Modelo de Aprendizaje Automático y Análisis de Factores de Abandono de Clientes para Mejorar su Retención en el Sector de Comercio Electrónico

Elaborado por:

Andrés David Camacho Arango

Diana Mireya Peña Sánchez

Olga Lucia Pabón Peña

Universidad Ean

Especialización en Machine Learning

Especialización en Gerencia de Calidad e Innovación

Seminario de Investigación

Bogotá

28/01/2026

Tabla de Contenido

Resumen

Este proyecto de investigación busca desarrollar un modelo de aprendizaje automático que permita predecir el abandono de clientes (churn) en el sector de comercio electrónico minorista de moda. Debido a que el abandono no se evidencia de forma explícita ni de manera inmediata en este tipo de comercio, el estudio busca analizar cuáles son las variables que tienen un mayor impacto en el abandono, con el fin de diseñar estrategias que optimicen los recursos de las campañas de retención. La metodología propuesta empleará datos fidedignos provenientes de mencionado sector y luego de anonimizarlos se procederá a emplear tres algoritmos de aprendizaje automático: Regresión Logística, Random Forest y XGBoost, ampliamente utilizados en diversos estudios. Realizada la validación experimental de mencionados algoritmos, el modelo que obtuvo los mejores valores en las métricas de desempeño fue el XGBoost, con una exactitud (Accuracy) del 96 %, un F1-Score de 97 % y un ROC-AUC de 99,33 % demostrando una alta capacidad para pronosticar el abandono; respecto al Random Forest que obtuvo una exactitud del 95 % y finalmente la Regresión Logística con una exactitud del 90 %. La precisión y sensibilidad sobresaliente del XGBoost evidencian su gran robustez brindando un soporte sólido y confiable para la formulación o diseño de estrategias de retención traduciéndolo en optimización de costos y aumento en la rentabilidad.

Palabras Clave

Predicción de Churn, Machine Learning, Ingeniería de Características, IA Explicable y Estrategias de Retención.

Planteamiento del Problema

Antecedentes del Problema

El comercio electrónico en Colombia según cifras del Observatorio de E-Commerce de MINTIC, es un sector en constate evolución teniendo en cuenta los cambios que presenta la economía mundial, se consolida como un canal de ventas fundamental en la economía digital. En 2023, el valor total de las ventas en plataformas digitales en el país alcanzó los COP 62.1 billones, esto significó un crecimiento del 12.58 % en comparación con el año anterior 2022 (COP 55.2 billones) (Ministerio de Tecnologías de la Información y las Comunicaciones (MINTIC), 2024). Este crecimiento elevado se evidencia en la cantidad de transacciones, que superó la meta del Plan Nacional de Desarrollo 2018-2022 (290 millones de transacciones digitales), registrándose 332.4 millones de transacciones en 2022 según cifras del Observatorio. Sin embargo, a medida que el mercado crece, continúan desafíos en la experiencia del cliente que justifican la necesidad de modelos predictivos, debido a que los consumidores aún manifiestan preocupaciones como la preferencia por la compra presencial (87.6 %), la desconfianza en los procesos logísticos y de devoluciones (55.5 %), y la inseguridad al proporcionar información bancaria en línea (55.1 %) (MINTIC, 2024). Por lo tanto, en este entorno de alto volumen transaccional y competencia, el desarrollo de un modelo de aprendizaje automático se vuelve crítico para analizar los patrones de comportamiento de los clientes y poder identificar de forma temprana y precisa a aquellos usuarios con alta probabilidad de abandono, asegurando así la retención de valor en la plataforma de comercio electrónico.

La pandemia COVID 19 tuvo impacto radical en el comportamiento de los consumidores en ecommerce, se generó un aumento significativo en las compras en plataformas digitales (Tariq et al., 2022), a raíz de esto muchas empresas enfrentan dificultades para adaptarse a las nuevas expectativas de los clientes, bien sea por recursos limitados o uso no optimo de los datos que recolectan de las transacciones, esta situación incrementó la necesidad de predecir el abandono de clientes como indicador clave de supervivencia empresarial. El churn implica que

un cliente deja de comprar o cancela una suscripción según la plataforma, lo cual afecta directamente la rentabilidad. El análisis de trazabilidad de los consumidores en los comercios, es decir, clics en anuncios, abandono del carrito, compras, descargas, permite identificar patrones de comportamiento asociados al abandono. La literatura refleja que la predicción de churn está directamente relacionada con la experiencia del cliente en plataformas digitales y sugiere la integración de machine learning y big data para anticipar el abandono y diseñar estrategias de retención más efectivas.

En un estudio reciente de vanguardia, (Asfe et al., 2025) se abordó la complejidad de los patrones de comportamiento en el comercio electrónico mediante la propuesta de una arquitectura denominada MNeuralTab. Esta investigación tuvo como objetivo superar las limitaciones de los modelos tradicionales (como la regresión logística) integrando múltiples redes neuronales profundas y TabNet bajo un enfoque de meta-modelo. Utilizando el conjunto de datos públicos de Olist Online (Brasil) y REES46, los autores aplicaron el Análisis de Componentes de Vecindad (NCA) para la selección de características relevantes. Los resultados obtenidos fueron sobresalientes: el modelo propuesto alcanzó una exactitud (Accuracy) del 99.62 %, una precisión del 99.32 % y un Área Bajo la Curva (AUC) de 0.98 sobre el conjunto de datos de Olist, superando significativamente a modelos base como XGBoost (98.35 % de exactitud) y Random Forest.

Descripción del Problema

En mercados altamente competitivos actualmente como lo es el comercio electrónico (ecommerce), los clientes cambian con facilidad de proveedor generando el fenómeno de deserción, retener clientes es más rentable que adquirir nuevos, pero las campañas de retención dependen de modelos de predicción de abandono de clientes (churn), estos modelos tradicionalmente se han optimizado con métricas estadísticas como precisión, F1 (media armónica) o área bajo la curva (AUC) sin considerar directamente los costos de clasificación

errónea ni los beneficios de una clasificación correcta (Liu et al., 2024). Esto provoca que, aunque los modelos sean precisos, no necesariamente maximicen la rentabilidad de las campañas. Además, la mayoría de las aplicaciones en predicción de churn no optimizan adecuadamente sus hiperparámetros, o lo hacen mediante métodos exhaustivos, lo que implica altos costos computacionales y resultados subóptimos. Se ha ignorado especialmente el peso de clase, un parámetro crítico para manejar el desbalance entre clientes que desertan y los que permanecen. En consecuencia, existe una brecha entre la capacidad predictiva de los modelos y su utilidad práctica para la maximización de beneficios empresariales.

En el comercio electrónico las reseñas y calificaciones a vendedores es un factor decisivo en el cliente para realizar una compra incluso es tan relevante como las reseñas de los productos. Varias investigaciones han abordado la participación de clientes y vendedores de forma unidimensional (Batta et al., 2023) es decir analizando un solo canal como redes sociales y plataformas de ventas sin considerar la interacción cruzada entre ambos.

El sector de ecommerce sufre de un problema crítico en la retención de clientes expresado en una alta tasa de churn (Pondel et al., 2021). Las tácticas existentes para retener son usualmente reactivas y son puestas en práctica sólo después de la pérdida del cliente. Existe la posibilidad abierta para que las empresas del sector busquen mejorar por medio de datos recabados por las transacciones de los usuarios, mediante la implementación de sistemas proactivos basados en aprendizaje automático (Machine Learning) y aprendizaje profundo (Deep Learning) capaces de manejar el volumen y la complejidad de los datos, identificar los factores de riesgo de churn con alta precisión y permitir la segmentación temprana de clientes en riesgo para aplicar intervenciones de retención personalizadas y optimizadas.

Los costos de adquisición de clientes (CAC) son significativamente más altos que los costos de retención, haciendo que esta sea una prioridad económica sobre la captación (Matuszelański & Kopczewska, 2022). Sin embargo, a diferencia de empresas con modelos de

suscripción donde el abandono es explícito, en los ecommerce el abandono es "silencioso"; los clientes simplemente dejan de comprar sin previo aviso, lo que dificulta su detección temprana.

Muchas empresas de ecommerce operan de manera reactiva, intentando recuperar clientes cuando ya ha pasado demasiado tiempo desde su última interacción. La falta de una identificación proactiva de los factores de riesgo impide la implementación de estrategias de

fidelización personalizadas y eficientes. Por tanto, existe la necesidad de desarrollar una solución analítica que no solo prediga quién se irá, sino que explique el por qué, permitiendo optimizar el valor de vida del cliente (CLV).

Pregunta de Investigación

¿Cuál es el modelo de aprendizaje automático más preciso para predecir el abandono de clientes en un entorno de comercio electrónico, y qué factores determinantes permiten diseñar estrategias proactivas de retención de clientes?

Objetivos

Objetivo General

Entrenar un modelo de aprendizaje automático a través del análisis de patrones de comportamiento de los clientes que permita identificar de forma temprana y precisa a los usuarios con alta probabilidad de abandono en la plataforma de comercio electrónico.

Objetivos Específicos

- Estructurar los datos de comportamiento de compra de clientes de la plataforma de comercio electrónico en el sector minorista de la moda con el fin de obtener los datos base para el entrenamiento del modelo.

- Entrenar un modelo de aprendizaje automático basado en técnicas de clasificación que a partir del análisis de variables determine cuales son las más influyentes en la probabilidad de que un cliente abandone el comercio electrónico.
- Analizar el rendimiento de los diferentes modelos de aprendizaje automático utilizando métricas de evaluación para definir el de mayor eficacia para la predicción de abandono de clientes.
- Formular un conjunto de recomendaciones estratégicas de retención de clientes basadas en el análisis de variables y las predicciones del modelo, que permita al comercio electrónico diseñar campañas de retención más efectivas.

Marco Teórico

Estado del Arte

Análisis Comparativo de Modelos de Agrupación para Predicción de Pérdida de Clientes

Un estudio reciente de (Boozary et al., 2025) comparó el desempeño de distintos modelos de aprendizaje automático para predecir el abandono de clientes (churn), con enfoque en los métodos de agrupación. Se utilizó un set de datos del sector telecomunicaciones de 10,000 registros, el estudio se realizó con modelos clásicos y avanzados.

Los resultados del estudio demostraron indicadores con alto desempeño en los métodos de agrupación, como Random Forest y XGBoost, sobre los clasificadores o modelos tradicionales. XGBoost obtuvo un alto rendimiento con una exactitud del 99.99 % y un área bajo la curva ROC (AUC) exacta de 1.0, demostrando una capacidad muy avanzada para identificar clientes en riesgo. El estudio también reveló que las variables más relevantes en el modelo para la predicción fueron, la antigüedad del cliente (tenure), el tipo de contrato y los cargos mensuales.

La investigación concluye que estos enfoques avanzados son muy efectivos para manejar datos complejos e inestables, proporcionando pronósticos confiables y asertivos para estrategias de retención.

Evaluación Integral de los Modelos de Aprendizaje Automático y Aprendizaje Profundo para la Predicción de la Rotación de Clientes

La investigación de (AbdelAziz et al., 2025) evaluó modelos de Machine Learning y Deep Learning para predecir el abandono de clientes en sectores como seguros, telecomunicaciones y servicios de internet. Se compararon tres arquitecturas relevantes: el tradicional XGBoost, una red neuronal convolucional (CNN) y un modelo híbrido de Ensemble Deep Learning.

Los resultados mostraron que el rendimiento óptimo cambia según el tipo de sector. En seguros, el modelo de Ensemble Deep Learning obtuvo el valor más alto en precisión (95.96 %). Para los proveedores de internet, el modelo de XGBoost fue el más efectivo (95.36 %). En el sector de telecomunicaciones, tanto CNN como XGBoost alcanzaron el mejor desempeño, con una precisión y puntuación F1 del 98.42 %.

La investigación concluye que los métodos de predicción de abandono, especialmente el enfoque de Ensemble Deep Learning, XGBoost y CNN son eficientes en set de datos de seguros, comunicaciones y servicios de internet, sin embargo, hay limitaciones en factores externos como tendencias del mercado y condición económica, debido a la falta de datos disponibles que puede limitar la generalización de modelos, se sugiere la incorporación de variables externas para futuras investigaciones.

