CALIDAD DE SERVICIO	1.368	526	1.894
GESTIÓN COBRANZA	1.445	142	1.587
MEJOR OFERTA	558	119	677
COMPORTAMIENTO INDEBIDO P...	35	14	49
INCUMPLIMIENTO VISITA DE MA...	30	7	37
Total	23.660	9.150	32.810

Figura 12. Motivos terminaciones

7.2 OBTENCIÓN Y RECOLECCIÓN DE LOS DATOS

La información se obtuvo principalmente del sistema OSS(Sistema de Soporte Operativo), de allí se obtuvo la base de datos de clientes activos, mantenimientos ejecutados y terminaciones completadas, se encontraron en algunas variables valores nulos que deben pasar por el proceso de preparación de los datos, la información comprende datos desde octubre de 2021 a octubre de 2024, esta fue recibida en formato de Excel.

Un compromiso al momento de la recolección de datos fue la protección de la información personal de los clientes, ya que esta contiene datos como números de cedula, nombres, direcciones y teléfonos de contacto, información sensible y que de no ser manipulada con ética y responsabilidad, puede generar afectaciones a los clientes como reputacionales de la compañía, por lo anterior firmamos un acuerdo de confidencialidad entre las personas que participamos en el proyecto y la empresa.

7.3 PREPARACION DE LOS DATOS

En esta sección se describe inicialmente la estructura del conjunto de datos, detallando las variables disponibles, sus tipos y características generales. Posteriormente, se documentan las transformaciones aplicadas para limpiar, ajustar y enriquecer la base, con el fin de dejarla lista para las etapas de análisis y modelado.

7.3.1 DESCRIPCIÓN INICIAL DEL CONJUNTO DE DATOS

La base de datos utilizada para este trabajo está compuesta por 137361 registros y 15 variables, las cuales recogen información de clientes del servicio de internet entre septiembre de 2021 y octubre de 2024. Estas variables incluyen datos demográficos, comerciales, transaccionales y de comportamiento asociados al cliente.

A continuación, se presenta el diccionario de datos que describe el significado de cada variable y su tipo:

Variable	Tipo	Descripción
departamento	String	Departamento donde están ubicados los clientes.
municipio	String	Municipio donde están ubicados los clientes.
tipo_cliente	String	Tipo de venta asociada al cliente.
id_cuenta	String	Identificador de la cuenta del cliente en el sistema.
cedula	String	Identificador único del cliente.
estrato	String	Estrato socioeconómico del cliente.
tipo_plan	String	Tipo de plan adquirido por el cliente.
canal_venta	String	Canal de adquisición del cliente.
churn	Binario	Indica si el cliente se ha retirado del servicio o no.
fecha_instalacion	Date	Fecha de instalación del servicio de internet.
fecha_terminacion	Date	Fecha de cancelación o terminación del servicio de internet.
antiguedad	Numérico	Antigüedad del cliente en la empresa, medida en meses. Se obtiene entre la diferencia de la fecha de terminación y fecha de inicio.
orden_post_venta	String	Tipo de servicio postventa requerido por el cliente.
motivo_terminacion	String	Motivo por el cual el cliente canceló el servicio.
precio_plan	Numérico	Precio mensual según el tipo de plan contratado (100, 200, 400 y 600 Mbps) y estrato socioeconómico del cliente.
total_pagado	Numérico	Valor total pagado por los meses de permanencia. Se obtiene de la multiplicación entre el precio del plan y la antigüedad.

Tabla 1. Diccionario de datos y tipos de variables

En esta r599199evisión preliminar, se identificaron los siguientes tipos de datos en la base original:

- 10 variables categóricas (tipo cadena de texto o categoría), tales como: departamento, municipio, tipo_cliente, estrato, tipo_plan, canal_venta, orden_post_venta, motivo_terminacion.
- 2 variables tipo fecha: fecha_instalacion y fecha_terminacion.
- 3 variables numéricas continuas: antiguedad, precio_plan y total_pagado.
- 1 variable binaria: churn, que representa el abandono del cliente.

Se identificó además que 9 variables presentan valores nulos, lo cual será tenido en cuenta en el proceso de limpieza y preparación de los datos.

Variable	Tipo	Nulos	% Nulos
departamento	String	0	0.00
municipio	String	0	0.00
tipo_cliente	String	10738	7.81
id_cuenta	String	0	0.00
cedula	String	2169	1.57
estrato	String	7060	5.13
tipo_plan	String	0	0.00
canal_venta	String	26382	19.20
churn	Binary	0	0.00
fecha_instalacion	Date	152	0.11
fecha_terminacion	Date	0	0.00
antiguedad	Int	152	0.11
orden_post_venta	String	0	0.00
motivo_terminacion	String	2127	1.54
precio_plan	Int	7061	5.14
total_pagado	Int	7401	5.38

Tabla 2. Cantidad y proporción de valores nulos por variable

Se realiza la descripción de las variables numéricas presentes en la base de datos con el objetivo de identificar su comportamiento general, evaluar posibles sesgos, rangos de valores y presencia de valores atípicos. Esta revisión permite detectar distribuciones no normales y orientar decisiones posteriores en cuanto a transformaciones, escalamiento o imputaciones necesarias. A continuación, se presentan las principales estadísticas descriptivas para estas variables.

Variable	Media	Desviación	min	25%	50%	75%	max
antiguedad	13.9	11.4	0	5	10	20	61
precio_plan	66365.2	12602.7	55900	55900	67800	67800	141491
total_pagado	940410.7	828146.5	55900	335400	678000	1290912	7640514

Tabla 3. Análisis descriptivo variables numéricas

De igual forma, se realiza un análisis descriptivo de las variables categóricas mediante una tabla resumen que incluye el total de observaciones no nulas, el número de categorías únicas y la categoría dominante (moda) junto con su frecuencia. Este análisis permite identificar qué categorías dominan en cada variable, facilitando la detección de posibles desbalances o sesgos en la distribución. La información obtenida sirve como base para tomar decisiones sobre el tratamiento posterior de estas variables dentro del flujo de análisis predictivo.

Variable	Frecuencia	Categorías	Moda	Frecuencia moda
departamento	137361	2	VALLE DEL CAUCA	94278

municipio	137361	30	IBAGUÉ	30635
tipo_cliente	126623	2	INQUILINO	92595
estrato	130301	6	2	75999
tipo_plan	137361	4	200/200	69294
canal_venta	110979	12	PAP	53353
orden_post_venta	137361	3	no_encontrado	120864
motivo_terminacion	135234	12	SIN_RETIRO	101221

Tabla 4. Análisis descriptivo variables categóricas

Se analiza la distribución de la variable objetivo del modelo, que en este caso es churn, la cual representa si el cliente se ha retirado del servicio o no. Se identifica que el 76.1% de los clientes permanecen activos y el 23.9% ha desertado. Este desbalance sugiere la necesidad de aplicar técnicas de balanceo durante la etapa de modelado.

Churn	Total clientes
No	104551
Si	32810

Tabla 5. Distribución variable objetivo

7.3.2 IMPUTACIÓN DE VALORES FALTANTES

Para el tratamiento de valores nulos, se aplicaron distintas estrategias según el tipo y el contexto de las variables:

- Variables categóricas

En las variables tipo_cliente (10738), motivo_terminacion (2127), estrato (7060) y canal_venta (26382), se identificaron valores faltantes. Dado que estas variables son categóricas, se optó por imputar los valores nulos utilizando la moda, es decir, el valor más frecuente dentro de cada una. Esta técnica es ampliamente utilizada en el tratamiento de variables categóricas cuando no se dispone de fuentes externas confiables para una imputación basada en reglas o modelos supervisados. Además, esta decisión se respalda en el hecho de que el porcentaje de datos nulos en estas variables es menor al 20% (8%, 2%, 5% y 19% respectivamente), umbral que suele considerarse aceptable en prácticas de análisis de datos, ya que no compromete significativamente la representatividad de la variable ni introduce sesgos graves en el modelo. Imputar con la moda en estos casos permite conservar la distribución original de la variable sin eliminar registros valiosos, manteniendo así el tamaño y la diversidad del conjunto de datos para las etapas posteriores del análisis.

- Variables numéricas

La variable antigüedad con 152 valores nulos, lo que representa apenas el 0.11% del total de registros. Aunque su distribución muestra una ligera asimetría positiva, con coeficiente de asimetría de 1, se optó por imputar los valores faltantes con la media. Esta decisión se justifica porque la proporción de datos nulos es mínima y la diferencia entre la media y la mediana (media = 13.9, mediana = 10) no es lo suficientemente significativa como para distorsionar la información. Esta decisión permite conservar la tendencia general de la variable y mantener la coherencia con otras transformaciones numéricas del conjunto de datos. Además, los modelos de predicción considerados en el análisis no se ven afectados por esta elección, en especial los basados en árboles, que son robustos ante sesgos y escalas.

Para completar los valores faltantes en la variable precio_plan, se utiliza una tabla con información externa que relaciona el estrato socioeconómico y el tipo de plan contratado con su respectivo precio. Se cruza esta información a la base de datos mediante una combinación basada en estas dos variables. Posteriormente, los valores faltantes en la variable precio_plan son reemplazados utilizando estos valores de referencia.

Finalmente, en la variable total_pagado, los valores nulos son calculados multiplicando el precio del plan por la antigüedad del cliente, bajo el supuesto de que el cliente ha pagado el mismo valor del plan durante todo el tiempo de antigüedad registrado.

7.3.3 VERIFICACIÓN Y DEPURACIÓN FINAL DEL CONJUNTO DE DATOS

Tras la imputación de valores faltantes, se lleva a cabo una fase de verificación y depuración del conjunto de datos, con el objetivo de dejarlo completamente listo para el análisis exploratorio y la modelación.

Para evitar sesgos y reducir la complejidad del modelo, se eliminan columnas que no aportan valor al análisis:

- Identificadores personales, como la cedula y id_cuenta, que no contienen información predictiva útil y pueden generar problemas de privacidad.
- Fechas de instalación y terminación, que ya fueron utilizadas para calcular otras variables relevantes como la antigüedad del cliente, y por tanto ya no se requieren directamente.
- **motivo_terminacion**, ya que solo aplica a clientes que desertaron, por lo que presenta una alta proporción de valores nulos en aquellos casos donde el cliente continúa activo, afectando su utilidad como predictor general.

Se valida nuevamente la cantidad de datos faltantes para asegurar que se haya corregido el problema de forma efectiva y el tipo de cada variable, asegurándose de que todas estén correctamente clasificadas según su naturaleza:

- Las variables numéricas clave como antigüedad, precio_plan y total_pagado, son transformadas al tipo “entero”.
- Las variables categóricas como estrato, tipo_cliente y canal_venta son convertidas al tipo “categoría”, optimizando su interpretación por parte del modelo.

Variable	Tipo	%Nulos
departamento	category	0.00
municipio	category	0.00
tipo_cliente	category	0.00
estrato	category	0.00
tipo_plan	category	0.00
canal_venta	category	0.00
churn	int64	0.00
antiguedad	int64	0.00
orden_post_venta	category	0.00
precio_plan	int64	0.00
total_pagado	int64	0.00

Tabla 6. Cantidad de valores nulos y tipo por cada variable

Como resultado de este proceso de depuración y limpieza, la base final quedó compuesta por 11 variables, todas seleccionadas por su relevancia potencial en la predicción de la deserción de clientes (churn). Con esta validación se da por finalizada la etapa de preparación de los datos, lo que permite continuar con el análisis exploratorio y el desarrollo de los modelos predictivos de forma adecuada.

7.3.4 ANÁLISIS UNIVARIADO DE VARIABLES NUMÉRICAS

Se realizó un análisis exploratorio de las variables numéricas con el fin de conocer su comportamiento, tendencia central y dispersión. Las estadísticas descriptivas revelaron que variables como antiguedad, precio_plan y total_pagado presentan distribuciones asimétricas hacia la derecha, con presencia de valores extremos. Estos patrones fueron visualizados mediante histogramas y diagramas de caja, lo que permitió identificar la necesidad de posibles transformaciones futuras. Además, se confirmó que estas variables no siguen una distribución normal, de acuerdo con los resultados de la prueba de Kolmogorov-Smirnov.

En esta fase del análisis se examinan individualmente las variables numéricas del conjunto de datos para entender mejor su comportamiento y distribución.

Se aplica la prueba estadística Kolmogorov-Smirnov, cuyo objetivo es determinar si una variable numérica sigue una distribución normal. Los resultados fueron los siguientes:

Variable	Estadístico	Valor p
antiguedad	0.135	0.00

precio_plan	0.342	0.00
total_pagado	0.141	0.00

Tabla 7. Prueba Kolmogorov-Smirnov

- En todos los casos, los valores p fueron inferiores a 0.05, lo que indica que ninguna de las variables numéricas analizadas tiene distribución normal.
- Esto sugiere que, para ciertas técnicas estadísticas o de modelado que asumen normalidad, podría ser necesario aplicar transformaciones, o preferir modelos que no dependan de esta suposición.

Se utilizó un nivel de significancia de 0.05 en la prueba de Kolmogorov-Smirnov debido a que es un umbral convencionalmente aceptado en estadística para tomar decisiones sobre la evidencia empírica [14]. La prueba de Kolmogorov-Smirnov evalúa la hipótesis nula de que la variable sigue una distribución normal. Al establecer un valor p de 0.05, se adopta un equilibrio entre el riesgo de cometer un error tipo I (rechazar la normalidad cuando es cierta) y la sensibilidad para detectar desviaciones importantes respecto a la normalidad [15]. Si el valor p resultante es menor a 0.05, se considera que la evidencia empírica es estadísticamente suficiente para rechazar la hipótesis de normalidad, conforme a los criterios comunes en pruebas de bondad de ajuste [16].

Variable Antigüedad: La mayoría de los usuarios tienen una antigüedad baja, especialmente concentrada entre 0 y 10 meses. A medida que aumenta la antigüedad, la cantidad de usuarios disminuye de forma progresiva, lo cual sugiere una distribución asimétrica a la derecha (sesgo positivo). Se observan valores atípicos hacia la derecha del boxplot (mayores de 40 meses), aunque en menor proporción.

La mayoría de los clientes son relativamente nuevos, con una menor proporción de usuarios de larga permanencia.

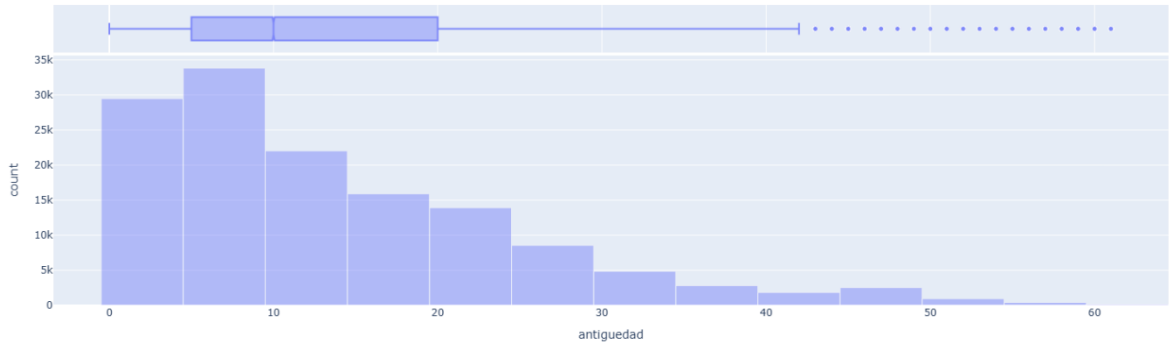

Figura 13. Distribución variable antigüedad

Variable Precio del plan: Hay una fuerte concentración de usuarios con planes entre $60.000 y $70.000. Posteriormente, se observa una caída en la frecuencia para valores superiores. El boxplot muestra varios valores extremos por encima de los $90.000, lo que indica que aunque son pocos, existen clientes con planes significativamente más costosos (hasta cerca de $140.000).

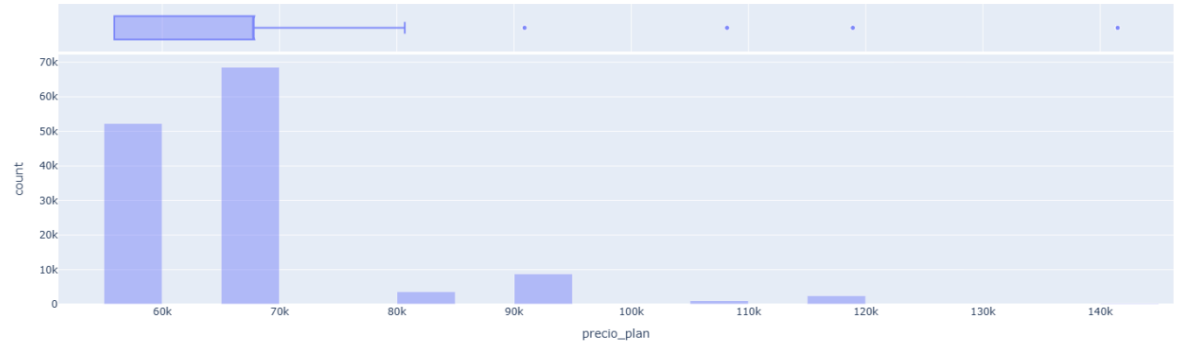

Figura 14. Distribución variable precio del plan

Variable Total Pagado: La distribución es notablemente asimétrica hacia la derecha. La mayoría de los usuarios han pagado montos totales bajos (entre $0 y $1.000.000), y la frecuencia disminuye conforme aumentan los montos. Hay una gran cantidad de valores atípicos hacia la derecha del boxplot, con pagos que superan incluso los $7.000.000, lo que evidencia alta dispersión en esta variable.

El total pagado depende de factores como la antigüedad y el tipo de plan. La alta variabilidad sugiere que esta variable podría tener un impacto relevante en modelos predictivos, pero también que podría requerir transformaciones para estabilizar su comportamiento.

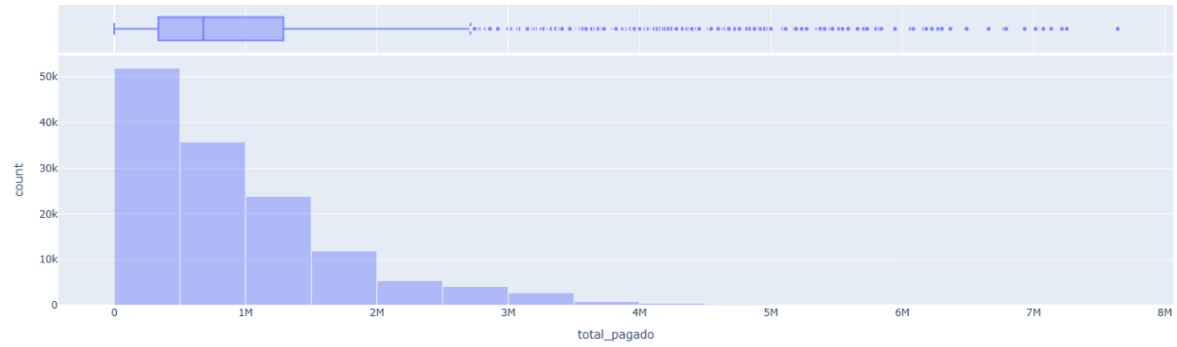

Figura 15. Distribución variable total pagado

7.3.5 ANÁLISIS UNIVARIADO DE VARIABLES CATEGÓRICAS

Para las variables categóricas se realizó un análisis exploratorio con el fin de conocer la distribución de las diferentes categorías presentes en el conjunto de datos. Esto permite identificar patrones de comportamiento, preferencias y concentraciones en determinadas categorías.

Variable Departamento: Valle del Cauca concentra cerca del 70% de los usuarios y Tolima agrupa el restante 30%.

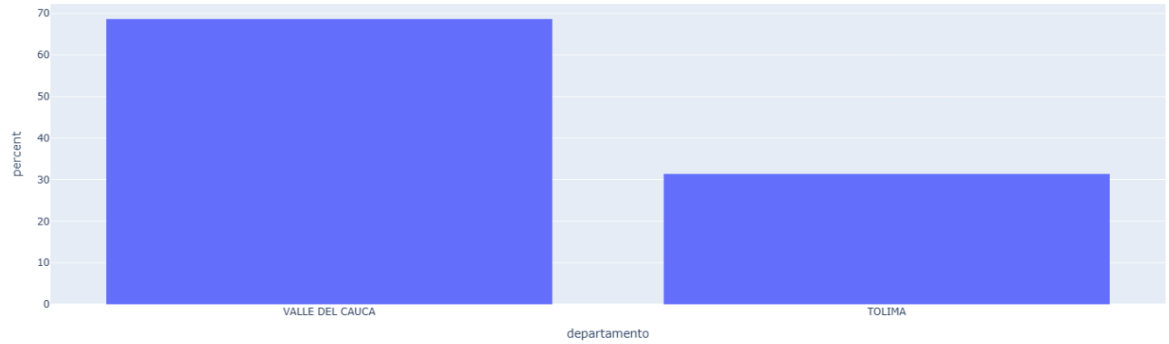

Figura 16. Distribución variable departamento

Variable Tipo de Cliente: La mayoría de los clientes son inquilinos (75%), lo que podría indicar una mayor rotación o diferente comportamiento de consumo frente a los propietarios (25%).

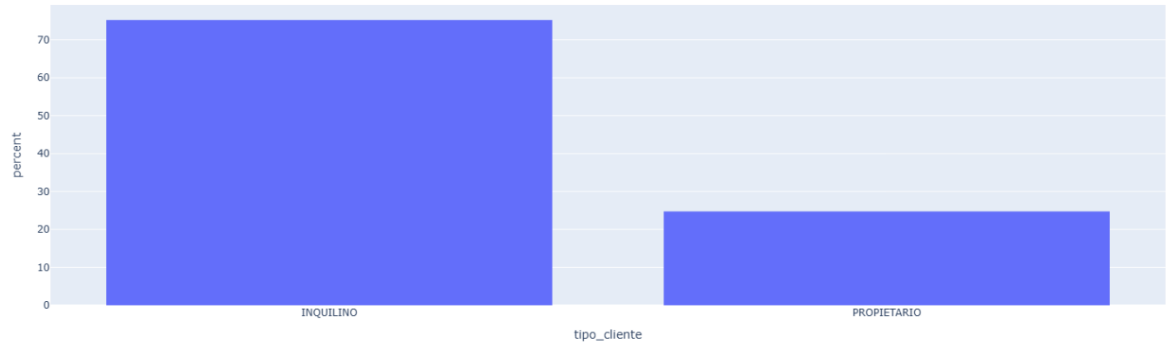

Figura 17. Distribución variable tipo de cliente

Variable Estrato: Los clientes son mayoritariamente de estratos bajos (60% estrato 2). El estrato 3 tiene aproximadamente el 25%, y el resto de los estratos están muy por debajo. Estratos 5 y 6 son casi inexistentes.

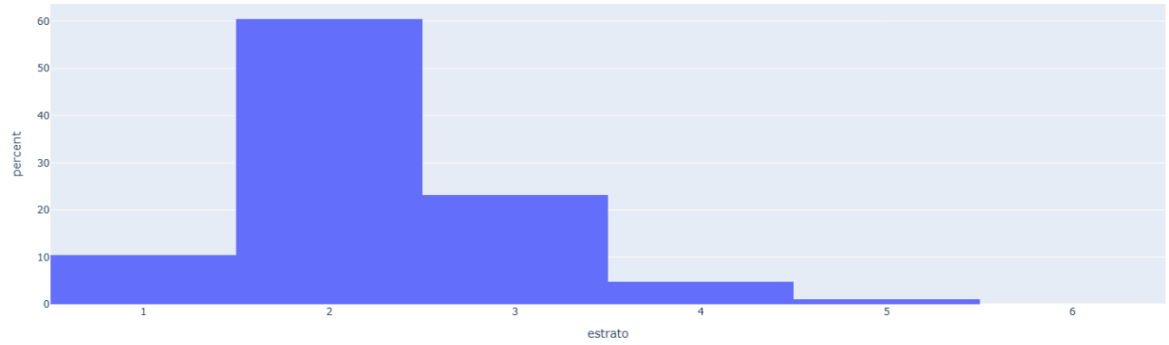

Figura 18. Distribución variable estrato

Variable Tipo de Plan: Hay una fuerte concentración en los planes de menor capacidad, lo que puede estar asociado a perfil económico. Los planes 100/100 y 200/200 son los más populares, representando aproximadamente el 40% y 50% (respectivamente) de los clientes. Los planes 400/400 y 600/600 tienen baja participación, lo que indica que los usuarios prefieren planes más económicos o básicos.