Comparación de Métodos para Manejar Datos Desbalanceados en la Predicción de CHURN con Selección de Características Utilizando los Frameworks SHAP y mRMR

El estudio (Tam et al., 2025) comparó métodos para tratar datos desbalanceados en la predicción de la deserción de clientes (churn) en los sectores bancario y de comercio

electrónico. Los investigadores aplicaron una estrategia de selección de características utilizando los métodos SHAP y mRMR, y luego evaluaron distintas técnicas de balanceo, como métodos de remuestreo (Oversampling, Undersampling e híbridos) y el ajuste de ponderación de clases (Class Weight) dentro de los algoritmos, probándolos con modelos individuales y de conjunto.

Los principales hallazgos indican que los modelos de conjunto (como Bagging y Boosting) mostraron mayor robustez frente al desbalance de datos, superando consistentemente a los modelos únicos. En cuanto a las técnicas de balanceo, el Oversampling presentó el mejor rendimiento general para los conjuntos de datos medianos analizados, mientras que el Undersampling mejoró el Recall a costa de reducir la Precision. El ajuste de Class Weight también demostró ser altamente efectivo.

Dos modelos destacaron por su desempeño excepcional, logrando métricas como Accuracy, Precision, Recall y F1-score superiores a 0.9: ROS-CatBoost (que combina Oversampling con CatBoost) y CW-XGBoost (que utiliza ponderación de clases con XGBoost). Para abordar la falta de transparencia típica de estos modelos de conjunto, el estudio propone emplear el marco SHAP de IA explicable, el cual permite generar interpretaciones tanto globales como individuales de las predicciones realizadas.

Aunque es conocido que los modelos como XGBoost y Random Forest, son algoritmos excelentes para pronosticar cuando un cliente dejará de comprar, todavía se presentan fallas porque no se cuenta con datos completos o se presentan desordenes en la información lo cual limita la aplicación de dichos algoritmos en el mundo real. Por tanto, es necesario integrar la tecnología con la toma de decisiones basadas en la calidad y la innovación.

Calidad e Innovación en la analítica predictiva

Es importante resaltar que la analítica predictiva si bien es una herramienta técnica que sirve para modelar datos, también conduce a la innovación de procesos los cuales pueden

robustecer la calidad del servicio en el comercio electrónico. Con base a este criterio, la pérdida o deserción de un cliente puede revelar una falla en la misma requiriendo tomar acciones estratégicas que impliquen asegurar el éxito del negocio. El ciclo de mejora continua PHVA (Planear, hacer, verificar y actuar), diseñado por Deming, busca la toma de acciones preventivas y correctivas que permitan trasformar datos en decisiones estratégicas, mejorando la eficiencia en el uso de los recursos disminuyendo costos derivados de las fallas de calidad como lo cita el manual de OSLO (OECD, 2018) y además generando ventajas competitivas que fortalecen la calidad y promueven la innovación (Casanova-Villalba et al., 2023).

Marco Conceptual

Fundamentos del Fenómeno de Estudio

Conceptualización y Tipología del Abandono (churn) en Ecommerce.

El abandono de clientes, comúnmente denominado churn, es una métrica fundamental en la gestión de relaciones con el cliente o Customer Relationship Management (CRM). Sin embargo, su conceptualización varía drásticamente según el modelo de negocio. Según la distinción clásica de (Fader & Hardie, 2007) los entornos comerciales se dividen en contractuales y no contractuales. En un escenario contractual (ej. Telecomunicaciones, banca, streaming), el abandono es observable: el cliente cancela el servicio. Por el contrario, en el comercio electrónico minorista (retail), el entorno es no contractual; el abandono es un evento "silencioso" y latente, donde el cliente simplemente deja de realizar transacciones sin previo aviso.

En este contexto, el abandono de clientes se convierte en un desafío probabilidad. (Matuszelański & Kopczewska, 2022) argumentan que, en el comercio electrónico, las tasas de retención pueden ser muy bajas con cifras de un solo dígito en la mayoría de los casos, lo que

obliga a redefinir el abandono no como la cancelación de un servicio, sino como la ausencia de una siguiente compra tras una transacción anterior que en la mayoría de los casos es la primera que realiza el cliente. Esta distinción es fundamental para esta investigación, pues desplaza el foco de la investigación desde la prevención de cancelaciones hacia la predicción de la probabilidad de recompra.

Impacto Económico: Relación CLV (Customer Lifetime Value) vs Costos de Adquisición (CAC).

El Valor de Vida del Cliente (CLV) es una métrica clave para evaluar la contribución económica de un cliente a lo largo de toda su relación con una empresa, incluyendo desde su primera compra hasta todas las transacciones posteriores. Esta medida permite entender que tan rentable es el cliente.

Otra métrica importante es el Costo de Adquisición de Clientes (CAC)**,** que refleja la inversión en marketing y ventas destinada a captar un cliente, como gastos en publicidad, promociones y gestión de leads. (Ali & Shaban, 2024) El CAC es determinante para evaluar que tan rentable es atraer nuevos clientes versus el CLV.

Contexto Específico del Comercio Electrónico de Retail de Moda: Dinámicas del Sector.

El comercio electrónico minorista, especialmente en el sector de la moda, enfrenta desafíos particulares en su dinámica que lo hacen muy propenso al abandono de clientes. Esto se debe a que la decisión de compra en moda es profundamente personal y subjetiva, ligada a la percepción, la imagen, el ajuste y la talla esperada, y no tanto a una funcionalidad objetiva. Para poder desarrollar un modelo de aprendizaje predictivo, es necesario identificar las

variables que impactan negativamente la intención de compra, un aspecto clave para la retención. En este sentido, el principal obstáculo es el Riesgo Percibido por el cliente. (Margalina, 2021) explica que una desconexión entre lo que la plataforma promete y lo que el cliente recibe alimenta este riesgo, siendo la razón fundamental de que no vuelvan a comprar y terminen abandonando la marca. Por lo tanto, el modelo predictivo de aprendizaje automático debe medir este riesgo de forma práctica (operacional), usando indicadores como las tasas de devolución, las consultas al servicio de soporte o los tiempos de entrega. Esto permite clasificar y priorizar a los clientes más insatisfechos para poder ejecutar estrategias de retención de manera proactiva.

Bases de la Ciencia de Datos y Analítica Predictiva

Para el desarrollo del presente trabajo se optará por el estándar CRISP-DM(Cross-Industry Standard Process for Data Mining) sin embargo lo propuesto por Martinez y otros (Martinez-Plumed et al., 2021), se preferirá un enfoque más flexible con la finalidad de permitir iterar entre la exploración de los datos y el modelado predictivo ya que no solo se busca predecir el retiro (churn) sino también descubrir factores latentes; dicho estándar es ampliamente empleado por la industria y para el desarrollo de investigaciones.

La principal fortaleza radica en que el modelo CRIPSP-DM organiza el proceso de análisis de datos en fases bien definidas, con la finalidad de asegurar la alineación entre los objetivos del negocio y las soluciones basadas en datos. El modelo se compone de seis etapas interrelacionadas, de naturaleza iterativa las cuales son:

Etapa 1 Comprensión del negocio: La primera fase consiste en definir de manera inequívoca la problemática desde la perspectiva del negocio con la finalidad de identificar los objetivos estratégicos, las necesidades que se desean resolver y los criterios de éxito del proyecto.

Etapa 2 Compresión de los datos: Una vez definidos los objetivos en la etapa anterior, se procede a la exploración inicial de los datos obtenidos o disponibles. Esta fase se constituye de la recopilación de los datos, la descripción de sus características principales y la identificación de posibles anomalías en los mismos como pueden ser valores faltantes, inconsistencias desbalance, valores atípicos entre clases.

Etapa 3 Preparación de los datos: Esta fase es considerada una de las más críticas del proceso. En ella se realizan tareas como la limpieza de datos, la selección de variables elevantes, la transformación de atributos y la construcción de nuevas variables mediante técnicas de ingeniería de características.

Etapa 4 Modelado: En esta fase se seleccionan y aplican técnicas estadísticas o algoritmos de aprendizaje automático los cuales permitirán dar respuesta al problema planteado. Es por ello que en esta fase subyace de manera implícita la elección del modelo, el ajuste de sus parámetros y la evaluación preliminar de su desempeño.

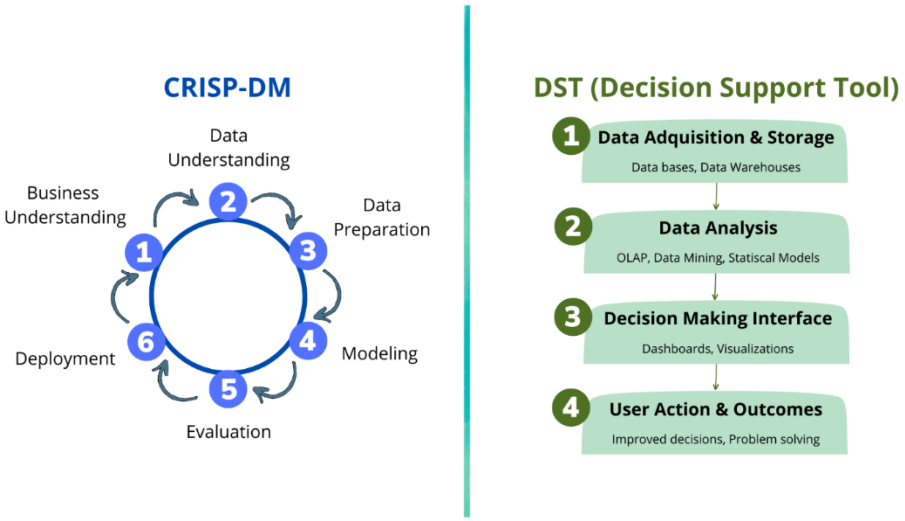
Etapa 5 Evaluación: El fin de la presente etapa es la determinación de si el modelo seleccionado e implementado cumple con los objetivos definidos en la fase inicial, mediante el análisis de los resultados obtenidos a partir de métricas de desempeño, así como su utilidad práctica y su consistencia teórica.

Etapa 6 Implementación: La última fase corresponde a la puesta en producción del modelo, ya sea mediante su implementación en una infraestructura productiva o a través de la generación de recomendaciones y estrategias basadas en los resultados obtenidos.

Las anteriores etapas se pueden describir en figura 1 a continuación:

Figura 1

Comparación Metodología CRIP-DM vs DST



Nota: Autoría propia

A diferencia de la minería de datos tradicional guiada estrictamente por objetivos la ciencia de datos moderna incorpora fases exploratorias iterativas. Dado que este proyecto busca no solo predecir el churn, sino también descubrir factores latentes, se adopta este enfoque flexible que permite iterar entre la exploración del valor de los datos y el modelado predictivo, adaptándose a la incertidumbre inherente al comportamiento del consumidor.

Evolución de la Analítica.

La analítica de datos en el contexto del mercadeo (marketing) ha venido experimentando una constante evolución, pasando de una función meramente descriptiva (que explica lo que ya pasó) a ser predictiva y prescriptiva logrando anticipar y recomendar acciones a tomar. De hecho, hay autores que proponen que el campo de la analítica de marketing se dedica a usar datos —incluyendo Big Data y Machine Learning— para incidir directamente en el desempeño empresarial, la segmentación de clientes y la lealtad (Petrescu & Krishen, 2023). Es precisamente esta capacidad predictiva la que le da valor a la ciencia de datos, ya que permite anticipar problemas críticos como la pérdida de clientes. Así en lugar de analizar el abandono de

clientes cuando ya ocurrió, se empieza a prevenir activamente el abandono, por lo cual las empresas construyen una ventaja competitiva real, estratégica y duradera en el mundo digital.

Aprendizaje Automático Supervisado para Clasificación Binaria: Fundamentos y Aplicación al Problema de Churn.