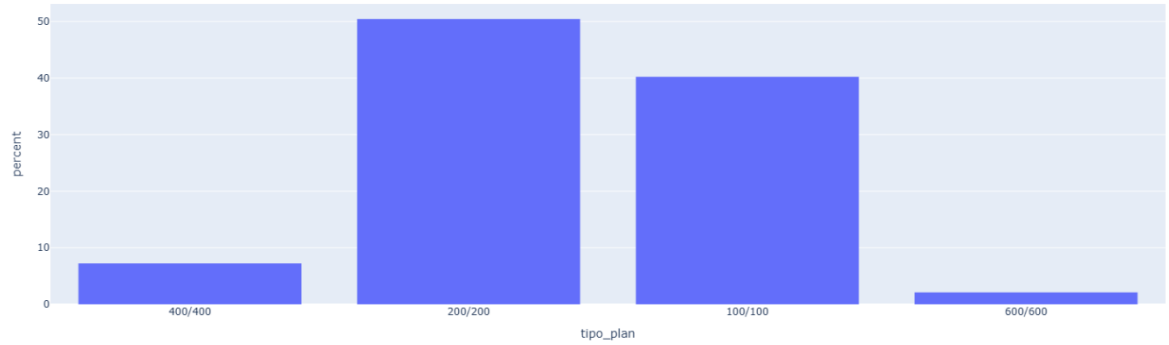

Figura 19. Distribución variable tipo de plan

Variable Canal de Venta: El canal PAP (Puerta a Puerta) es el más utilizado, con un casi 60% de participación. Los canales como Punto de Venta, Telefónico-Otro y WhatsApp Business tienen participaciones menores (5-15%). Otros canales como Tienda física, E-commerce y Proyectos de Conexión no muestran actividad significativa.

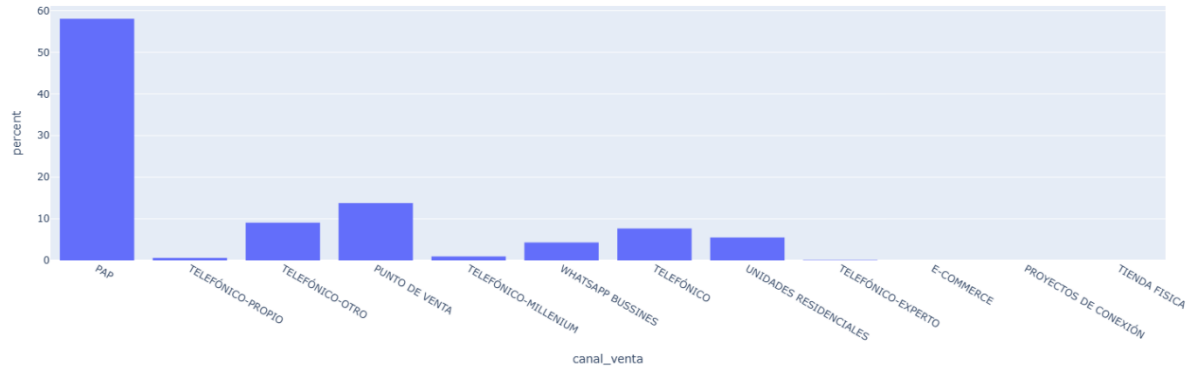

Figura 20. Distribución variable canal de venta

Variable Municipio: Los municipios con más clientes son Ibagué (22%) y Palmira (20%). Juntos, representan casi la mitad de los clientes. Hay una gran cantidad de municipios con porcentajes bajos (< 5%), incluso muchos por debajo del 1%.

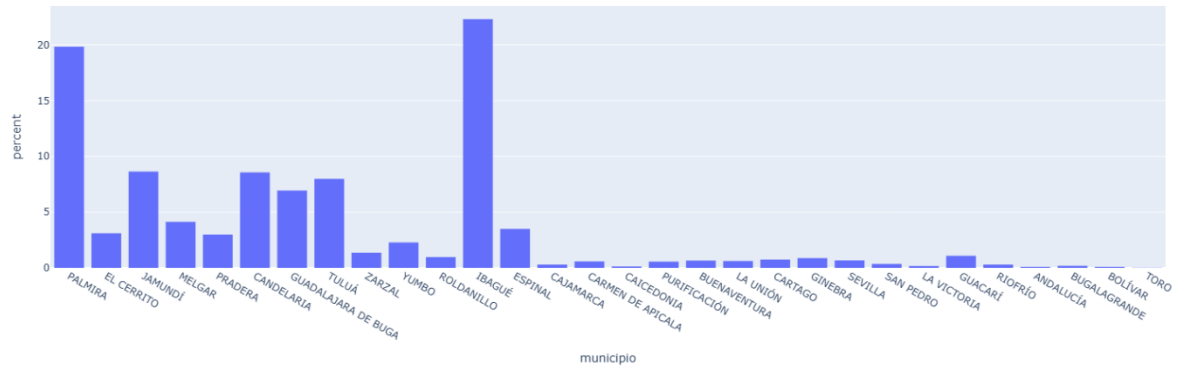

Figura 21. Distribución variable municipio

7.3.6 ANÁLISIS DE CORRELACIÓN

En la figura 22 encontramos la matriz para las variables numéricas, usa la correlación de Pearson para mostrar relaciones:

- churn (abandono del cliente): Correlación negativa con antigüedad (-0.16) y también con total_pagado (-0.15) y precio_plan (-0.017), aunque muy baja.
- antigüedad y total_pagado: correlación alta de 0.98, lo cual es esperable (cuanto más tiempo lleva el cliente, más ha pagado).
- precio_plan y total_pagado: correlación moderada (0.23), tiene sentido ya que planes más caros implican mayor gasto.

La única relación fuerte es entre antigüedad y total pagado. Las correlaciones con churn son bajas, pero podrían ser significativas en un modelo predictivo.

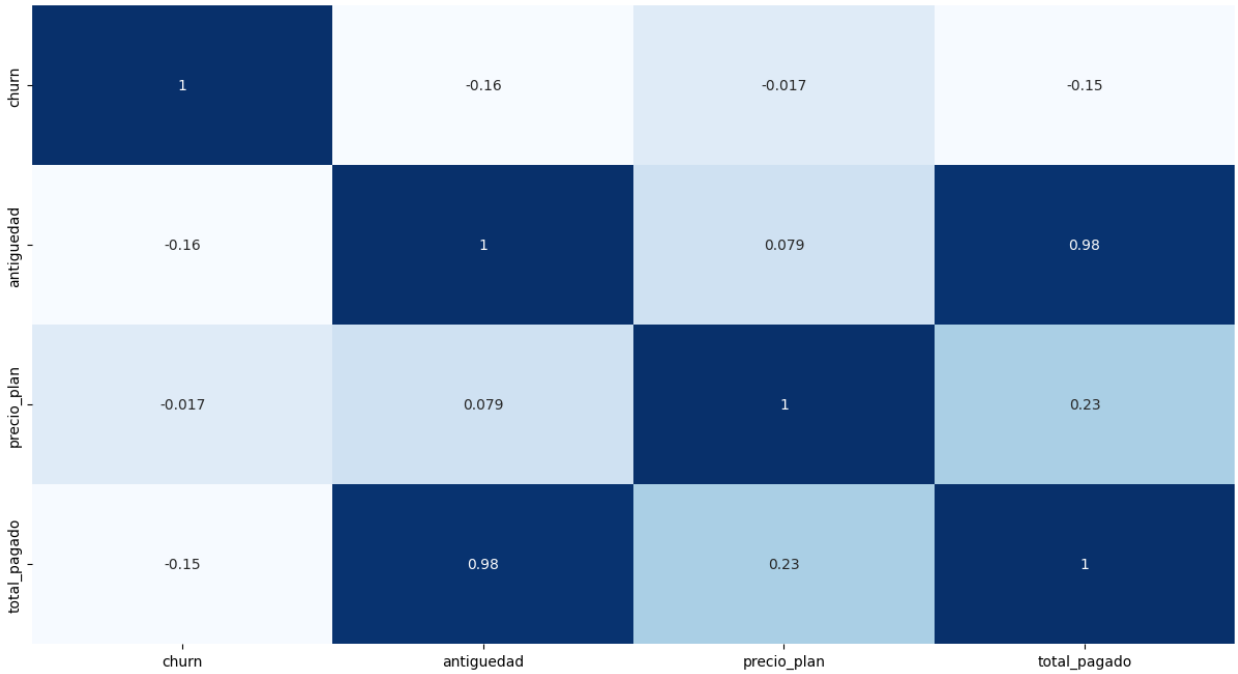

Figura 22. Matriz de correlación variables numéricas

En la figura 23 encontramos la matriz para las variables categóricas, esta matriz representa la correlación utilizando la métrica Cramér's V. Los valores van de 0 (sin relación) a 1 (relación perfecta).

- Departamento y municipio: Tienen una correlación de 1.0, lo cual es lógico ya que un municipio pertenece a un único departamento.
- Las demás correlaciones son bastante bajas (< 0.3), lo que sugiere que las variables categóricas no están altamente relacionadas entre sí.
- No se observa multicolinealidad preocupante (salvo en departamento vs municipio, lo cual es esperable).

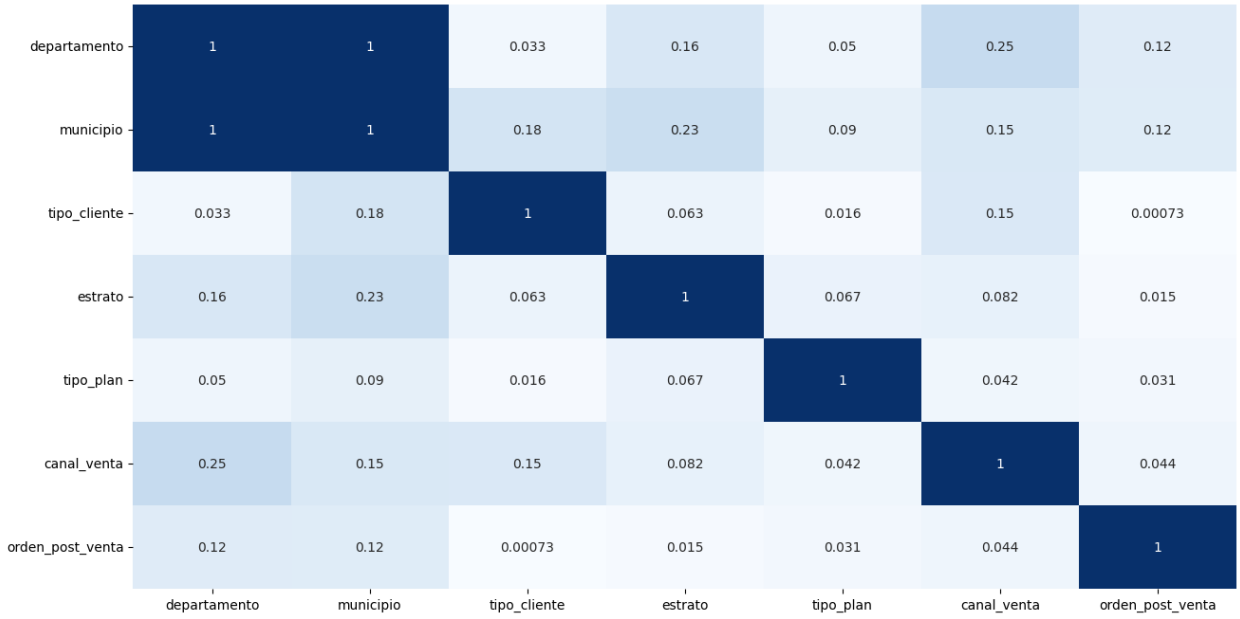

Figura 23. Matriz de correlación variables categóricas

En este análisis se evaluaron las relaciones entre variables categóricas y variables numéricas. Para ello, se utilizó la prueba de Kruskal-Wallis, una prueba estadística no paramétrica adecuada cuando se comparan distribuciones de una variable numérica entre varios grupos categóricos y no se puede asumir normalidad en los datos [17].

Los resultados muestran (Tabla 8), que la mayoría de las combinaciones evaluadas presentan diferencias significativas (valor p < 0.05), lo que indica que las variables categóricas analizadas influyen en las variables numéricas [17]. Algunas relaciones destacadas por su alto nivel de significancia y relevancia para el negocio incluyen:

Variable Categórica	Variable Numérica	Estadístico	Valor p
tipo_plan	precio_plan	132544.70	0.00
canal_venta	antiguedad	17193.45	0.00
canal_venta	total_pagado	16007.08	0.00
municipio	antiguedad	10077.84	0.00
municipio	total_pagado	9698.22	0.00
tipo_plan	total_pagado	7501.48	0.00
estrato	precio_plan	7076.03	0.00

Tabla 8. Prueba Kruskal-Wallis

Se adopta un valor p de 0.05 como umbral de significancia en la prueba de Kruskal-Wallis para establecer si existen diferencias estadísticamente significativas en la distribución de una variable numérica entre distintos grupos definidos por una variable categórica. Este nivel de significancia indica que existe un 5% de probabilidad de rechazar la hipótesis nula por error, es decir, concluir que hay diferencias entre los grupos cuando en realidad no las hay (error tipo I) [15]. El uso de este umbral facilita la identificación de patrones relevantes y consistentes desde el punto de vista

estadístico y de negocio, asegurando al mismo tiempo un control razonable sobre la probabilidad de obtener falsos positivos en las comparaciones múltiples [16] [17].

- Tipo de plan con precio del plan y total pagado: Existe una diferencia clara entre los tipos de planes en cuanto a su valor económico.
- Canal de venta con antigüedad y total pagado: El canal por el cual se adquiere un cliente impacta significativamente su antigüedad y cuánto paga.
- Municipio con antigüedad y total pagado: La ubicación geográfica también marca diferencias importantes en el comportamiento del cliente.
- Estrato con precio del plan y total pagado: Se observan diferencias socioeconómicas que influyen en la capacidad de pago y elección de plan.

7.4 MODELADO

Con base en el análisis exploratorio de los datos y los resultados del estudio de correlación, se optó por descartar ciertos atributos debido a su alta correlación con otras variables, con el fin de prevenir problemas de multicolinealidad. Entre ellos se encuentran:

- **Departamento:** Esta variable presenta una correlación perfecta (100%) con el campo Municipio, lo que indica que una es completamente redundante respecto a la otra.

- **Municipio:** Fue eliminado por redundancia frente a Departamento.

- **Antigüedad:** Tiene una correlación muy alta (98%) con Total pagado. Por su claridad interpretativa y relevancia, se conservó esta variable.

- **Total pagado:** Debido a su fuerte correlación con Antigüedad, se descartó para evitar duplicidad de información.

Por lo tanto, el dataset final con el que se desarrolló el modelo quedó conformado por las siguientes variables:

Variable	Tipo	%Nulos
departamento	category	0.00
tipo_cliente	category	0.00
estrato	category	0.00
tipo_plan	category	0.00
canal_venta	category	0.00
churn	int64	0.00
antiguedad	int64	0.00
orden_post_venta	category	0.00
precio_plan	int64	0.00

Tabla 9. Variables dataset final

Aunque la variable churn figura con tipo int64, en realidad representa una variable categórica binaria, ya que sus valores indican si un cliente se mantuvo (0) o canceló el servicio (1).

En este análisis, churn se define como la variable objetivo, pues corresponde al comportamiento que se desea predecir. Por otro lado, las demás variables seleccionadas en el apartado anterior constituyen las variables independientes, las cuales se emplean como predictores para estimar la probabilidad de que un cliente abandone el servicio.

7.4.1 MODELOS SELECCIONADOS PARA LA PREDICCIÓN DEL CHURN

En el caso de este proyecto, se optó por utilizar una selección de modelos de aprendizaje supervisado ampliamente reconocidos en estudios previos sobre predicción de abandono de clientes. Específicamente, se emplearon Regresión Logística, Árboles de Decisión, Random Forest y XGBoost, todos ellos con un sólido respaldo en la literatura por su efectividad en problemas de clasificación binaria como el churn.

Esta elección metodológica se basa en el análisis de antecedentes de investigaciones similares, donde estos algoritmos han demostrado un buen desempeño en términos de precisión, interpretabilidad y robustez frente a datos desbalanceados, comunes en este tipo de problemas. Por ejemplo, modelos como XGBoost y Random Forest han sido frecuentemente destacados por su capacidad para manejar relaciones no lineales y captar interacciones complejas entre variables, mientras que la Regresión Logística y los Árboles de Decisión ofrecen interpretaciones claras y resultados competitivos con menor complejidad computacional.

La combinación de estos modelos permite no solo comparar distintos enfoques de predicción, sino también identificar cuál se adapta mejor a las particularidades de los datos disponibles y a los objetivos del negocio, facilitando así la toma de decisiones estratégicas para reducir el abandono de clientes de manera proactiva y basada en resultados obtenidos mediante técnicas de análisis de datos.

7.4.2 MODELADO DE DATOS

Previo al entrenamiento de los modelos, se definieron las variables predictoras (independientes) y la variable objetivo (churn). Posteriormente, se realizó una partición del conjunto de datos, destinando el 80% para entrenamiento y el 20% restante para prueba. Esta división se llevó a cabo mediante un muestreo estratificado, garantizando que la proporción entre clases en la variable churn se mantuviera representativa en ambos subconjuntos.

a) **Conjunto de entrenamiento:** Contiene un total de 109,888 registros, correspondiente al 80% del dataset original. Este conjunto fue utilizado para entrenar los modelos predictivos. Al analizar la distribución de la variable objetivo churn, se identificó un fuerte desbalance: 104,551 registros pertenecen a la clase "0" (clientes que no abandonaron), frente a 32,810 de la clase "1" (clientes que sí abandonaron).

Para mitigar este desbalance en el conjunto de entrenamiento, se implementó la técnica de undersampling, mediante RandomUnderSampler con la estrategia 'majority'. Esta técnica reduce aleatoriamente los registros de la clase mayoritaria para igualarlos con los de la clase minoritaria. Su aplicación busca evitar el sesgo del modelo hacia la clase dominante, mejorando su capacidad para detectar clientes que podrían abandonar. El balance de clases resultante permite un entrenamiento más equitativo y mejora métricas relevantes como recall y F1-score.

b) **Conjunto de prueba:** Compuesto por 27,473 registros (el 20% restante del dataset), este conjunto fue reservado exclusivamente para validar el desempeño de los modelos sobre datos no vistos. La distribución original de la variable churn se conservó para simular un entorno de predicción realista.

Para evaluar el desempeño de los modelos aplicados, se consideraron diversas métricas que permiten medir la calidad de las predicciones. Las más relevantes, en orden de prioridad, son las siguientes:

- **Sensibilidad (Recall)**: mide la capacidad del modelo para identificar correctamente los casos positivos, es decir, los clientes que efectivamente abandonan el servicio.
- **Accuracy:** representa el porcentaje total de predicciones correctas, considerando tanto los casos positivos como negativos.
- **Especificidad**: refleja la proporción de verdaderos negativos correctamente identificados, es decir, clientes que no abandonan el servicio.
- **Curva ROC y AUC (Área bajo la curva)**: utilizadas como métricas globales de rendimiento del modelo, ya que permiten visualizar el equilibrio entre la tasa de verdaderos positivos y la tasa de falsos positivos.

7.5 EVALUACIÓN

Una vez completado el entrenamiento y validación de los modelos planteados, se muestran a continuación los resultados obtenidos junto con las métricas de desempeño utilizadas en el proyecto para su evaluación.

	Modelo	Accuracy	Precision	Recall	F1-Score	Specificity
0	Logistic Regression	0.651913	0.378982	0.716093	0.495649	0.631773
1	Decision Tree	0.572198	0.329636	0.765315	0.460797	0.511597
2	XGBoost	0.685873	0.413400	0.752210	0.533564	0.665057
3	Random Forest	0.661377	0.390736	0.746876	0.513059	0.634546

Tabla 10. Resultados por modelo

Con base en los resultados obtenidos *(Tabla 10)*, se confirma que el modelo XGBoost continúa destacándose como el de mejor desempeño general en la predicción del abandono de clientes. Este modelo alcanzó la mayor puntuación en métricas clave como Accuracy (0.685), Precision (0.413) y F1-Score (0.533), lo que evidencia su capacidad para lograr un equilibrio sólido entre la correcta identificación de clientes propensos a abandonar y la reducción de falsos positivos. Además, obtuvo un Recall de 0.752, lo que indica una alta sensibilidad en la detección de clientes que efectivamente se retiran del servicio, mientras que su Specificity (0.665) se mantuvo elevada, mostrando un buen control sobre los falsos positivos.

El modelo de Random Forest también presentó un rendimiento competitivo, destacándose especialmente por su estabilidad en métricas como Recall (0.746), F1-Score (0.513) y Specificity (0.634) *(Tabla 10)*. Aunque no superó a XGBoost, mostró un desempeño equilibrado y consistente, lo que lo convierte en una opción viable cuando se prioriza la robustez del modelo.

Por otro lado, el modelo de Árbol de Decisión logró el mayor valor de Recall (0.765), lo que indica que fue el más eficaz al momento de detectar clientes que abandonan. No obstante, sus valores reducidos de Precision (0.329) y Specificity (0.511) *(Tabla 10)* revelan una elevada proporción de falsos positivos, lo cual podría generar esfuerzos de retención innecesarios. En cuanto a la Regresión Logística, si bien no lidera en ninguna métrica, mantiene resultados estables en todas ellas, con un Recall de 0.716 y Specificity de 0.631 *(Tabla 10)*, siendo su principal ventaja la interpretabilidad y facilidad de implementación.

Dado el desbalance de clases presente en el dataset, se priorizaron las métricas de Recall y F1-Score en la evaluación del rendimiento, ya que estas permiten medir con mayor precisión la capacidad del modelo para identificar clientes en riesgo de abandono sin comprometer excesivamente la tasa de falsos positivos. En este contexto, XGBoost se posiciona como la mejor alternativa para desplegar un sistema predictivo efectivo y confiable.

7.5.1 EVALUACIÓN DEL MODELO CON MEJOR RENDIMIENTO (XGBOOST)

El modelo XGBoost fue el que presentó el mejor desempeño general entre los modelos evaluados, destacándose especialmente en métricas clave como el F1-Score y el Recall. Aunque las diferencias respecto a otros algoritmos fueron moderadas, su equilibrio entre sensibilidad y especificidad lo posiciona como la mejor opción para la predicción del abandono de clientes. En particular, la Curva ROC evidencia una buena capacidad de discriminación del modelo, con un AUC (Área Bajo la Curva) de 0.79 (Figura 25), lo que indica que el modelo es capaz de distinguir correctamente entre clientes que abandonan y los que permanecen en aproximadamente el 79% de los casos.

En la figura 24 se presenta la matriz de confusión correspondiente al modelo XGBoost, utilizada para evaluar su desempeño sobre el conjunto de validación. En ella se observa que el modelo logró predecir correctamente a 13.907 clientes que no abandonaron el servicio (Verdaderos Negativos), así como a 4.936 clientes que sí abandonaron y fueron identificados correctamente (Verdaderos Positivos). No obstante, también se registraron 7.004 Falsos

Positivos, es decir, clientes que el modelo predijo erróneamente como desertores cuando en realidad permanecieron activos, y 1.626 Falsos Negativos, correspondientes a clientes que abandonaron pero que el modelo no logró anticipar.