Es ampliamente expuesto por la literatura que el objetivo del Aprendizaje Automático (ML), consiste en desarrollar sistemas que pueden aprender automáticamente a partir de un conjunto de datos, sin necesidad de recibir instrucciones programadas paso a paso; cuyo objetivo principal consiste en el desarrollo y el análisis de algoritmos idóneos para el aprendizaje a partir de datos históricos, mediante la generación de predicciones con datos recién obtenidos y posteriormente evaluar la efectividad de dicho proceso de aprendizaje (Velasco Rebolledo, 2024). Según el tipo de datos de entrada, es posible clasificar las tareas de aprendizaje en supervisado y no supervisado.

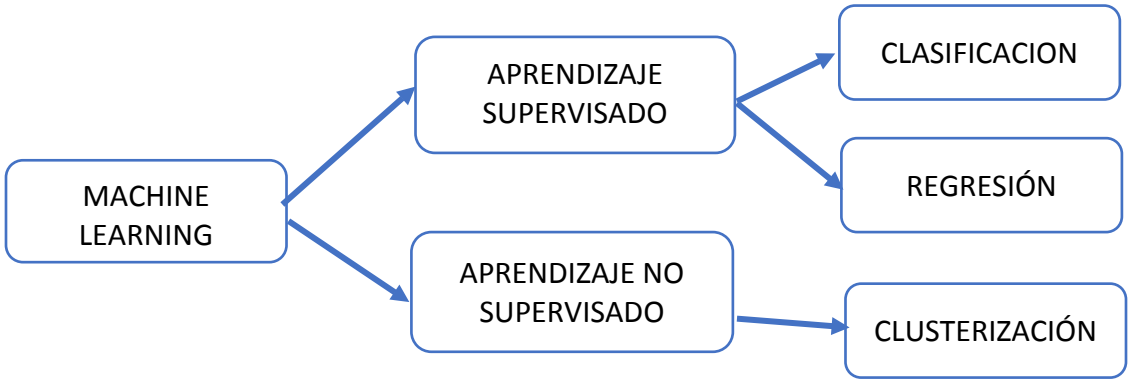
En el aprendizaje supervisado, se busca entrenar a un determinado modelo para hallar patrones en datos etiquetados y así poder definir reglas que asocien entradas con salidas específicas. Un caso de uso sería una plataforma de encuestas que categoriza preferencias de los diferentes usuarios mediante el empleo de algoritmos como la regresión lineal o clasificación.

La clasificación consiste en asignar etiquetas predefinidas a nuevos datos recopilados (como textos o imágenes) basándose en sus características. Para alcanzar dicho objetivo, el modelo se entrena con un subconjunto de datos ya clasificados y luego se valida con otro conjunto cuyas etiquetas son desconocidas, con el fin de maximizar su precisión predictiva mediante el empleo de análisis estadísticos, optimizando dicho modelo para minimizar el error de la predicción.

En la figura 2 se detalla los tipos de aprendizaje automático y la clasificación de técnicas por cada uno.

Figura 2

Aprendizaje Supervisado No Supervisado



Nota: Tomado de Machine Learning, Velasco Rebolledo, Jacinto. 2024.

En la siguiente tabla se detalla los tipos de algoritmo de ML y las categorías asociadas a cada tipo.

Tabla 1

Algoritmos y categoría asociada

Algoritmo	Categoría
K-Means	Clustering
Gaussian Mixtures	Clustering
Ordinary Least Squares (OLS)	Regresión, Selección de Características
Naive Bayes (NB)	Clasificación
K-Nearest Neighbors (k-NN)	Clasificación, Regresión
Support vector machines (SVM)	Clasificación, Regresión
Decision Trees (DT)	Clasificación, Regresión
Random Forest (RF)	Clasificación, Regresión
Recurrent Neural Networks (RNN)	Clasificación, Regresión

Nota: Tomado de Machine Learning, Velasco Rebolledo, Jacinto. 2024.

Las técnicas de clasificación basadas en aprendizaje supervisado, como las Máquinas de Vectores de Soporte (SVM) y los K-Vecinos Más Cercanos (KNN), ofrecen una solución con alto desempeño en casos de uso como comercio electrónico, debido a que permiten representar las relaciones complejas no lineales de los datos que un modelo estadístico tradicional no podría representar.

Algoritmos y Técnicas de Modelado Predictivo

Modelos Clásicos y de Ensamble: Regresión Logística, Árboles de Decisión, Random Forest, XGBoost.

La regresión logística se clasifica como una técnica estadística tradicional utilizada en analítica predictiva. En el contexto de la investigación (Kasemrat et al., 2025), proporciona información sobre cómo influye varias características en la probabilidad de que un cliente realice una compra, este modelo sirvió como base para la clasificación binaria, que busca determinar si un cliente realizará o no una compra, es una técnica valiosa por su interpretabilidad y facilidad de uso.

Los árboles de decisión consisten en un modelo de aprendizaje supervisado el cual funciona como un mapa lógico de reglas sencillas que permitirían realizar una predicción sobre una variable especifica. Dicho procedimiento se ejecuta de forma recursiva, dividiendo la información mediante umbrales óptimos alcanzando un criterio de parada definido. Dependiendo del tipo de variable se puede clasificar en dos tipos: Si los tipos de datos son discretos (es decir discontinuos), podemos denominar que se trata de un árbol de clasificación, pero si la variable es continua, se trataría de un árbol de regresión.

El algoritmo Random Forest es un método que se fundamenta en la creación de múltiples árboles de decisión independientes durante la etapa de entrenamiento del modelo, cada uno generado a partir de una muestra aleatoria del conjunto de datos original (Velasco

Rebolledo, 2024). Cada árbol se pondera en función de tasa de error, para dar un mayor peso en la estimación final, esta técnica puede ser compleja debido a la combinación de múltiples arboles lo que toma mayor tiempo de entrenamiento y predicción.

El Extreme Gradient Boosting (XGBoost) es una técnica de aprendizaje supervisado basada en el principio de gradient boosting, esta técnica sobresale en la optimización de la velocidad de entrenamiento y precisión del modelo. Se clasifica como un método de conjunto (ensemble) que permite realizar la selección de características y aplicar técnicas de regularización para prevenir sobreajuste y mejorar la capacidad de generalización del modelo a nuevos datos.

Deep Learning: Redes Neuronales

El aprendizaje profundo es un subcampo de la IA que utiliza diferentes estructuras de redes neuronales que permiten aprendizaje más significativo sobre los datos.

Las redes neuronales es una técnica programada para imitar el funcionamiento de las redes neuronales biológicas, (Aguado López, 2020), se caracteriza por la capacidad de aprendizaje en patrones complejos, una red neuronal consta de una capa de entrada que recibe datos, múltiples capas ocultas que aprenden características mediante pesos y sesgos y una capa de salida que genera la predicción final, los pesos y sesgos se ajustan en el entrenamiento para minimizar la perdida de entre las predicciones (Velasco Rebolledo, 2024). Las aplicaciones de esta técnica van desde reconocimiento de voz, clasificación de imágenes, predicción en el sistema financiero y comercio electrónico entre otros.

Optimización de Hiperparámetros.

La elección del algoritmo y su configuración (parámetros) son determinantes para el rendimiento predictivo. Si bien modelos clásicos como la regresión logística ofrecen

interpretabilidad, estudios recientes en retail demuestran que algoritmos de ensamblaje como XGBoost capturan mejor las relaciones no linealidades y las interacciones complejas entre variables de comportamiento o conductas.

La optimización de parámetros está dada para para encontrar la combinación más eficaz de parámetros para un modelo, la selección robusta de hiperparámetros permite evitar el sobreajuste (overfitting) y maximizar la capacidad de generalización del modelo, en lugar de utilizar métodos como la búsqueda en cuadrícula (Grid Search), que puede llegar a ser computacionalmente costoso e ineficiente, (Snoek et al., 2012), se fundamenta el uso de la Optimización Bayesiana, este método modela la función de rendimiento del algoritmo como un proceso Gaussiano, construye un modelo probabilístico que utiliza toda la información de evaluaciones anteriores para decidir donde evaluar la función y no solo confiar en las aproximaciones locales, esto logra superar las limitaciones de las búsquedas por cuadricula, permitiendo encontrar la configuración óptima de hiperparámetros con un número menor de evaluaciones, este método permite optimizar el uso de tiempo y recursos.

Preparación y Transformación de Datos (Feature Engineering)

Calidad, Tratamiento y Normalización de Datos.

El análisis exploratorio de datos (EDA) es una etapa indispensable en la ciencia de datos, pues permite comprender la información básica antes de emplear técnicas de modelado de datos más complejas. Su objetivo es examinar la naturaleza y estructura de los datos para no afectar la calidad de los análisis posteriores y mejorar la precisión de los modelos, (Velasco Rebolledo, 2024). El proceso inicia con un análisis de la estructura del conjunto de datos, identificando las variables, categoría de los datos y la posible presencia de valores faltantes o nulos. Posteriormente, se calcula estadísticas descriptivas (como la media, mediana, desviación estándar, valores máximos y mínimos) para obtener una visión inicial de distribución y variabilidad.

La visualización de la distribución de características es un componente importante del EDA. A través de histogramas, diagramas de caja y gráficos de densidad se puede observar el comportamiento de las variables numéricas, detectando patrones, tendencias o anomalías (Velasco Rebolledo, 2024), para variables categóricas, se analizan las frecuencias de cada categoría mediante gráficos de barras o circulares.

Un aspecto fundamental es el manejo de valores atípicos y datos faltantes, ya que pueden afectar la calidad del análisis. Según el caso, se decide eliminarlos, imputarlos o aplicar otras estrategias para reducir su impacto, garantizar integridad y resultados consistentes.

Transformación de Datos (Feature Engineering).

La capacidad predictiva del modelo depende intrínsecamente de la calidad de las variables de entrada (feature engineering).

Modelo RFM y Clustering: La literatura sugiere transformar los datos transaccionales crudos utilizando el modelo RFM (Recencia, Frecuencia, Valor Monetario). Haga clic o pulse aquí para escribir texto. demuestran que la segmentación de clientes basada en RFM, potenciada por algoritmos de agrupamiento basados en densidad como DBSCAN, permite identificar patrones de comportamiento y aislar valores atípicos (ruido) de manera más efectiva que los métodos de partición tradicionales como K-Means.

Variables Espaciales y Geodemográficas: Una innovación reciente en la predicción de churn es la inclusión del contexto geográfico. (Matuszelański & Kopczewska, 2022) validan que enriquecer los datos transaccionales con información censal y ubicación espacial mejora la precisión del modelo, permitiendo diferenciar el riesgo de abandono entre clientes de zonas urbanas densas y zonas rurales.

Manejo de Big Data en Entornos Distribuidos

El éxito de cualquier modelo predictivo en el comercio electrónico depende de un manejo adecuado del Big Data en entornos distribuidos. Esto es indispensable dada la velocidad, el gran volumen y la variedad de los datos transaccionales. En este sentido, sostienen que el paso más importante para anticipar el abandono no reside en la elección del algoritmo, sino en la fase de Preparación y Transformación de Datos, conocida como Feature Engineering. La función de esta etapa es tomar los datos brutos sobre el comportamiento del cliente y convertirlos en características realmente informativas. Al hacer esto, el modelo puede aprender de manera más efectiva, lograr predicciones precisas, y transformar la analítica de Big Data en una fuente concreta de ventaja competitiva.

Evaluación y Validación de Modelos

Métricas para Problemas Desbalanceados.

Para la evaluación y validación de los modelos de predicción de abandono, es esencial ser consciente de las limitaciones del Accuracy (precisión general). En la práctica, los datos de abandono suelen estar desbalanceados, pues los clientes que realmente abandonan son una minoría, y en esta situación el Accuracy no es suficiente. Por ello, el trabajo de (De et al., 2023) subraya la importancia de dar prioridad a métricas más robustas, como el Recall, el F1-Score y el ROC-AUC. Estas métricas son las que garantizan que el modelo no solo sea preciso en general, sino que sea realmente eficiente identificando a la clase minoritaria: los clientes que están en riesgo de irse. Este es, en definitiva, el objetivo estratégico de cualquier proyecto de retención.