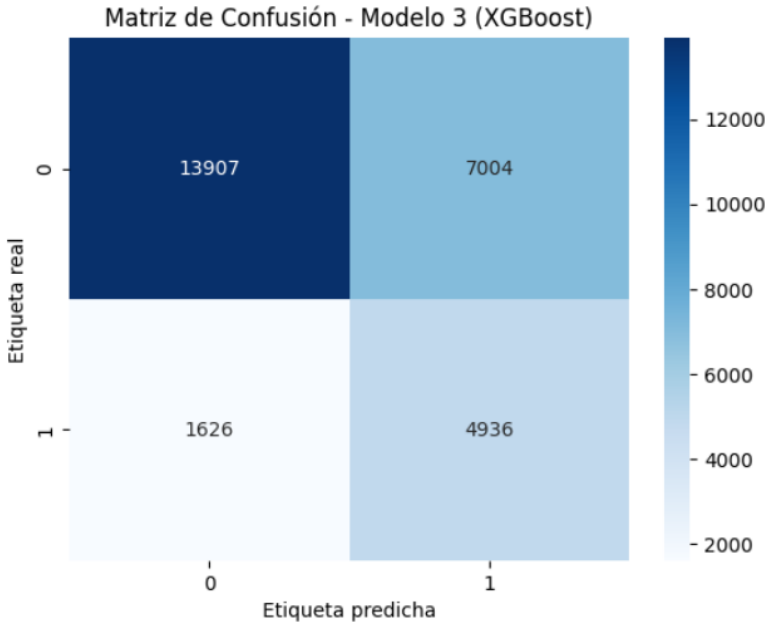

Figura 24. Matriz de confusión XGBoost

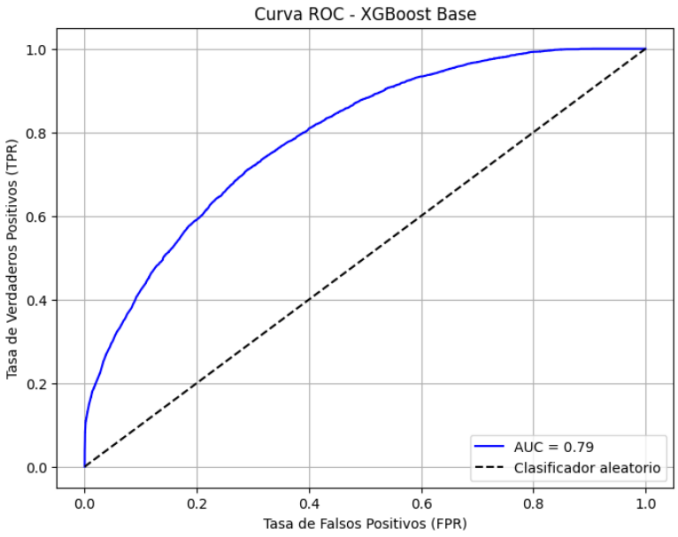

Figura 25. Curva ROC XGBoost

Este análisis permite evidenciar que el modelo mantiene una buena capacidad de detección de abandonos (Recall) sin comprometer de manera excesiva la tasa de falsos positivos, reflejando un balance adecuado entre sensibilidad y especificidad. Cabe destacar que la matriz fue generada a partir del conjunto de prueba, el cual representa el 20% del total de datos,

manteniendo la proporción original de clases gracias a la estrategia de estratificación aplicada durante la partición del dataset.

7.5.2 EVALUACIÓN DEL MODELO CON EL SEGUNDO MEJOR RENDIMIENTO (RANDOM FOREST)

Con base en la matriz de confusión y la curva ROC obtenidas para el modelo Random Forest, se puede concluir que este algoritmo presentó un desempeño sólido en la predicción del abandono de clientes, ubicándose como el segundo mejor modelo del análisis. La matriz de confusión (Figura 26) muestra que el modelo identificó correctamente a 4.901 clientes que efectivamente abandonaron (verdaderos positivos) y a 13.269 que no lo hicieron (verdaderos negativos), lo cual refleja una adecuada capacidad de clasificación en ambos escenarios. Aunque se detectaron 7.642 falsos positivos y 1.661 falsos negativos, el equilibrio entre estos valores fue razonablemente manejado por el modelo.

En cuanto a la curva ROC, el área bajo la curva (AUC) fue de 0.76 *(Figura 27)*, lo que indica un buen rendimiento general. Este valor demuestra que el modelo logra mantener una buena separación entre las clases, siendo significativamente superior al umbral del azar representado por la diagonal (AUC = 0.5). El modelo consigue un balance entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR), manteniéndose consistentemente por encima de la línea de referencia, lo cual respalda su capacidad discriminativa.

En conjunto, tanto los valores observados en la matriz de confusión como el AUC obtenido reafirman que Random Forest es una opción confiable y robusta para identificar clientes propensos al abandono. Aunque fue superado levemente por XGBoost en ciertas métricas, su rendimiento balanceado lo convierte en una alternativa efectiva dentro del contexto del proyecto.

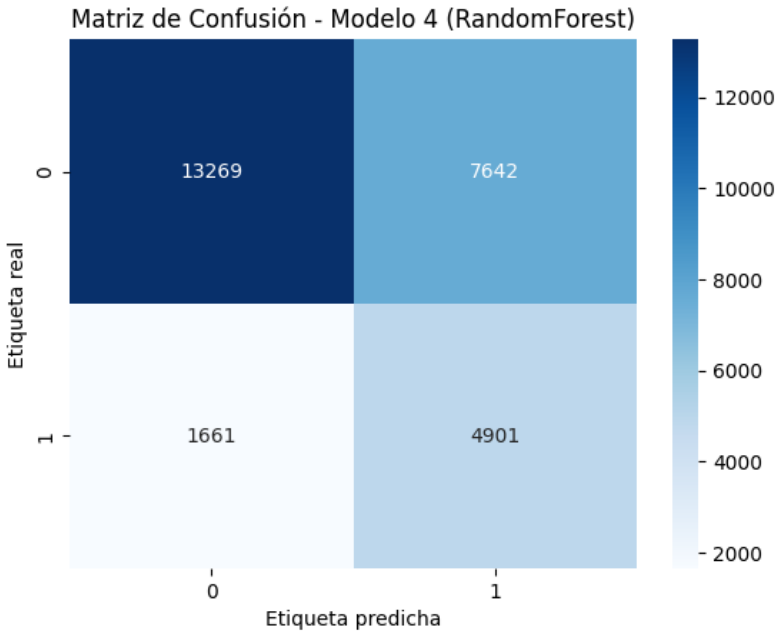

Figura 26. Matriz de confusión Random Forest

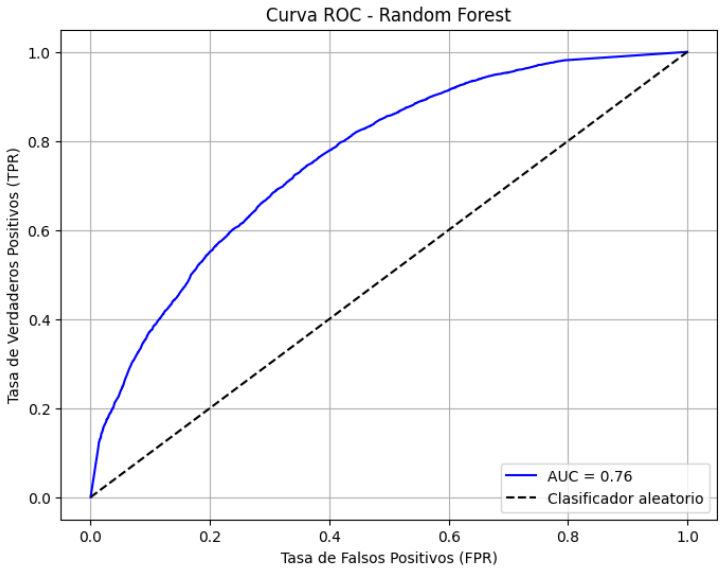

Figura 27. Curva ROC Random Forest

7.5.3 EVALUACIÓN DEL MODELO CON DESEMPEÑO INTERMEDIO (LOGISTIC REGRESSION)

Con base en la matriz de confusión y la curva ROC, el modelo de Regresión Logística se posiciona como el tercer mejor modelo en la tarea de predicción del abandono de clientes. En la matriz de confusión (Figura 28) se observa que logró clasificar correctamente a 4.699 clientes que efectivamente abandonaron el servicio (verdaderos positivos) y a 13.211 que no

lo hicieron (verdaderos negativos), aunque también incurrió en 7.700 falsos positivos y 1.863 falsos negativos. Esto indica una capacidad moderada para identificar a los clientes en riesgo, pero con una tendencia a generar algunas alertas innecesarias.

La curva ROC (Figura 29) para este modelo muestra un área bajo la curva (AUC) de 0.74, lo cual sugiere un rendimiento general aceptable. Aunque no tan alto como el obtenido por XGBoost (0.79) o Random Forest (0.76), este valor indica que el modelo tiene una buena habilidad para diferenciar entre los clientes que abandonan y los que permanecen activos.

Si bien la Regresión Logística no fue el modelo con mejor desempeño en términos de precisión o sensibilidad, ofrece una ventaja importante: su interpretabilidad. Esta característica puede ser de gran valor en contextos donde se requiere comprender con claridad cuáles variables están influyendo en la predicción. Por lo tanto, su inclusión como tercer mejor modelo se justifica no solo por su rendimiento cuantitativo, sino también por el valor cualitativo que aporta en la interpretación y transparencia del modelo predictivo.

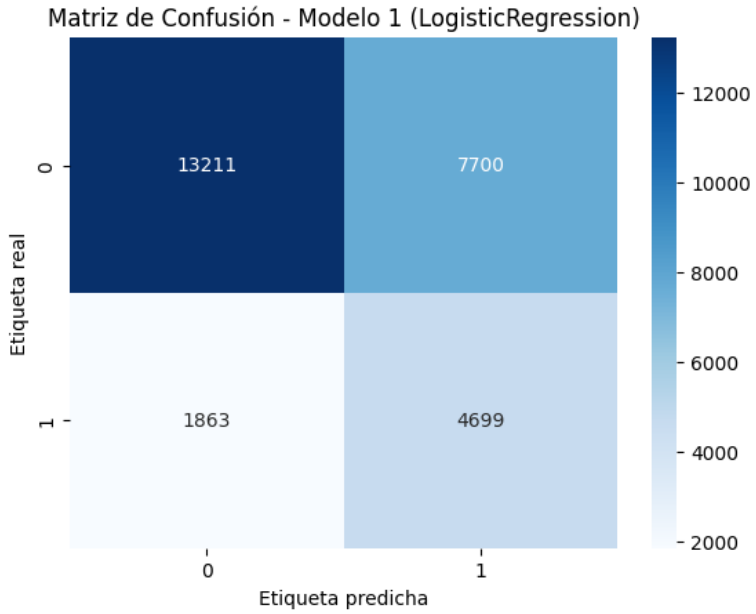

Figura 28. Matriz de confusión Logistic Regression

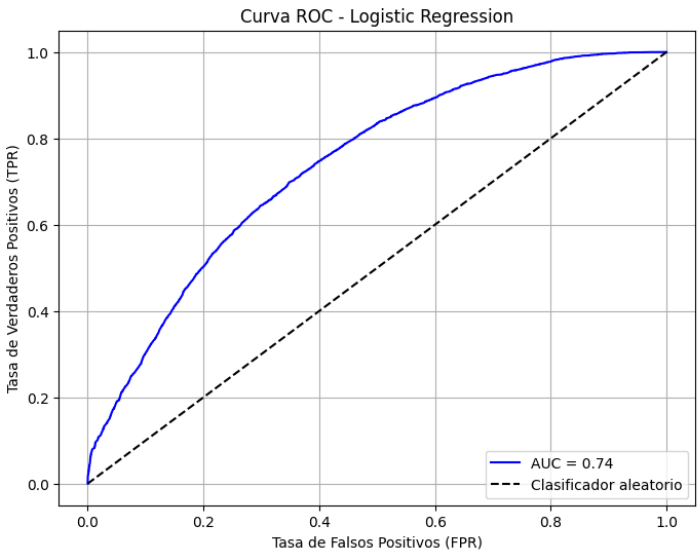

Figura 29. Curva ROC Logistic Regression

7.5.4 EVALUACIÓN DEL MODELO CON MENOR RENDIMIENTO (DECISION TREE)

Con base en los resultados obtenidos por el modelo Decision Tree, se evidencia que fue el de menor rendimiento entre los evaluados. A pesar de que logró identificar correctamente una proporción significativa de clientes que abandonaron (Recall de 0.765), lo hizo a expensas de una alta tasa de falsos positivos, lo cual se ve reflejado en su baja precisión (0.329) y en su baja especificidad (0.511) (Tabla 10). Esto significa que clasificó incorrectamente a muchos clientes activos como si fueran propensos a abandonar, lo que podría derivar en intervenciones innecesarias y mal uso de recursos.

Además, la matriz de confusión (Figura 30) muestra que el modelo etiquetó erróneamente a más de 10.000 clientes que no abandonaron como si lo fueran (falsos positivos), y aunque también detectó correctamente a 5.022 clientes que sí abandonaron (verdaderos positivos), el número de errores sigue siendo significativo. Esto indica un desbalance en la capacidad del modelo para diferenciar correctamente entre ambas clases.

Por último, la curva ROC (Figura 31) refuerza este diagnóstico. El área bajo la curva (AUC) fue de 0.66, la más baja de todos los modelos probados, lo que evidencia una menor capacidad discriminativa. En conjunto, estos resultados posicionan al Decision Tree como la opción menos adecuada para el problema de predicción de abandono en este caso, debido a su bajo desempeño general y pobre equilibrio entre sensibilidad y precisión.

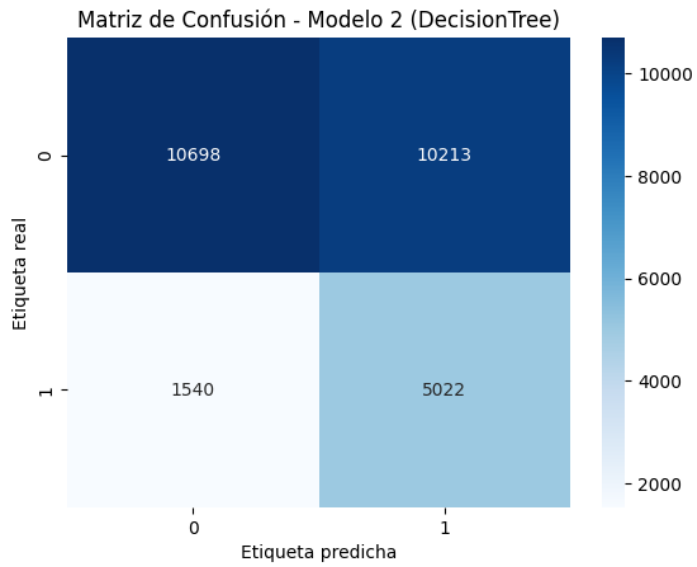

Figura 30. Matriz de confusión Decision Tree

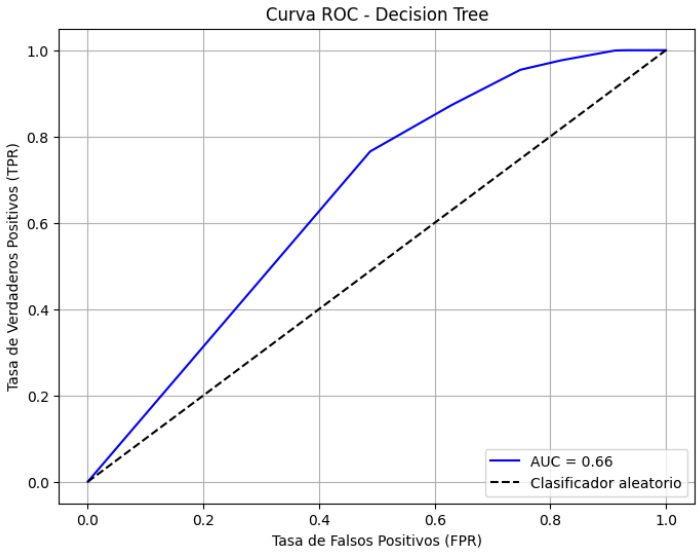

Figura 31. Curva ROC Decision Tree

7.5.5 AJUSTE DE HIPERPARAMETROS PARA EL MEJOR MODELO (XGBOOST)

Para mejorar el desempeño del modelo XGBoost, se realizó un ajuste de hiperparámetros mediante validación cruzada. Los hiperparámetros evaluados fueron los siguientes:

- **n_estimators** (número de árboles): permite controlar la complejidad del modelo; se evaluaron valores moderados para evitar sobreajuste y mejorar la generalización.

- **max_depth** (profundidad máxima): regula la complejidad de cada árbol; árboles más profundos pueden capturar patrones más complejos, pero también incrementan el riesgo de sobreajuste.
- **learning_rate** (tasa de aprendizaje): controla la contribución de cada árbol al modelo final; tasas más bajas permiten un aprendizaje más gradual y preciso.
- **subsample** (fracción de muestras por árbol): introduce aleatoriedad al usar solo una parte del conjunto de entrenamiento, lo que ayuda a reducir el sobreajuste.
- **scale_pos_weight (peso para la clase positiva):** ajusta la importancia de la clase minoritaria en problemas desbalanceados; al aumentar este valor se le da más peso a los errores en la clase positiva, lo que mejora el recall y el F1-score en contextos donde detectar esa clase es prioritario (por ejemplo, clientes que hacen churn).

Se empleó la técnica de GridSearchCV, una búsqueda sistemática que permite encontrar la mejor combinación de hiperparámetros para un modelo de machine learning. Esta técnica evalúa todas las combinaciones posibles dentro de una malla (grid) predefinida y selecciona aquella que ofrece el mejor desempeño según una métrica específica; en este caso, el recall. Además, utiliza validación cruzada para asegurar que los resultados sean robustos y no dependan de una sola partición de los datos.

Los principales parámetros utilizados fueron los siguientes:

- **estimator=model_3**: define el modelo base (XGBoost) sobre el cual se realizará la búsqueda de hiperparámetros.
- **param_grid=param_grid_xgb**: especifica el conjunto de combinaciones de hiperparámetros que se evaluarán.
- **scoring='recall'**: selecciona la métrica a optimizar durante la búsqueda. Se eligió *recall* por ser clave en problemas de churn, donde es más importante detectar correctamente a los clientes que abandonan.
- **cv=3**: aplica validación cruzada con 3 particiones, lo que permite evaluar el desempeño del modelo de forma robusta y con un costo computacional razonable.
- **verbose=1**: muestra el progreso del proceso de búsqueda en la consola, útil para monitorear la ejecución.
- **n_jobs=-1**: utiliza todos los núcleos disponibles del procesador para paralelizar las búsquedas, acelerando el tiempo de ejecución.

La decisión de utilizar validación cruzada con tres folds, en lugar de un número mayor, se fundamenta principalmente en la relación entre el tamaño muestral, el tiempo computacional y la estabilidad de las métricas. Aunque se cuenta con un conjunto de datos considerable, el entrenamiento de modelos como XGBoost implica una alta demanda computacional, especialmente cuando se realizan múltiples iteraciones de ajuste de hiperparámetros y se optimiza con métricas específicas como el recall, F1-score, etc. En este contexto, se priorizó un balance entre obtener una estimación suficientemente representativa de la generalización del

modelo, así mismo mantener un tiempo de entrenamiento razonable para poder iterar y ajustar parámetros sin comprometer recursos ni extender excesivamente los tiempos de análisis.

En la ejecución, el tiempo promedio de entrenamiento por iteración utilizando 3 folds fue de aproximadamente 15 minutos, mientras que al aumentar a 5 folds, dicho tiempo se incrementó a cerca de 40 minutos por iteración. Además, la variabilidad de las métricas entre folds se mantuvo baja, con una desviación estándar del F1-score de solo 0.03, lo que indica una estabilidad suficiente en los resultados sin necesidad de incrementar el número de folds y, por ende, el costo computacional.

Estudios previos (por ejemplo, Kohavi [14]) sugieren que, cuando se dispone de una muestra grande, usar menos folds (como 3 o 5) es suficiente para estabilizar las métricas, ya que la varianza disminuye gracias al tamaño muestral y aumentar el número de folds solo reduce ligeramente el sesgo, mientras el costo computacional se incrementa considerablemente. Por eso, se optó por usar tres folds como un punto intermedio que permitiera validar correctamente el modelo, facilitar las pruebas y ajustar los parámetros de forma continua.

Una vez definida esta estrategia de validación y optimización, se procedió a correr el modelo utilizando GridSearch y los hiperparámetros seleccionados, obteniéndose los siguientes resultados:

Reporte de clasificación para XGBoost ajustado:

	precision	recall	f1-score	support
0	0.89	0.70	0.78	20911
1	0.43	0.72	0.54	6562
accuracy			0.70	27473
macro avg	0.66	0.71	0.66	27473
weighted avg	0.78	0.70	0.72	27473

Tabla 11. Resultados de métricas para XGBoost ajustado

El modelo XGBoost ajustado logró un recall del 72% para la clase de clientes que abandonan, cumpliendo el objetivo principal de identificar correctamente estos casos. La precisión fue del 43%, indicando una mayor tasa de falsos positivos, pero aceptable en contextos donde es preferible intervenir antes que no actuar. El modelo alcanzó una accuracy del 71% y un F1-score de 0.54 para la clase positiva, reflejando un buen equilibrio general. (Tabla 11)

Se realizó una comparación entre el modelo base y el modelo ajustado mediante hiperparámetros, la cual se presenta en la siguiente tabla.

	Modelo	Accuracy	Precision	Recall	F1-Score	Specificity
0	XGBoost (base)	0.687220	0.415071	0.756324	0.535990	0.665535
1	XGBoost (ajustado)	0.703345	0.428301	0.722798	0.537877	0.697241

Tabla 12. Resultados de métricas para XGBoost ajustado vs XGBoost base

La comparación (Tabla 12) muestra que el modelo ajustado por hiperparámetros obtuvo mejores resultados en accuracy, precisión, F1-score y specificity, logrando un desempeño más equilibrado. Aunque el recall disminuyó ligeramente, se mantiene en un nivel adecuado (72.2%) para detectar clientes que abandonan. En un contexto de churn, este ajuste es favorable, ya que mejora la precisión sin sacrificar significativamente la sensibilidad.

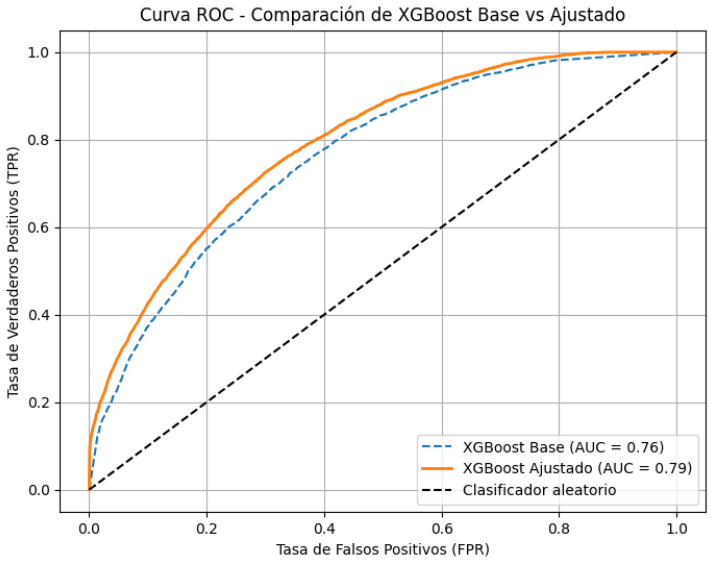

Figura 32. Curva ROC XGBoost base vs Ajustado

La curva ROC (Figura32) muestra que el modelo XGBoost ajustado supera al modelo base, alcanzando un AUC de 0.79 frente a 0.76. Esto indica una mejor capacidad del modelo ajustado para diferenciar entre clientes que abandonan y los que no. El ajuste de hiperparámetros permitió mejorar la discriminación del modelo, optimizando su rendimiento global sin comprometer la estabilidad.

7.5.6 SEGMENTACIÓN DE CLIENTES

7.5.6.1 CLUSTERING BIVARIADO UTILIZANDO ANTIGÜEDAD Y PROBABILIDAD DE CHURN

Para clasificar a los clientes en niveles de riesgo de cancelación, se utilizó el modelo de machine learning ajustado previamente. Este modelo generó una probabilidad de churn para cada cliente del conjunto de test, es decir, una estimación de la probabilidad de que un cliente cancele el servicio.