Técnicas de Validación: Hold-out, Cross-Validation.

Como técnica para evaluación del modelo se utiliza la división de datos, por lo general se utiliza 80, 20, o 70, 30, es decir 80 % de los datos se utilizan para entrenamiento, es decir para que el modelo aprenda y 20 % para evaluar el modelo, esta técnica se conoce como Hold-

out. La evaluación depende de una única partición lo que puede generar sesgo si la división no es representativa, Por ello al trabajar con conjunto de datos más pequeños se aplica la técnica conocida como validación cruzada (Cross-validation). Este método consiste en dividir el dataset original en múltiples subconjuntos, que se alternan como datos de entrenamiento, prueba y validación en las distintas fases del proceso (Velasco Rebolledo, 2024). La elección del algoritmo predictor debe adaptarse a las características específicas de los datos, buscando un equilibrio entre la precisión y atributos como su interpretabilidad, simplicidad, eficiencia en el entrenamiento y capacidad para operar en tiempo real. Una vez divididos los datos en conjuntos de entrenamiento y prueba, se aplica y optimiza la función de predicción mediante validación cruzada, refinándola hasta alcanzar un margen de error satisfactorio. El modelo definitivo puede consistir en un único predictor o integrar una combinación de varios, previamente evaluados.

Desbalance de Clases y Rentabilidad.

El dataset típico de comercio electrónico presenta un severo desbalance de clases, donde la clase minoritaria (clientes que recompran/no abandonan) es la de mayor interés.

Manejo del Desbalance (SMOTE): Para mitigar el sesgo hacia la clase mayoritaria, se aplica la técnica SMOTE (Synthetic Minority Over-sampling Technique) propuesta por Haga clic o pulse aquí para escribir texto. A diferencia del sobremuestreo simple, SMOTE genera instancias sintéticas interpolando entre vecinos cercanos de la clase minoritaria en el espacio de características, lo que expande las regiones de decisión y mejora la generalización del clasificador.

Métricas Orientadas al Beneficio (EMPC): Tradicionalmente, los modelos se evalúan mediante AUC o Accuracy. Sin embargo, (Höppner et al., 2017) argumentan que estas métricas asumen erróneamente que todos los errores de clasificación tienen el mismo costo. En un contexto de retención, un Falso Negativo (no detectar a un cliente que se va) conlleva la

pérdida del valor de vida del cliente (CLV), mientras que un Falso Positivo implica un costo de incentivo innecesario. Por tanto, se propone el uso de la EMPC (Expected Maximum Profit Measure for Customer Churn) para seleccionar el modelo que maximice la rentabilidad esperada de la campaña, alineando la técnica con los objetivos de negocio.

Interpretabilidad Y Acción Estratégica (XAI)

Inteligencia Artificial Explicable (XAI).

Finalmente, para que el modelo sea una herramienta de soporte a la decisión, debe ser transparente. (Molnar, 2025, pp. 15–33) define la interpretabilidad como el grado en que un humano puede entender la causa de una decisión. Dado el uso de modelos de "caja negra" como XGBoost, es necesario implementar técnicas de Inteligencia Artificial Explicable (XAI). Herramientas como los gráficos de Dependencia Parcial (PDP) permiten visualizar el efecto marginal de una característica (ej. precio, tiempo de entrega) sobre la probabilidad de abandono, facilitando el diseño de estrategias de retención basadas en evidencia causal y no solo correlacional.

Técnicas de Interpretación: SHAP – LIME.

LIME (Local Interpretable Model-Agnostic Explanations) es un enfoque novedoso que funciona como un algoritmo de Inteligencia Artificial Explicable (XAI).

Estos métodos son universales, lo que significa que funcionan para explicar cualquier modelo de Inteligencia Artificial, sin importar qué tan complejo o moderno sea. LIME actúa como una lupa que analiza una decisión específica, (Hassan et al., 2025) se enfoca en un caso individual para identificar exactamente qué datos influyeron más en ese resultado puntual, traduciendo procesos complejos a un lenguaje fácil de entender.

Por otro lado, SHAP ofrece una interpretación más profunda y matemáticamente exacta. Basado en la teoría de juegos, este método calcula cuánto contribuyó cada pieza de información al resultado final. Mientras LIME nos da una explicación rápida y local, SHAP proporciona una base sólida y confiable. Su funcionamiento consiste en asignar un valor de importancia a cada característica individual, según su contribución a una predicción específica. Esto permite clarificar el peso de cada variable e identificar los factores clave que más influyen en el resultado del modelo.

Vinculación con Segmentación Proactiva de Clientes en Riesgo.

La última etapa de la analítica predictiva es la interpretabilidad y acción estratégica (XAI), cuya función es que los modelos de Machine Learning no se limiten a predecir el abandono con alta precisión, sino que también puedan explicar por qué lo hacen. (Martínez & Segarra Marlon, 2024) enfatizan que el valor real para la empresa no está solo en la predicción, sino en la capacidad del modelo para determinar la Importancia de variables y así lograr caracterizar a los clientes que están en riesgo. Este conocimiento es indispensable para una retención exitosa, ya que permite realizar una segmentación proactiva y diseñar intervenciones específicas. De esta forma, la predicción pasa de ser una herramienta técnica para convertirse en una palanca fundamental de la estrategia de negocio.

Marco Conceptual y Operativo

Contexto del Ecommerce en Latinoamérica.

El comercio electrónico en América Latina está en plena expansión, impulsado por la creciente conectividad y los hábitos de consumo que cambiaron tras la pandemia. Se espera que este dinamismo continúe a medio plazo.

En crecimiento general, se proyecta que las ventas minoristas totales crezcan a una tasa anual compuesta (TCAC) del 6 % entre 2024 y 2028, mientras que el comercio en línea lo hará a un ritmo mucho más acelerado (11 % TCAC)(HKTDC RESEARCH, 2025). Como resultado, se estima que la participación de las ventas en línea aumente del 12.3 % en 2023 al 15.9 % en 2028 (HKTDC RESEARCH, 2025).

El liderazgo regional, lo concentran Brasil y México con la mayor parte de la actividad económica y del mercado digital. En conjunto, representaron alrededor de dos tercios de las ventas de comercio electrónico regional en 2023. Los seis principales mercados (Brasil, México, Argentina, Chile, Colombia y Perú) sumaron más de 110 mil millones de dólares en ventas ese año (HKTDC RESEARCH, 2025).

La plataforma dominante es Mercado Libre es el marketplace líder con una cuota del 26 % en valor de ventas minoristas (2023) (HKTDC RESEARCH, 2025). Otras plataformas importantes son Amazon (5 %) y Magazine Luiza (3 %). Se observa la creciente presencia de plataformas asiáticas como AliExpress y Shopee (2 % cada una), atrayendo a consumidores con productos de buena relación calidad-precio.

Categorías de productos, en 2023, los electrodomésticos y la electrónica fueron los productos más vendidos en línea (22 % del valor), seguidos por la moda (14 %) y los alimentos (8 %). Sin embargo, las proyecciones de crecimiento más altas para 2024-2028 están en categorías de compra frecuente, como salud y belleza (TCAC 12 %), alimentos (TCAC 11 %) y otros productos (TCAC 13 %) (HKTDC RESEARCH, 2025).

Entorno digital y regulaciones, **l**a región tiene una población joven y altamente conectada. El marketing digital es fundamental: el 93 % de las empresas utiliza redes sociales, siendo Facebook e Instagram las más populares (82 % de uso cada una) (HKTDC RESEARCH, 2025). En cuanto a regulaciones, los umbrales de importación sin impuestos

(valor de *minimis*) varían significativamente entre países, desde los 30 dólares en Chile hasta los 200 dólares en Perú.

Definiciones Operativas Clave.

Abandono de Cliente (Churn No Contractual): Fenómeno de deserción en entornos minoristas donde no existe un contrato formal. Se define operativamente según (Fader & Hardie, 2007) como un cese de actividad "silencioso", clasificando como churner (clase 1) al cliente que no realiza recompras en una ventana de tiempo determinada (ej. 6 meses).

Desbalance de Clases (Class Imbalance): Disparidad significativa en la distribución de la variable objetivo, común en retail donde la tasa de abandono o recompra es minoritaria. Operativamente, esta condición justifica el uso de técnicas de muestreo (SMOTE) y métricas de rentabilidad (EMPC) en lugar del Accuracy tradicional.

Ingeniería de Características (Feature Engineering): Proceso de transformación de datos crudos en variables predictivas. En esta investigación, implica convertir logs transaccionales en métricas de comportamiento RFM (Recencia, Frecuencia, Monto) y variables espaciales derivadas mediante clustering, fundamentales para capturar patrones de consumo.

Modelo RFM: Técnica de segmentación que evalúa la lealtad del cliente basándose en qué tan recientemente compró (R), con qué frecuencia (F) y cuánto gastó (M). Estas variables calculadas serán los inputs principales del modelo predictivo debido a su alta correlación con la retención futura.

XGBoost (Extreme Gradient Boosting): Algoritmo de ensamblaje de árboles de decisión optimizado para velocidad y rendimiento. Se selecciona como modelo central por su capacidad para manejar datos tabulares complejos y relaciones no lineales, optimizando sus hiperparámetros para maximizar la detección de abandono.

Optimización Bayesiana: Método de ajuste de hiperparámetros que utiliza modelos probabilísticos para encontrar la configuración óptima del algoritmo de forma más eficiente que la búsqueda exhaustiva (Grid Search). Se utilizará para ajustar la tasa de aprendizaje y profundidad de los árboles del modelo XGBoost.

SMOTE (Synthetic Minority Over-sampling Technique): Técnica para corregir el desbalance de clases generando datos sintéticos de la clase minoritaria (abandono) mediante interpolación de vecinos cercanos, evitando el sesgo del modelo hacia la clase mayoritaria.

EMPC (Expected Maximum Profit Measure for Customer Churn): Métrica de evaluación orientada al negocio que supera a las métricas estadísticas tradicionales al integrar la matriz de costos y beneficios. Su objetivo es maximizar la rentabilidad esperada, ponderando económicamente los errores de clasificación según el valor del cliente.

Valor de Vida del Cliente (CLV): Estimación del beneficio neto futuro atribuible a un cliente. Operativamente, actúa como el "costo de oportunidad" en la función de costos del EMPC; un falso negativo implica la pérdida total de este valor.

Falso Positivo (Error Tipo I): Error donde se clasifica un cliente leal como en riesgo de abandono. Operativamente, representa un desperdicio de recursos de marketing (costo de contacto e incentivos) en usuarios que no los necesitaban.

Falso Negativo (Error Tipo II): Error crítico donde el modelo predice que un cliente permanecerá, cuando en realidad abandona. Según (Höppner et al., 2017) es el error más costoso pues impide ejecutar acciones de retención, resultando en la pérdida del cliente y su CLV asociado.

Valores SHAP (Shapley Additive explanations): Método basado en teoría de juegos para interpretar modelos de "caja negra". Se utilizará para explicar individualmente por qué un cliente tiene alto riesgo de abandono, facilitando la creación de estrategias de retención personalizadas.

DBSCAN: Algoritmo de agrupamiento basado en densidad utilizado en la fase de preprocesamiento para clasificar la ubicación geográfica de los clientes, permitiendo derivar variables demográficas sin depender de códigos postales rígidos.

Hiperparámetro: Configuración externa del modelo (ej. profundidad del árbol) que no se aprende de los datos, sino que debe fijarse previamente. Su correcta optimización es crucial para equilibrar la complejidad del modelo y su capacidad de generalización.