Posteriormente, se definió una función de tipificación que asigna un nivel de riesgo basado en estos valores probabilísticos:

- **Muy alto**: Probabilidad mayor o igual a 0.8
- **Alto**: Probabilidad entre 0.6 y 0.8

- **Medio**: Probabilidad entre 0.4 y 0.6
- **Bajo**: Probabilidad entre 0.2 y 0.4
- **Muy bajo**: Probabilidad menor a 0.2

Cada cliente fue clasificado de manera automática en una de estas cinco categorías, lo que permite segmentar la base de clientes en función de su propensión al churn. Esta tipificación facilita el diseño de estrategias específicas de retención enfocadas en los grupos de mayor riesgo.

Luego de tipificar a los clientes según su probabilidad de churn, se realizó un análisis específico sobre aquellos clientes que efectivamente cancelaron el servicio (churn = 1). Se contabilizó la cantidad de clientes en cada nivel de riesgo (Muy bajo, Bajo, Medio, Alto y Muy alto).

El total de clientes que realizaron churn fue de 6,562. De estos, el 84.72% correspondió a los niveles de riesgo Medio, Alto o Muy alto, es decir, categorías que representan un riesgo significativo de cancelación según el modelo.

Este resultado valida la capacidad predictiva de la tipificación, ya que una alta proporción de clientes que finalmente cancelaron ya habían sido identificados como clientes con riesgo moderado o alto. Esto permite enfocar esfuerzos de retención de manera más eficiente en los segmentos de mayor riesgo.

```
riesgo_churn
Alto        2598
Medio       1814
Muy alto    1147
Bajo         831
Muy bajo     172
Name: count, dtype: int64

Total de churn: 6562

Porcentaje de Alto, Muy alto y Medio: 84.72%
```

Tabla 13. Resultado de predicciones del modelo en el conjunto de test

Para segmentar a los clientes en grupos con características similares, se utilizó el algoritmo de K-Means basado en las variables de antigüedad y probabilidad de churn. Antes de aplicar el modelo, se realizó un escalamiento de las variables para asegurar que ambas tuvieran la misma importancia en el análisis.

Con el fin de determinar el número adecuado de clústeres (k), se aplicó el método del codo. Este método evalúa la "inercia", una medida de la distancia interna dentro de los clústeres, en función del número de clústeres. Al graficar esta relación, se identifica el punto donde la disminución de la inercia empieza a ser menos pronunciada, indicando el número óptimo de clústeres. Este punto de inflexión es conocido como el "codo".

Este análisis permitió seleccionar un valor de k que logra un balance entre precisión en la segmentación y simplicidad del modelo, optimizando así la identificación de patrones relevantes en los datos de clientes.

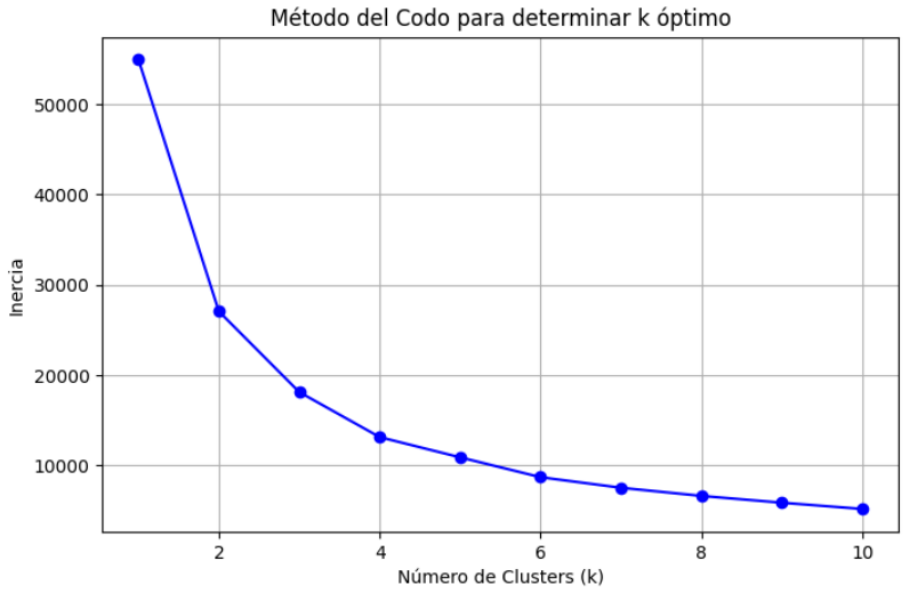

Figura 33. Método del codo para determinar número óptimo de clusters

Una vez determinado el número óptimo de clústeres mediante el método del codo, se procedió a aplicar el algoritmo de K-Means con k=4. Este método de aprendizaje no supervisado agrupa a los clientes en función de la antigüedad y la probabilidad de churn, previamente escaladas para asegurar que ambas variables tuvieran igual peso en la formación de los clústeres.

Cada cliente fue asignado a uno de los cuatro clústeres generados, lo que permite identificar patrones de comportamiento comunes dentro de los grupos. Para facilitar la interpretación de los resultados, se realizó una visualización en dos dimensiones, donde el color representa el clúster al que pertenece cada cliente. Esta segmentación proporciona una visión clara de los diferentes perfiles de clientes según su antigüedad y riesgo de cancelación, permitiendo diseñar estrategias más precisas y personalizadas de retención.

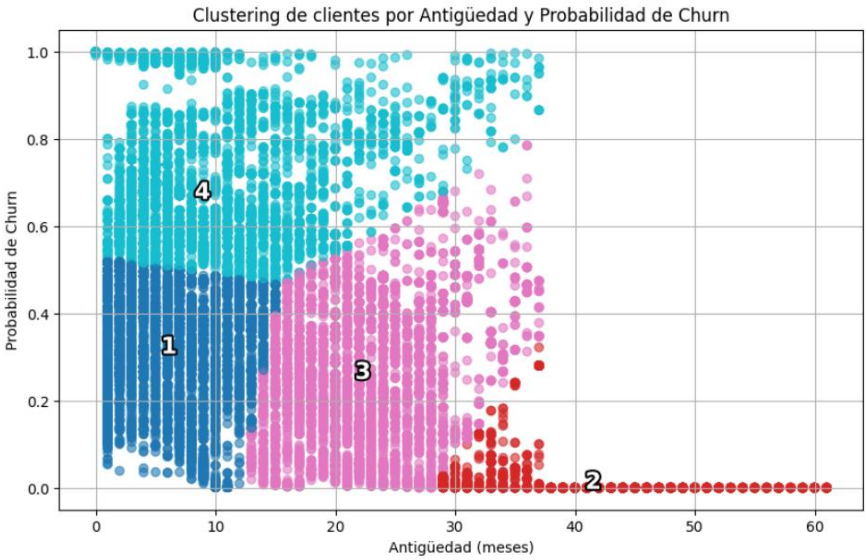

Figura 34. Segmentación de clientes por antigüedad y probabilidad de churn

La gráfica (Figura 34) muestra la segmentación de clientes utilizando el algoritmo K-Means, considerando las variables antigüedad (en meses) y probabilidad de churn. Cada color representa un clúster distinto.

Se observan patrones claros:

- Clientes con mayor antigüedad (más de 30 meses) tienden a presentar una probabilidad de churn baja, concentrándose en un clúster homogéneo con baja dispersión.
- Clientes con antigüedad intermedia (entre 15 y 30 meses) presentan mayor variabilidad en la probabilidad de churn, agrupándose en un clúster de riesgo moderado.
- Clientes más recientes (menos de 15 meses) se dividen en dos grupos: uno con baja probabilidad de churn y otro con probabilidad elevada, evidenciando una mayor propensión a cancelar el servicio en etapas tempranas de su ciclo de vida.

Esta segmentación permite identificar perfiles de riesgo diferenciados y enfocar estrategias de retención de manera más precisa, priorizando acciones en los grupos con mayor vulnerabilidad al churn.

7.5.6.2 CLUSTERING MULTIVARIABLE Y VISUALIZACIÓN CON ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

Además de la segmentación basada únicamente en antigüedad y probabilidad de churn, se amplió el análisis utilizando múltiples variables relevantes para caracterizar de manera más completa a los clientes. Se consideraron variables numéricas como antigüedad, precio del plan, probabilidad de churn, así como variables categóricas como tipo de cliente, estrato, canal de venta, tipo de plan y departamento.

Las variables categóricas fueron transformadas mediante codificación One-Hot Encoding y, posteriormente, todo el conjunto de datos fue escalado para asegurar que cada variable tuviera el mismo peso en el proceso de agrupamiento. Aplicando el algoritmo K-Means sobre este conjunto de datos multivariable, se segmentó a los clientes en cuatro grupos bien diferenciados. Sin embargo, dada la alta dimensionalidad de los datos, se utilizó el Análisis de Componentes Principales (PCA) para reducir la información a dos componentes principales, facilitando así su visualización en un espacio bidimensional.

La gráfica (Figura 35) obtenida muestra la distribución de los clientes en función de estos dos componentes principales, donde cada color representa un clúster distinto. Se observan formaciones definidas y patrones de agrupamiento claros, lo que confirma que los clientes presentan similitudes en sus características demográficas, comerciales y de comportamiento.

El uso de PCA no solo permitió una visualización clara de la segmentación, sino que también evidenció la existencia de estructuras naturales en los datos, reforzando la solidez de la segmentación multivariable realizada. Esta segmentación proporciona una comprensión profunda del perfil de los clientes y ofrece una base robusta para desarrollar estrategias de marketing y retención personalizadas, dirigidas a las necesidades específicas de cada grupo identificado.

En resumen, mediante el uso combinado de técnicas de clustering y reducción de dimensionalidad, se logró una segmentación precisa y visualmente interpretable de los clientes, optimizando el enfoque estratégico para la gestión del riesgo de cancelación.

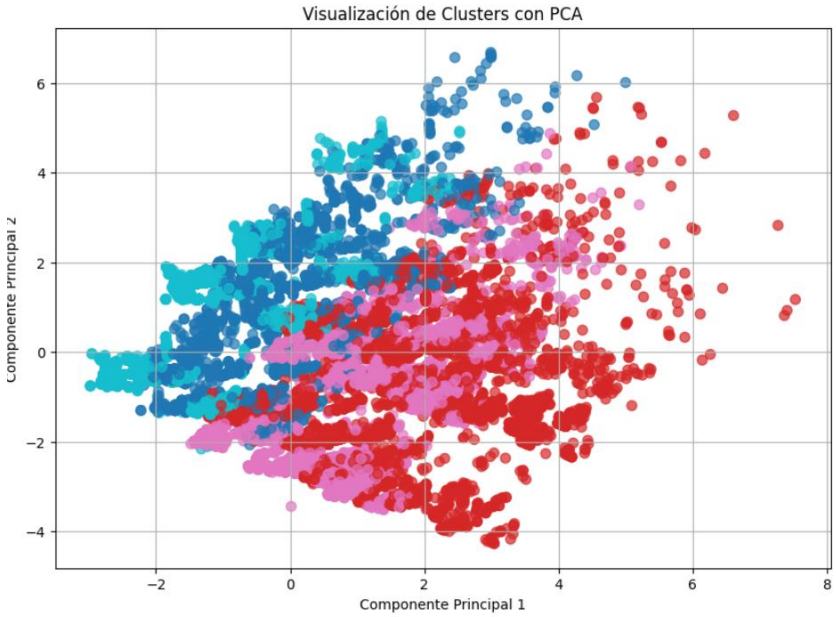

Figura 35. Análisis por componentes principales multivariado

7.5.6.3 ANÁLISIS DE PERFILES DE CLIENTES POR CLÚSTER

A partir del análisis multivariable, se identificaron cuatro clústeres de clientes con características diferenciadas:

- **Clúster 0:** Clientes inquilinos con planes de velocidad media (200/200) y una antigüedad promedio de 15.25 meses. Su probabilidad de churn es relativamente alta (45.44%), indicando un grupo con riesgo moderado-alto de cancelación.
- **Clúster 1:** Clientes inquilinos que cuentan con los planes de mayor capacidad (400/400). A pesar de tener una antigüedad similar (14.10 meses), presentan una probabilidad de churn más baja (39.32%) comparada con otros grupos de inquilinos, lo cual podría asociarse al beneficio percibido por el alto ancho de banda contratado.
- **Clúster 2:** Clientes inquilinos con los planes más básicos (100/100) y la menor antigüedad promedio (11.76 meses). Este grupo exhibe la mayor probabilidad de churn (47.50%),

indicando alta vulnerabilidad, posiblemente debido a insatisfacción temprana o menor fidelización.