De la Predicción a la Retención: Estrategias Basadas en Datos

La etapa final y más relevante del proceso es lograr la transición efectiva desde la simple predicción del abandono hacia la acción estratégica. (Rojas, 2024) explica que el valor de un modelo de Machine Learning va más allá de su precisión; su verdadero aporte radica en su capacidad para segmentar a los clientes según su perfil de riesgo. En la práctica, esto significa que el modelo es capaz de responder a tres preguntas esenciales para la empresa: 1) ¿A quién debe enviarse una oferta? 2) ¿Qué tipo de oferta (personalizada) debe enviarse? y 3) ¿Cuál es el momento oportuno para intervenir? Al aplicar esta estrategia dirigida, se asegura que los recursos de retención (como descuentos o correos win-back) se inviertan únicamente en los clientes correctos y en el momento preciso, optimizando así la tasa de fidelización.

Marco Normativo y Consideraciones Éticas

El desarrollo del presente proyecto de investigación aplicada, enfocado en la predicción de abandono de clientes (churn) mediante técnicas de Aprendizaje Automático, se rige bajo el ordenamiento jurídico colombiano y los lineamientos éticos internacionales para el tratamiento de datos masivos (Big Data) e Inteligencia Artificial. A continuación, se detallan los instrumentos legales que fundamentan la viabilidad y legalidad del estudio.

Protección de Datos Personales y Habeas Data.

Dado que la investigación procesa registros transaccionales de personas naturales, el marco principal es el derecho constitucional al Habeas Data (Art. 15 de la Constitución Política de Colombia) y su reglamentación mediante la (LEY ESTATUTARIA 1581 DE 2012, 2012).

Tratamiento de Datos y Anonimización.

En cumplimiento del principio de seguridad y confidencialidad consagrado en la Ley 1581 y el Decreto 1377 de 2013, los datos suministrados por las empresas del sector retail han sido sometidos a un protocolo estricto de anonimización irreversible previo a su ingreso en los modelos de entrenamiento.

Gestión de Variables Geoespaciales.

De conformidad con los conceptos de la Superintendencia de Industria y Comercio (SIC) sobre la minimización de datos, la variable de código postal no se utiliza como un identificador de ubicación domiciliaria específica. En su lugar, se emplea exclusivamente como un insumo estadístico para la derivación de variables agregadas (zona logística, densidad urbana/rural) y para la aplicación de algoritmos de agrupamiento (Clustering), garantizando que no sea posible la re-identificación o geolocalización precisa de los titulares de la información.

Propiedad Industrial y Secretos Empresariales.

Considerando que los datos provienen de fuentes privadas corporativas, el proyecto se adhiere al (Régimen Común Sobre Propiedad Industrial, 2000), específicamente en lo referente a la protección de Secretos Empresariales (Artículo 260).

Confidencialidad de la Fuente.

Para proteger la ventaja competitiva y las métricas internas de las marcas colaboradoras, se mantiene la reserva total sobre la identidad de las empresas proveedoras de los datos. Los resultados del modelo se presentan de manera agnóstica a la marca,

enfocándose en los patrones de comportamiento del consumidor y no en el desempeño comercial específico de una organización identificable.

Ética en Inteligencia Artificial (Explicabilidad).

El diseño metodológico del proyecto adopta los principios del Marco Ético para la Inteligencia Artificial en Colombia (MINTIC, 2021), priorizando la transparencia y la explicabilidad algorítmica.

Metodología

Enfoque

La presente investigación adopta un enfoque cuantitativo, con el propósito de analizar de manera objetiva y medible los factores asociados al abandono de clientes en el comercio electrónico retail-moda. Este enfoque permite convertir en indicadores cuantificables, las variables de estudio y evaluar su relación con el fenómeno del abandono de clientes en el sector de comercio electrónico minorista de moda mediante técnicas estadísticas y de analítica avanzada.

Diseño

Desde el punto de vista del diseño metodológico, la investigación es no experimental, dado que no se implementa y/o emplea una infraestructura o plataforma para la generación u obtención de los datos, y de corte transversal, ya que el análisis se realiza sobre datos recolectados en un único periodo temporal desde el 01 abril de 2024 al 31 de diciembre de 2025.

Así mismo se describen las etapas definidas para el procesamiento de la información de los datos en bruto obtenido; estos se someten a un proceso de anonimización con la finalidad de no comprometer la confidencialidad de las diferentes transacciones realizadas, luego se

realiza la carga de los mismos a una base de datos relacional y mediante la ejecución consultas de SQL se generarán las cifras necesarias para la generación del set de datos que alimentarán los algoritmos a emplear, luego se calcula la matriz de correlación con la finalidad de definir aquellas variables que alimentarán el entrenamiento de los algoritmos a emplear.

Alcance

El estudio presenta un alcance exploratorio, descriptivo y correlacional. En su fase exploratoria, la investigación permite identificar las variables relevantes asociadas al abandono de clientes dentro del contexto del comercio electrónico retail-moda a partir de los datos en bruto obtenidos de la operación de mencionado tipo de empresa. Desde el punto de vista descriptivo, se estudian las variables identificadas a partir de los datos disponibles con la finalidad de definir aquellas que se emplearán para el entrenamiento de los algoritmos de aprendizaje automático a usar. Finalmente, el alcance correlacional se orienta a examinar la relación existente entre las variables identificadas y el abandono de clientes mediante el empleo de métricas estadísticas (matriz de correlación). Este planteamiento se alinea con enfoques metodológicos aplicados en estudios de fidelización, como el realizado por (Rojas, 2024) apoyados en analítica cuantitativa y machine learning.

Descripción y Selección de Variables

La preparación de la base de datos se centró en convertir el historial operativo en una visión clara del comportamiento de cada cliente. El proceso tuvo como enfoque principal la unificación de la información de ventas y logística, transformando los registros brutos en indicadores de negocio fáciles de interpretar, como la frecuencia de compra, el gasto total y la calidad de la experiencia de entrega. Todo el proceso incluyó una limpieza exhaustiva para corregir errores en los datos y un estricto protocolo de seguridad para proteger la identidad de los usuarios, garantizando así información confiable y lista para el análisis.

Las variables y su respectiva descripción que se usarán para entrenar los modelos de machine Learning son los siguientes:

Tabla 2.

Variables y Descripción

Variable	Descripción conceptual	Descripción Operacional	Tipo de dato
brand	Identificador de la marca.	Mapeo del account_name de la base de datos transaccional a etiquetas anonimizadas.	Cualitativo Nominal (Categórico)
user_id	Identificador único del cliente.	Hash SHA256 del tipo y número de documento del cliente, convertido a hexadecimal.	Cualitativo Nominal (Identificador)
customer_tenure	Antigüedad del cliente en días (días transcurridos desde su primera compra hasta la fecha de corte).	Diferencia en días entre la fecha de corte ('2025-12-31') y la fecha de la primera compra histórica (MIN(creation_date)).	Cuantitativo Continuo (Intervalo)
recency	Recencia: Días transcurridos desde la última compra del cliente hasta la fecha de corte.	Diferencia en días entre la fecha de corte ('2025-12-31') y la fecha de la última compra registrada (MAX(creation_date)).	Cuantitativo Continuo (Intervalo)
frequency	Frecuencia: Cantidad total de órdenes distintas realizadas por el cliente.	Conteo total (SUM) de órdenes distintas (orders) realizadas en el periodo.	Cuantitativo Discreto
monetary_value	Valor Monetario: Suma total del GMV generado por el cliente.	Sumatoria total del GMV (Gross Merchandise Value) de todas las órdenes del cliente.	Cuantitativo Continuo (Ratio)
gmv_last_6m	Monto de ventas (GMV) acumulado en los últimos 6 meses.	Suma del GMV filtrando órdenes con fecha >= 6 meses antes de la fecha de corte ('2025-12-31').	Cuantitativo Continuo
gmv_previous_12m	Monto de ventas (GMV) acumulado en los 12 meses	Suma del GMV filtrando órdenes con fecha < 6 meses antes de la fecha de corte.	Cuantitativo Continuo

	anteriores al periodo reciente de 6 meses.		
avg_categories_in_order	Promedio de categorías únicas diferentes incluidas por orden (Variedad dentro de la cesta de compra).	Promedio (AVG) de la cantidad distinta de categorías (Category_Name) presentes en cada orden.	Cuantitativo Continuo
total_flag_discount	Cantidad total de órdenes que incluyeron algún descuento.	Sumatoria de flags donde la orden tuvo un items_discount distinto de 0.	Cuantitativo Discreto
total_shipping_cost	Monto total pagado por el cliente por concepto de envíos.	Sumatoria total del costo de envío pagado en todas las órdenes.	Cuantitativo Continuo
avg_shipping_cost	Costo promedio de envío pagado por orden.	Promedio (AVG) del costo de envío por orden.	Cuantitativo Continuo
total_flag_free_shipping	Cantidad de órdenes que tuvieron costo de envío cero.	Conteo de órdenes donde el costo de envío fue 0.	Cuantitativo Discreto
avg_lead_time	Tiempo promedio en días transcurrido entre la creación de la orden y el despacho real.	Promedio de días entre la creación de la orden y la fecha de despacho real (shipped_date_country) .	Cuantitativo Continuo
first_order_on_time	Indicador de primera experiencia: 1 si la primera orden histórica llegó a tiempo, 0 si se atrasó o no hay datos.	Flag (1/0) de la primera orden histórica (ORDER BY creation_date ASC LIMIT 1) indicando si se despachó a tiempo.	Cualitativo Nominal (Binario)
total_flag_shipped	Cantidad de órdenes que cumplieron con la promesa de entrega (Fecha de despacho <= Fecha estimada).	Sumatoria de órdenes que cumplieron la promesa (Fecha despacho <= Fecha estimada).	Cuantitativo Discreto
request_qty	Cantidad total de unidades solicitadas por el cliente en todas sus órdenes.	Suma total de la cantidad de ítems (qty) pedidos en todas las órdenes.	Cuantitativo Discreto
confirm_qty	Cantidad total de unidades confirmadas por el sistema en todas sus órdenes.	Suma total de la cantidad de ítems (confirm_qty) confirmados por el sistema.	Cuantitativo Discreto
first_order_in_full	Indicador de primera experiencia: 1 si la primera orden histórica	Flag (1/0) de la primera orden histórica indicando si se entregó	Cualitativo Nominal (Binario)

	llegó completa, 0 si tuvo quiebre o no hay datos.	completa (request_qty == confirm_qty).	
total_flag_qty	Cantidad de órdenes con cumplimiento perfecto de inventario.	Cantidad de órdenes donde lo solicitado fue igual a lo confirmado.	Cuantitativo Discreto
total_flag_delivered	Cantidad de órdenes que alcanzaron efectivamente el estado de despachado/entregado.	Sumatoria de órdenes que tienen fecha de creación en tabla de control de despacho (sbcs.created_at).	Cuantitativo Discreto
total_flag_return	Cantidad de órdenes que tienen asociado un proceso de devolución.	Conteo de órdenes que tienen un original_order_id asociado en la tabla de jerarquía de devoluciones.	Cuantitativo Discreto
avg_ticket	Ticket Promedio: Valor monetario promedio por compra realizada.	División segura (SAFE_DIVIDE) del monetary_value sobre la frequency.	Cuantitativo Continuo
delivery_delay_rate	Tasa de retraso: Porcentaje de órdenes que no cumplieron la fecha de entrega prometida.	Proporción de órdenes NO enviadas a tiempo (frequency - total_flag_shipped) sobre el total de órdenes.	Cuantitativo Continuo (Ratio 0-1)
return_rate	Tasa de devoluciones: Porcentaje de órdenes totales que resultaron en una devolución.	Proporción (SAFE_DIVIDE) de órdenes con devolución (total_flag_return) sobre el total de frecuencia.	Cuantitativo Continuo (Ratio 0-1)
discount_sensitivity	Sensibilidad al descuento: Porcentaje de compras realizadas bajo condiciones de descuento.	Proporción de órdenes compradas con descuento sobre el total de órdenes.	Cuantitativo Continuo (Ratio 0-1)
trend_score	Puntaje de tendencia: Relación de gasto reciente (6m) vs histórico (12m previos). >1 indica crecimiento, <1 indica decrecimiento.	Ratio entre gmv_last_6m y gmv_previous_12m. Si previo es 0 y actual > 0, se asigna 2.0 (boost).	Cuantitativo Continuo (Índice)
shipping_cost_ratio	Ratio de costo de envío: Relación entre el costo promedio de	Relación entre el costo promedio de envío y el valor monetario total del cliente.	Cuantitativo Continuo (Ratio)

	envío y el valor total de vida del cliente.		
is_one_timer	Flag de comprador único: 1 si el cliente ha realizado solo una compra, 0 si es recurrente.	Flag derivado: 1 si frequency es 1, de lo contrario 0.	Cualitativo Nominal (Binario)
churn	Target de fuga: 1 si la recencia es mayor a 90 días, 0 si ha comprado recientemente.	Flag derivado: 1 si recency > 90 días (relativo a la fecha de corte), de lo contrario 0.	Cualitativo Nominal (Binario)
Costo de Adquisición de Clientes (CAC)	Mide cuanto invierte una empresa para adquirir un nuevo cliente	CAC = Total de inversiones en Marketing y Ventas / Número de clientes nuevos adquiridos	Cuantitativo Continuo (Ratio)
Churn Rate – Tasa de Abandono	Tasa o porcentaje de clientes que dejar de utilizar o consumir un producto o servicio durante un periodo de tiempo	Churn Rate = Clientes perdidos durante x tiempo / Clientes al inicio del periodo) x 100	Cuantitativo Continuo (Ratio)
Customer Lifetime Value (CLV)	Valor de vida del cliente, estimación del beneficio neto futuro atribuible a un cliente, beneficio que un cliente genera durante su relación con la marca o comercio	El CLV se estima con la formula: CLV = (Valor promedio de compra x Frecuencia de Compra) x Duración promedio de vida del cliente – Costos de adquisición	Cuantitativo Continuo (Ratio)
Métricas del Modelo	Métricas que evalúan el Rendimiento del modelo: Exactitud (accuracy), Sensibilidad (recall), Precisión (precisión), F1-Score.	Exactitud: TN+TP / TN+TP+FN+FP, Sensibilidad: TP/ TP+FN, Precisión: TP/TP+FP, R1-Score: 2.Sensibilidad X Precisión / Sensibilidad+Precisión	Cuantitativo Continuo (Ratio)

Nota: Autoría propia

Selección de Métodos o Instrumentos para Recolección de Información

El desarrollo del modelo de aprendizaje automático se basa en la recopilación y análisis de datos históricos de clientes en comercio electrónico del sector retail - moda, se utiliza la base de datos anonimizada que contiene 91.921 registros de clientes, incluye información de

tipo transaccional e información logística, recaudada por la propia plataforma de ventas de las marcas y un sistema de administración de órdenes que almacena la operación logística de las mismas. La siguiente tabla relaciona los primeros registros de la base de datos.

Tabla 3.

Registros de la Base de Datos

	brand	customer_tenure	recency	frequency	monetary_value	gmv_last_6m	gmv_previous_12m
0	Brand_2	210	210	1	127440	127440	0
1	Brand_2	576	576	1	121140	0	121140
2	Brand_2	210	210	1	64990	64990	0
3	Brand_2	457	457	1	104355	0	104355
4	Brand_2	212	212	1	59990	59990	0

avg_categories_in_order	total_flag_discount	total_shipping_cost	...	total_flag_delivered	total_flag_return	avg_ticket
4.0	1	0	...	1.0	1.0	127440.0
1.0	1	0	...	1.0	1.0	121140.0
1.0	1	0	...	1.0	1.0	64990.0
3.0	1	0	...	1.0	1.0	104355.0
1.0	1	0	...	1.0	1.0	59990.0

delivery_delay_rate	return_rate	discount_sensitivity	trend_score	shipping_cost_ratio	is_one_timer	churn
0.0	1.0	1.0	2.0	0.0	1	1
0.0	1.0	1.0	0.0	0.0	1	1
0.0	1.0	1.0	2.0	0.0	1	1
0.0	1.0	1.0	0.0	0.0	1	1
0.0	1.0	1.0	2.0	0.0	1	1

Nota: Autoría propia

Para garantizar la privacidad absoluta de los clientes, se implementó un proceso de anonimización irreversible. Los documentos de identidad fueron sustituidos por códigos alfanuméricos únicos generados mediante un sistema de cifrado seguro y privado. Esto permite

analizar el historial y comportamiento de compra de cada usuario sin exponer jamás su identidad real ni sus datos personales.

Adicionalmente, se protegió la información estratégica de las empresas participantes reemplazando los nombres comerciales de las marcas por etiquetas genéricas. Esto asegura que el análisis se centre exclusivamente en los patrones operativos y de venta, eliminando sesgos y manteniendo la confidencialidad de cada negocio.

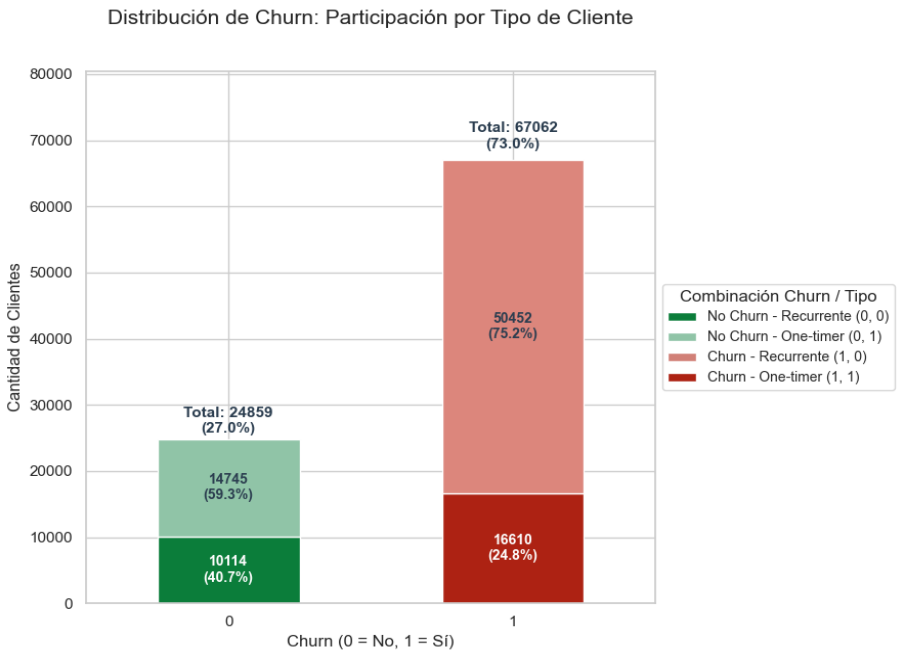
Técnicas para Análisis de Datos

Para analizar los datos se aplica el análisis exploratorio de datos (EDA) por sus siglas en inglés, esta etapa busca comprender los datos de forma sencilla, analizando la estructura y naturaleza de estos; se inicia examinando las primeras filas del dataset de datos, se identifica 91921 filas x 29 columnas, tipo de datos: números enteros (int64: 13 columnas), decimales (float64: 15 columnas) y texto (object: 1 columna) y 15981 valores faltantes para la variable trend_score, esto se debe a que la variable se calcula con el total de la venta por cliente desde 1 de abril hasta el 30 de septiembre de 2025, dividido por el total de la venta por cliente desde el 1 de abril de 2024 hasta el 31 de marzo de 2025, lo que da lugar a valores nulos cuando la venta del denominador es 0, como tratamiento a estos valores se agregó el valor -1 para que el modelo identifique que es un valor atípico y no lo asocie a ninguna venta, y 3 filas en variables numéricas con valores nulos en los cuales se reemplazó por la mediana de los datos. Luego se realizó una exploración utilizando estadísticas descriptivas, se identificó la media, desviación estándar, valores máximos y mínimos, cuartiles del 25 %, 50 %, y 75 % de los datos, detectando distribución y variabilidad.

Posteriormente para visualizar la distribución de churn: participación de todo tipo de cliente se utiliza un histograma que permite visualizar el tipo de churn si es recurrente o única vez. La siguiente gráfica ilustra la participación mencionada:

Figura 3

Distribución de Churn: Participación por Tipo de Cliente

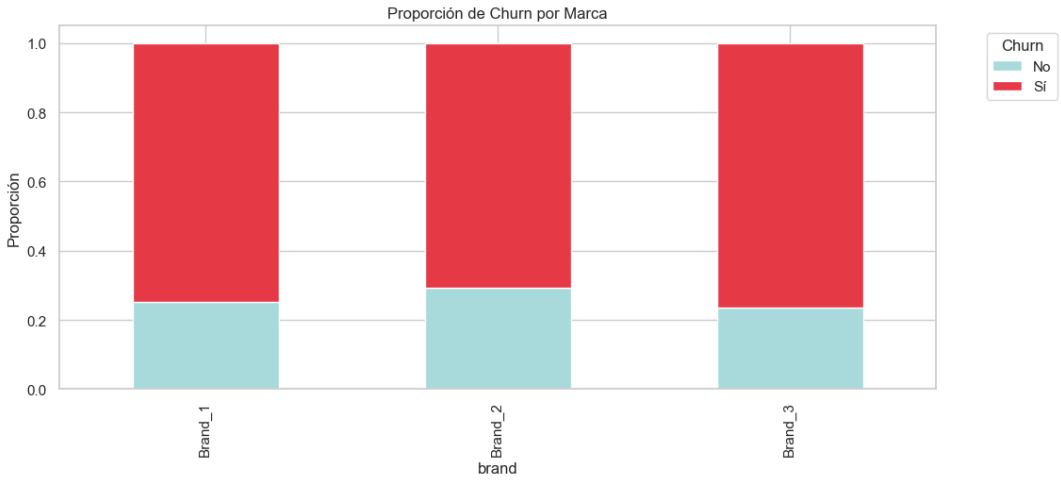


Nota: Autoría propia

Luego se analiza la proporción de churn en las 3 marcas estudiadas, la distribución de los datos identifica un patrón de comportamiento similar sin mayor variabilidad que representa la siguiente gráfica:

Figura 4

Proporción de Churn por Marca

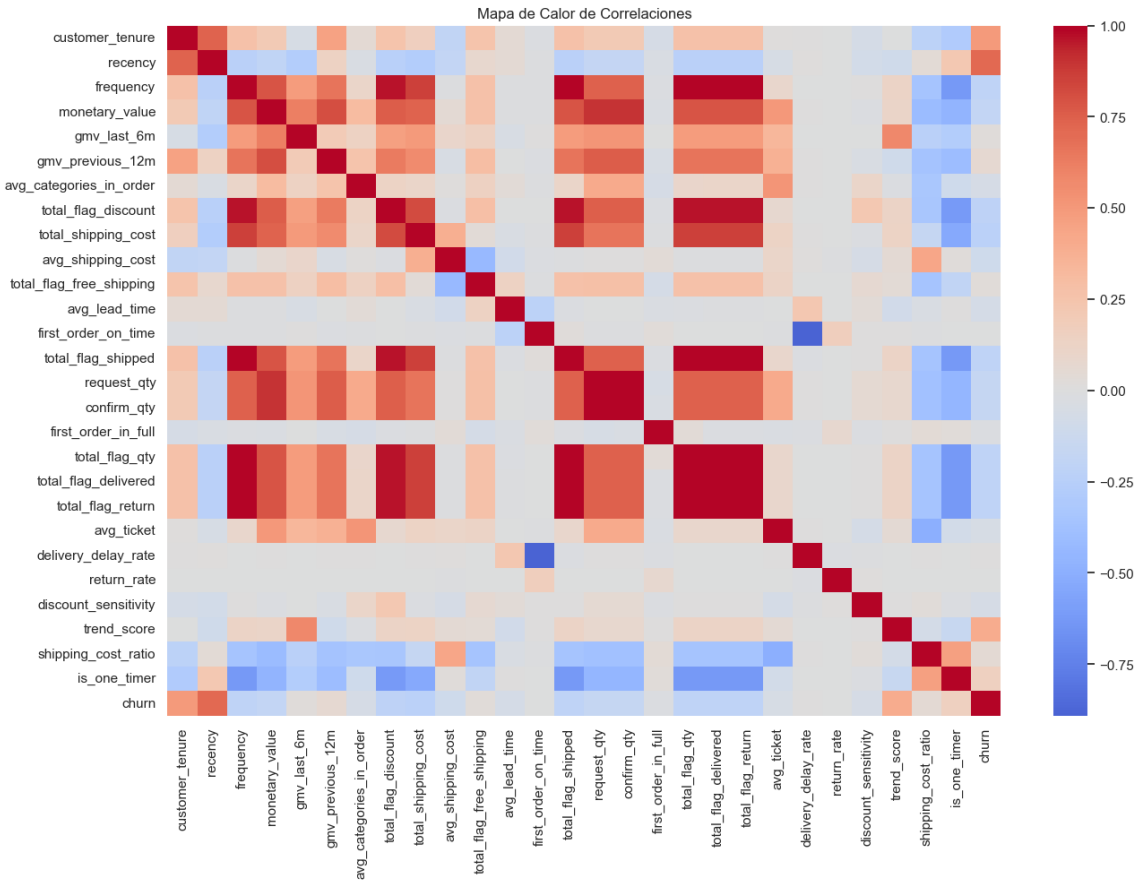


Nota: Autoría propia

Para analizar las variables se grafica un mapa de calor que permite identificar la correlación en las 28 variables, con el fin de identificar linealidad o multicolinealidad que permita seleccionar las variables más relevantes para entrenar los modelos, la gráfica indica que el color rojo intenso tiene una correlación fuerte positiva, el azul intenso correlación negativa fuerte, colores neutros (blanco / gris) correlación débil o nula.