- **Clúster 3:** Clientes propietarios con planes de velocidad media (200/200) y una antigüedad promedio de 14.68 meses. Son el grupo con la probabilidad de churn más baja (29.76%), reflejando mayor estabilidad y fidelidad en comparación con los inquilinos.

Cluster	Tipo de Cliente	Canal de Venta	Estrato	Tipo de Plan	Departamento	Antigüedad (meses)	Precio del Plan	Probabilidad de Churn
0	Inquilino	PAP	2	200/200	Valle del Cauca	15.25	68,455.14	0.4544
1	Inquilino	PAP	2	400/400	Valle del Cauca	14.10	99,205.01	0.3932
2	Inquilino	PAP	2	100/100	Valle del Cauca	11.76	56,458.66	0.4750
3	Propietario	PAP	2	200/200	Valle del Cauca	14.68	63,332.37	0.2976

Tabla 14. Perfilamiento de clientes

El análisis revela que los propietarios tienden a ser más estables, mientras que los *inquilinos* con menor capacidad contratada presentan mayor riesgo de cancelar el servicio. Estos resultados permiten diseñar estrategias de retención específicas para los grupos más vulnerables, enfocadas en incrementar la satisfacción de los inquilinos y consolidar la lealtad de los propietarios.

7.6 IMPLEMENTACIÓN

En esta última fase se plantea la puesta en marcha del modelo seleccionado en un entorno funcional que permita su uso práctico para generar alertas tempranas de abandono de clientes. El objetivo principal es contar con una herramienta automatizada que clasifique a los clientes según su riesgo de fuga y facilite la toma de decisiones oportunas por parte del área comercial.

Se prevé implementar el modelo de XGBoost, el cual fue identificado como el más adecuado por su buen rendimiento en métricas clave como Recall, F1-Score y Specificity. Para ello, se propone desarrollar un flujo automatizado con los siguientes componentes:

1. Carga de nuevos datos de clientes, provenientes de las bases de datos operativas.
2. Preprocesamiento automático, replicando las transformaciones realizadas en la etapa de entrenamiento (tratamiento de valores nulos, codificación de variables categóricas, etc.).
3. Generación de predicciones, donde el modelo evaluará el riesgo de abandono de cada cliente.
4. Clasificación por niveles de riesgo (alto, medio o bajo) según los umbrales definidos.
5. Visualización de resultados en un tablero de control, lo cual permitirá monitorear en tiempo real el estado de la base de clientes.

6. Almacenamiento de resultados y retroalimentación periódica, con el fin de mantener actualizado el modelo mediante reentrenamiento cada cierto periodo (por ejemplo, trimestralmente).

Este flujo podrá ser ejecutado de forma local en una primera etapa piloto, y posteriormente adaptado a un entorno en la nube o sistema integrado si se desea escalar su uso.

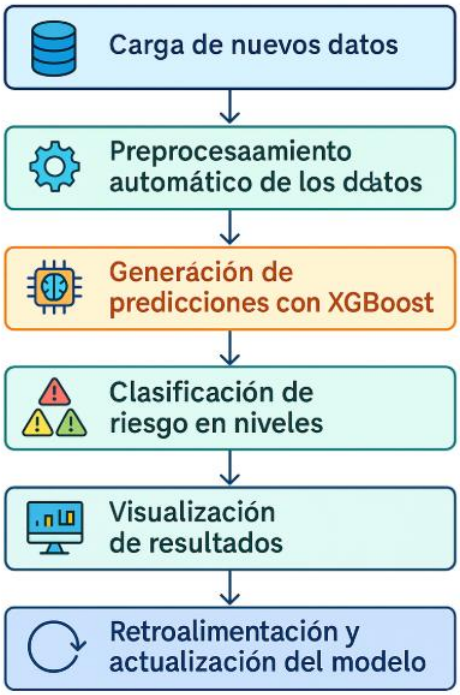

Figura 36. Flujo de implementación

8 CONCLUSIONES Y TRABAJOS FUTUROS

8.1 CONCLUSIONES

La ejecución del proyecto inicialmente y durante el análisis exploratorio ha evidenciado que la compañía debe trabajar en una política de gobierno de datos, actualmente no se cuenta con ello, en los datos se presentaron datos nulos en algunas de las variables que debieron ser imputados por la moda en el caso de las categóricas y el promedio para las numéricas, es necesario que se realice una normalización de los datos en la empresa y se garantice que esto no siga ocurriendo, también se logró evidenciar que la información se encuentra distribuida en distintos reportes, los cuales debieron ser unificados, se tuvieron dificultades para relacionar las bases de datos, dado que la llave primaria no estaba totalmente clara.

Con respecto al desarrollo de los modelos, se probaron 4 opciones, Regresión logística, Arboles de decisión, Random Forest y XGBoost, siendo este último es el que ha presentado los mejores resultados, con un Recall del 72% y una especificidad del 66%, se eligieron estas variables como las determinantes dadas las necesidades del proyecto, es importante que el modelo elegido sea el más acertado con la identificación de verdaderos positivos, porque puede ser la diferencia entre un cliente que se pueda mantener o uno fugado.

Análisis del impacto económico y justificación del umbral de decisión:

RESULTADO DE LA MATRIZ DE CONFUSIÓN CON EL MODELO XGBOOST:

	Predicho 0	**Predicho 1**
Real 0 (No abandona)	13.907	7.004
Real 1 (Si abandona)	1.626	4.936

- Falsos positivos (Error tipo I): **7.004**
- Falsos negativos (Error tipo II): **1.626**

COSTOS INVOLUCRADOS:

- **Costo de adquirir un nuevo cliente**: $1.200.000 COP
- **Costo de aplicar una campaña de retención innecesaria**:
  Se asume que el cliente migra de un plan promedio de $66.521 (valor promedio del ticket) a uno subsidiado de $39.900, implicando una pérdida mensual de ingresos de $26.621 COP.
  Considerando un horizonte de 6 meses para evaluar la efectividad de la retención:
  $26.621 × 6 = $159.726 COP por cliente retenido innecesariamente.

IMPACTO ECONÓMICO DE LOS ERRORES:

- **Costo total por errores tipo I** (7.004 falsos positivos):
  $159.726 × 7.004 = **$1.118.074.104 COP**

- **Costo total por errores tipo II** (1.626 falsos negativos):
  $1.200.000 × 1.626 = **$1.951.200.000 COP**

COMPARACIÓN DEL COSTO POR CLIENTE:

Costo por Falso Negativo = $1.200.000 = **7.51**
Costo por Falso Positivo $159.726

Por lo anterior concluimos que el costo de **no detectar a un cliente que abandona** (falso negativo) es **7.51 veces más alto por cliente** que el de aplicar una retención innecesaria (falso positivo). A pesar de tener menos falsos negativos en número absoluto, su impacto económico es **significativamente mayor**.

RELACIÓN CON LAS MÉTRICAS DE EVALUACIÓN:

El modelo alcanza una combinación equilibrada entre Precisión y Recall, aunque por el impacto económico, se prioriza minimizar los falsos negativos:

Precisión (clientes clasificados como en riesgo que efectivamente abandonan**):**

TP = True positives
FP = False positives
FN = False negatives

$$\frac{TP}{TP + FP} = \frac{4.936}{4.936 + 7.004} \approx 0.413$$

Recall (clientes que efectivamente abandonan que fueron detectados):

$$\frac{TP}{TP + FN} = \frac{4.936}{4.936 + 1.626} \approx 0.752$$

Para reflejar esta prioridad, usamos el Recall como métrica principal, le dimos mayor peso frente a la Precisión. Esto está alineado con el hecho de que **perder un cliente cuesta mucho más que aplicar una retención innecesaria**.

JUSTIFICACIÓN DEL UMBRAL DE DECISIÓN:

El umbral utilizado (implícito en esta matriz) permite un compromiso razonable: identificar correctamente a más del 75.2% de los clientes en riesgo de abandono, incluso a costa de algunas retenciones innecesarias. Esta elección se justifica objetivamente desde una perspectiva de maximización de valor económico, más allá de los indicadores estadísticos puros.

La compañía en adelante podrá usar esta información para anticiparse al abandono de un cliente y desde las áreas de Experiencia al Cliente y Mercadeo, diseñar estrategias enfocadas a la retención de clientes e iniciar un seguimiento al churn-rate en búsqueda de una disminución.

8.2 TRABAJOS FUTUROS

Es posible considerar un nuevo análisis una vez la compañía haya ejecutado un proceso de normalización de datos y cuente con una base más depurada. Adicionalmente, se recomienda evaluar las posibilidades técnicas de integración con el sistema OSS, de modo que los asesores del Contact Center puedan acceder a esta información durante las llamadas. Si el asesor dispone en tiempo real de datos sobre la probabilidad de abandono del cliente en línea, sería posible adaptar las técnicas de atención y aplicar intervenciones oportunas para prevenir la fuga.

Trabajar de manera articulada con COLOMBIA INTERNET ISP en la formulación de una política de gobierno de datos es fundamental para establecer lineamientos claros sobre la calidad, integridad, seguridad y disponibilidad de la información. Esta política debe contemplar también los ajustes estructurales necesarios en las bases de datos, con el fin de garantizar que la información relevante esté unificada, actualizada y sea confiable para los procesos analíticos. Durante este proceso, se recomienda evaluar la inclusión de nuevas variables que podrían enriquecer el análisis del churn-rate. Entre estas se destacan: sexo, cantidad de interacciones o solicitudes de mantenimiento, así como características del entorno del hogar, tales como número de personas con las que convive, motivo principal de uso del servicio (trabajo, estudio, entretenimiento, etc.), y la frecuencia de uso. Estos elementos permitirían construir perfiles más completos de los clientes y mejorar la capacidad predictiva del modelo.

9 REFERENCIAS BIBLIOGRÁFICAS

[1] P. &. C. J. Mathai, «Customer Churn Prediction: A Survey,» *International Journal of Advanced Research in Computer Science,* vol. 8, nº 5, 2017.

[2] F. &. G. L. Buttle, *Customer Lifetime Value,* 2015.

[3] R. I. Levin y D. S. Rubin, Estadística para administración y economía, Pearson Educación, 2010.

[4] R. &. N.-M. A. Caruana, «An Empirical Comparison of Supervised Learning Algotithms,» *Proceedings of the 23rd international conference on Machine learning - ICML '06,* pp. 161-168, 2006.

[5] M. Gopal, Applied Machine Learning, 1st Edition, 2019.

[6] L. Breiman, «Random forests,» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[7] A. S. G. A. y. A. S. A. O. Alshboul, «Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction,» *Sustainability,* vol. 14, nº 11, p. 6561, 2022.

[8] M. Narvaez, «Segmentación de Clientes: Qué es, Tipos y Ejemplos,» QuestionPro , [En línea]. Available: https://www.questionpro.com/blog/es/segmentacion-de-clientes/. [Último acceso: 10 06 2024].

[9] S. Cohen, «The basics of machine learning: strategies and techniques,» de *Artificial Intelligence and Deep Learning in Pathology*, S. Cohen, Ed., Elsevier, 2021, pp. 13-40.

[10] R. C. D. A. K. Jain, Algorithms for Clustering Data, Englewood Cliffs, NJ: Prentice Hall, 1988.

[11] A. F. E. Giraldo, *Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones,* Medellín: Universidad de Antioquia, Facultad de Ingeniería, Departamento de Ingeniería Electrónica y de Telecomunicaciones, 2019.

[12] S. F. S. Morales, *Modelo análisis predictivo para el cálculo de tasa de deserción en una empresa seguradora,* Bogotá: Fundación univeristaria Konard Lorenz, Facultad de postgrado, 2022.

[13] D. F. &. A. Motta, *Modelo de Abandono de Cliente en una Empresa de Créditos en Línea,* Bogotá: Tesis de Maestría, Departamento de Ingeniería Industrial, Universidad de lo Andes, 2019.

[14] R. A. Fisher, Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd, 1925.

[15] J. y. P. E. S. Neyman, «On the Problem of the Most Efficient Tests of Statistical Hypotheses,» *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,* vol. 231, nº 694–706, p. 289–337, 1933.

[16] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Boca Raton: Chapman & Hall/CRC, 2011.

[17] W. J. Conover, Practical Nonparametric Statistics, New York: Wiley, 1999.