Figura 5

Mapa de Calor de Correlaciones



Nota: Autoría propia

Por último se eliminan variables redundantes basado en la técnica VIF (Variance Inflation Factor) e información mutua (inf), VIF mide cuanto se incrementa la varianza de una variable debido a su correlación con otra, mientras que inf mide la dependencia de dos variables, se eliminan variables redundantes y poco útiles para el modelo que son las que tiene VIF alto y bajo inf, el resultado de aplicar esta técnica permite seleccionar 15 variables para entrenar los modelos: 'customer_tenure', 'frequency', 'avg_categories_in_order', 'avg_shipping_cost', 'total_flag_free_shipping', 'avg_lead_time', 'first_order_on_time', 'first_order_in_full', 'avg_ticket', 'delivery_delay_rate', 'discount_sensitivity', 'trend_score', 'shipping_cost_ratio', 'is_one_timer', 'request_qty'.

La siguiente figura ilustra la metodología propuesta para recolección, anonimización y análisis de datos para luego entrenar los modelos propuestos

Figura 6

Metodología Propuesta

Nota: Autoría propia

Entrenamiento de Modelos

Luego de realizar el preprocesamiento de los datos, se propone aplicar tres modelos de aprendizaje automático seleccionados de la revisión del estado del arte, los modelos a utilizar incluyen regresión logística, random forest y XGBOOST.

Modelo de Regresión Logística

El data set es imputado para valores nulos con la mediana de los datos, se aplica la técnica de *OneHotEncoder* para la variable categórica que representa las marcas '*brand*', a continuación se unen las variables a un pipeline final que incluye el prepocesamiento de datos y un modelo de regresión logística con balance de clases, luego se divide la data en entrenamiento 70 % y prueba 30 %, estratificando por la variable objetivo churn, el pipeline se ajusta con los datos de entrenamiento y continua con la predicción generando las etiquetas 'y_pred' y probabilidades 'y_proba sobre el conjunto de prueba. Continua la evaluación del modelo que incluye las métricas: precisión, recall, F1 y ROC-AUC como métrica principal.

Modelo de Random Forest

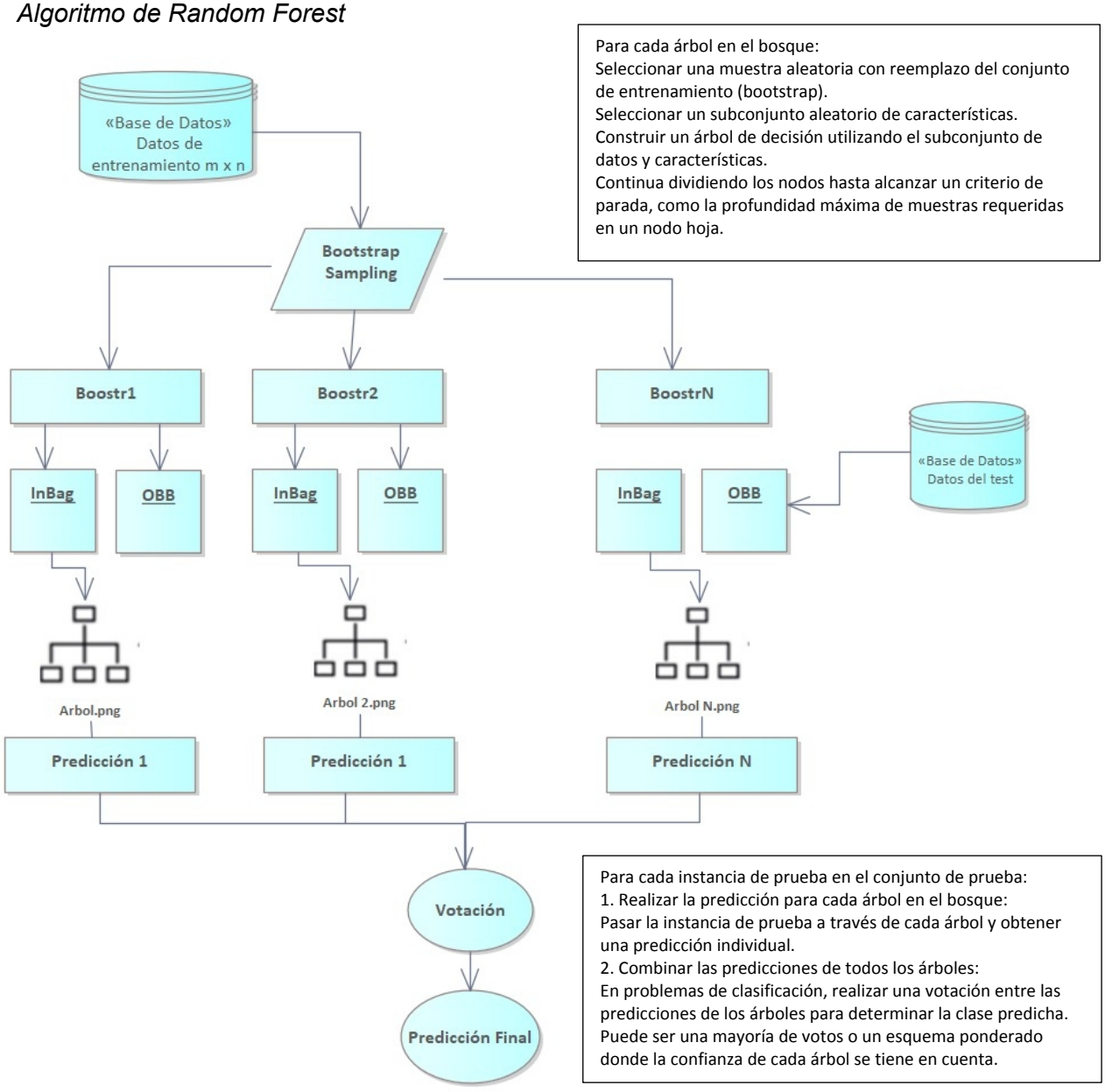
Una vez seleccionada las variables, se codifica la variable categórica 'brand' para etiquetar cada marca estudiada con un número identificador, posteriormente se definen las variables predictoras y la variable objetivo para el estudio 'churn', luego se divide el set de datos en entrenamiento 70 % y prueba 30 %, con los parámetros 'stratify=y' para mantener la proporción de clases y 'random_strate=42' para asegurar reproducibilidad. A continuación, se crea y entrena el modelo utilizando un radom forest classifier, con parámetros clave: 'n_estimators=200' número de árboles en el bosque, 'max_deph=None' sin límite de profundidad, los árboles crecen hasta que no haya más divisiones posibles, 'n_jobs=-1 usa todos los núcleos para acelerar el entrenamiento, por último, se generan predicciones sobre los

datos de prueba ‘X_test’, el resultado ‘y_pred’ es un vector para la predicción de clase churn (si/no).

En la figura 4 se representa la arquitectura de un algoritmo de Random Forest

Figura 7

Algoritmo de Random Forest



Nota: Tomado de Machine Learning, Velasco Rebolledo, Jacinto. 2024.

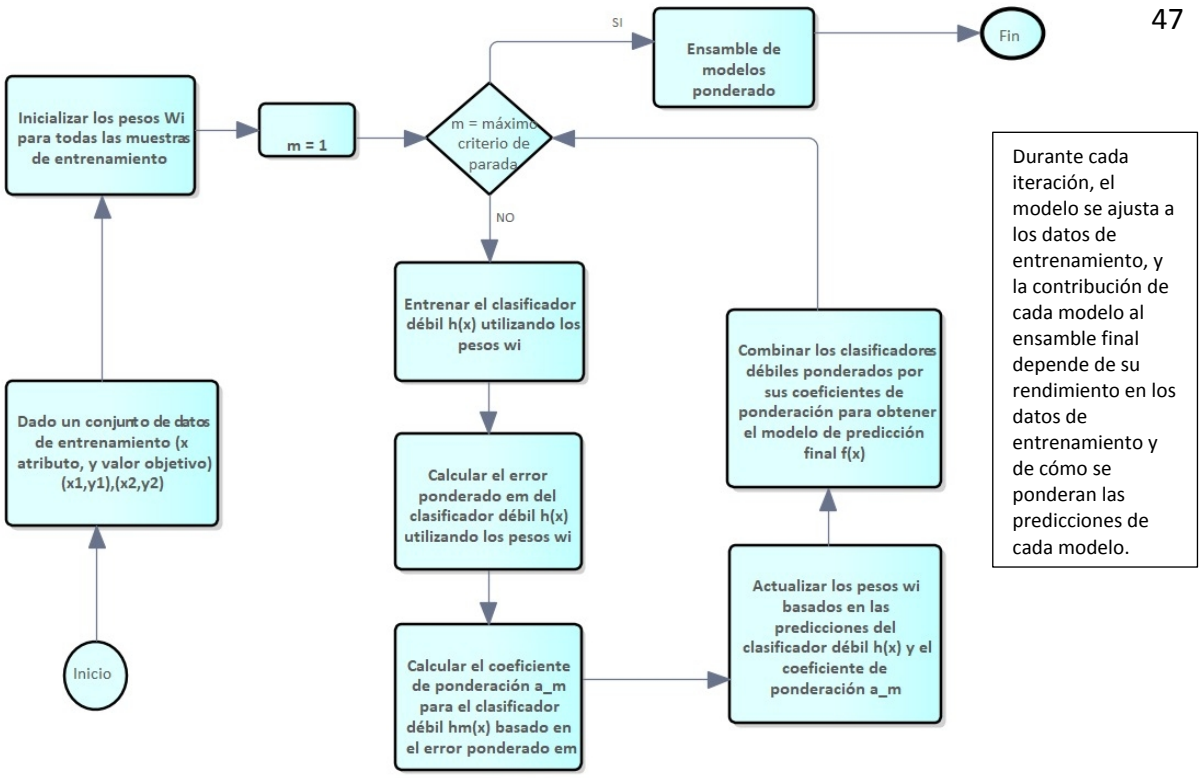
Modelo de XGBoost

Para iniciar se realiza un preprocesamiento con las variables numéricas y la variable categórica, igual que el proceso realizado en los anteriores algoritmos, luego se construye un pipeline que integra el preprocesamiento anterior y el clasificador XGBoost con hiperparámetros ajustados con optimización: ‘n_estimator=1200’ número de árboles, ‘learning_rate=0,01’ tasa de aprendizaje baja para mejor estabilidad, ‘max_depth=7’ profundidad máxima de los árboles, ‘subsample=0.08’ fracción de datos usada en cada iteración para reducir sobreajuste, ‘gamma=0,1’ penalización mínima para dividir nodos usado para controlar complejidad, ‘colsample_bytree=0,5’ fracción de variables utilizadas en cada árbol, ‘scale_pos_weight=1’ balance de clases, permite ajustar si hay desbalance, ‘eval_metric=logloss’ métrica de optimización basada en pérdida logarítmica. A continuación, se define la matriz de características y la variable objetivo, se divide la data en 70 / 30 manteniendo la misma proporción que los anteriores algoritmos, por último, se generan las predicciones y probabilidades, el paso final en todos los algoritmos es generar el reporte de métricas con el fin de evaluar los modelos.

En la figura 8 se representa el diagrama de flujo para desarrollar un algoritmo Boost

Figura 8

Algoritmo Boost



Nota: Tomado de Machine Learning, Velasco Rebolledo, Jacinto. 2024.

Análisis de Resultados

Acorde a los datos analizados de la base de datos del comercio electrónico sector moda, luego de entrenar los modelos seleccionados regresión logística, random forest, XGBOOST, las métricas de obtenidas de la evaluación arrojo los siguientes valores:

Tabla 4.

Métricas de los Modelos

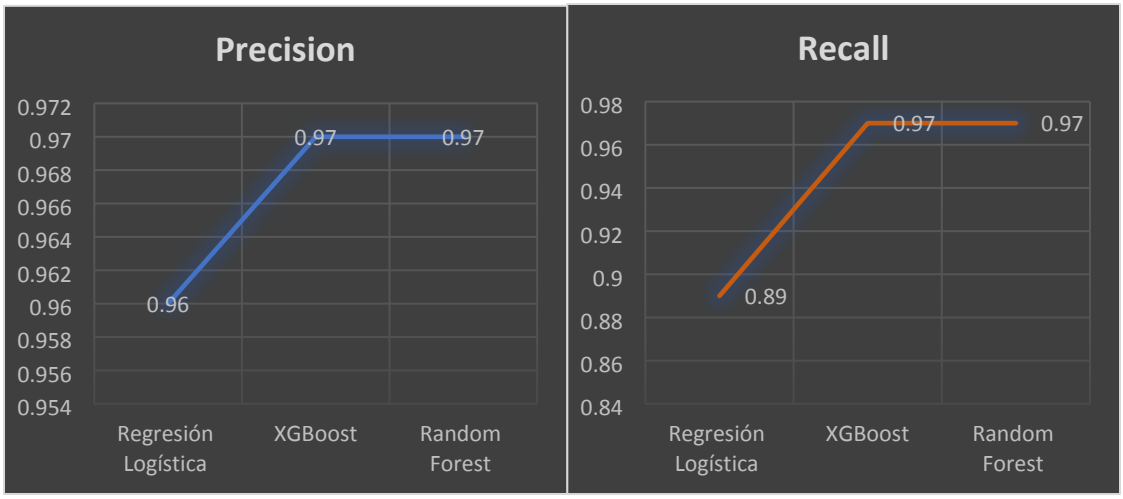
	Métrica				
Modelo	Precisión	Recall	F1-Score	Accuray	ROC-AUC
Regresión Logística	0,96	0,89	0,93	0,9	0,966
XGBoost	0,97	0,97	0,97	0,96	0,9933
Random Forest	0,97	0,97	0,97	0,95	0,9923

Nota: Autoría propia

La siguiente figura relaciona las métricas Precisión y Recall obtenidas por los modelos entrenados.

Figura 9

Métrica Precision - Recall



Nota: Autoría propia

La siguiente figura relaciona las métricas ROC-AUC y Accuracy obtenidas por los modelos entrenados.

Figura 10

Métrica Accuracy – ROC-AUC

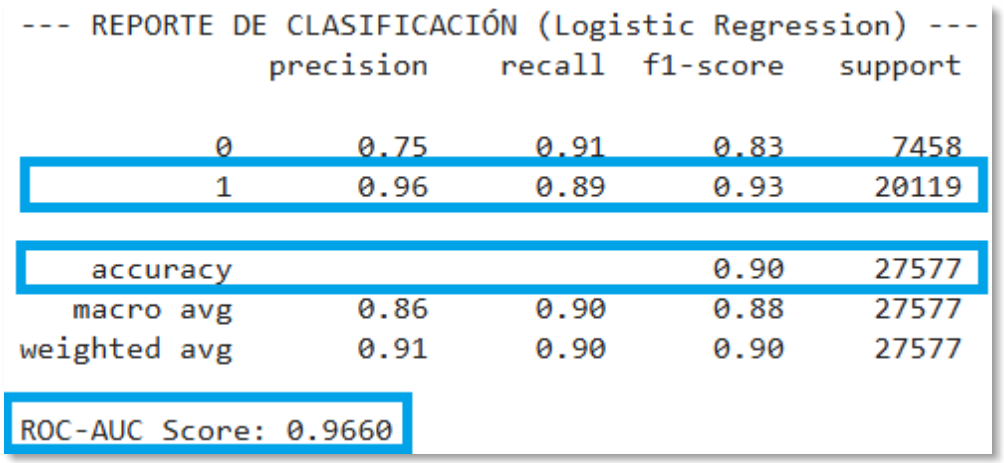
Nota: Autoría propia

Los resultados obtenidos reflejan un desempeño destacado de los modelos de aprendizaje automático evaluados para la predicción del abandono de clientes en el comercio electrónico retail-moda. En generales, los tres algoritmos —Regresión Logística, Random Forest y XGBoost— alcanzaron valores altos en las métricas de evaluación, lo que confirma que existen patrones en el comportamiento de los clientes que pueden ser explotados analíticamente para anticipar el churn.

La regresión logística, utilizada como modelo base por su interpretabilidad, alcanzó una precisión del 96 % y un recall del 89 %, con un F1-score de 93 % y un ROC-AUC de 96,6 %, representados en la siguiente tabla:

Tabla 5.

Métricas del Algoritmo Regresión Logística



--- REPORTE DE CLASIFICACIÓN (Logistic Regression) ---

	precision	recall	f1-score	support
0	0.75	0.91	0.83	7458
1	0.96	0.89	0.93	20119
accuracy			0.90	27577
macro avg	0.86	0.90	0.88	27577
weighted avg	0.91	0.90	0.90	27577

ROC-AUC Score: 0.9660

Nota: Autoría propia

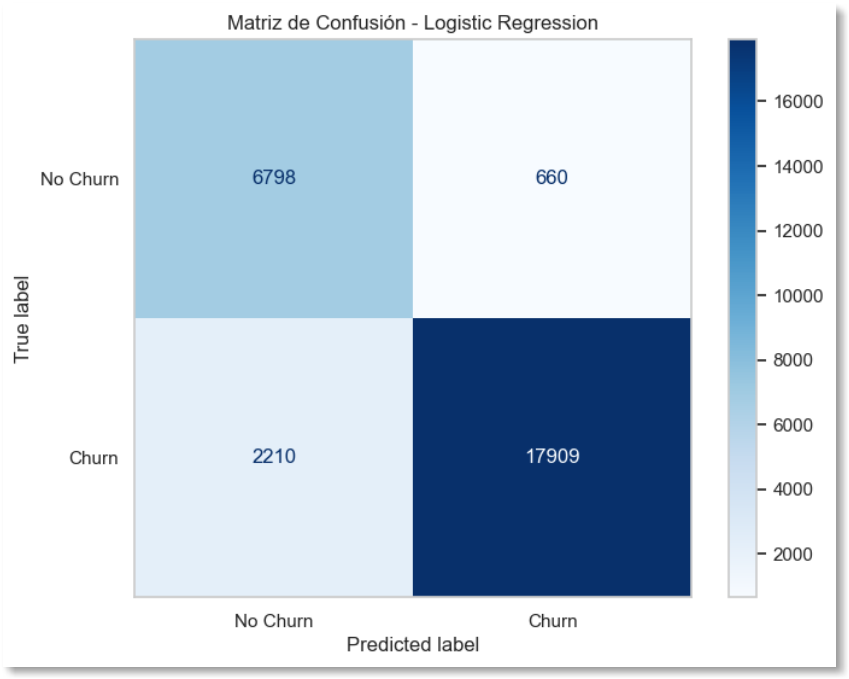
Aunque estos resultados son valiosos, el menor valor de recall indica una limitación en la identificación completa de clientes con probabilidad de abandono, lo que representa un

riesgo desde la perspectiva de negocio, pues los falsos negativos implican la pérdida directa del valor de vida del cliente (CLV).

La siguiente matriz de confusión representa los verdaderos positivos, verdaderos negativos, falsos positivos, falsos negativos, es decir los valores reales que los clientes que abandonaron o no el comercio

Figura 11

Matriz de Confusión – Regresión logística



Nota: Autoría propia

Por su parte, los modelos de Random Forest y XGBoost presentaron un desempeño significativamente superior. Ambos alcanzaron valores de precisión, recall y F1-score del 97 %, lo que muestra alta capacidad para identificar correctamente tanto a los clientes que abandonan como a los que permanecen activos. Este equilibrio entre precisión y sensibilidad

resulta muy relevante en contextos de datos desbalanceados, como es común en comercio electrónico, las tablas 5 y 6 representan las métricas mencionadas anteriormente.

Tabla 6.

Métricas del Algoritmo Random Forest

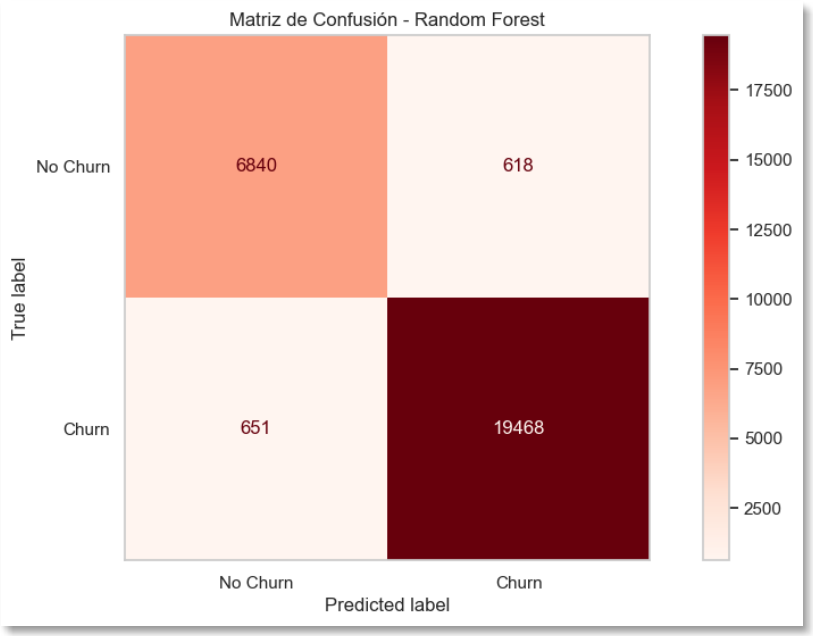
```
--- REPORTE DE CLASIFICACIÓN (Random Forest) ---
              precision    recall  f1-score   support

           0       0.91      0.92      0.92      7458
           1       0.97      0.97      0.97     20119

    accuracy                           0.95     27577
   macro avg       0.94      0.94      0.94     27577
weighted avg       0.95      0.95      0.95     27577

ROC-AUC Score: 0.9424
```

Nota: Autoría propia

La siguiente matriz de confusión representa los valores reales que los clientes que abandonaron o no el comercio para el algoritmo de Random Forest

Figura 12

Matriz de Confusión – Radom Forest



Nota: Autoría propia

Tabla 7.

Métricas del Algoritmo XGBoost

--- REPORTE DE CLASIFICACIÓN (XGBOOST) ---

	precision	recall	f1-score	support
0	0.92	0.92	0.92	7458
1	0.97	0.97	0.97	20119
accuracy			0.96	27577
macro avg	0.95	0.95	0.95	27577
weighted avg	0.96	0.96	0.96	27577

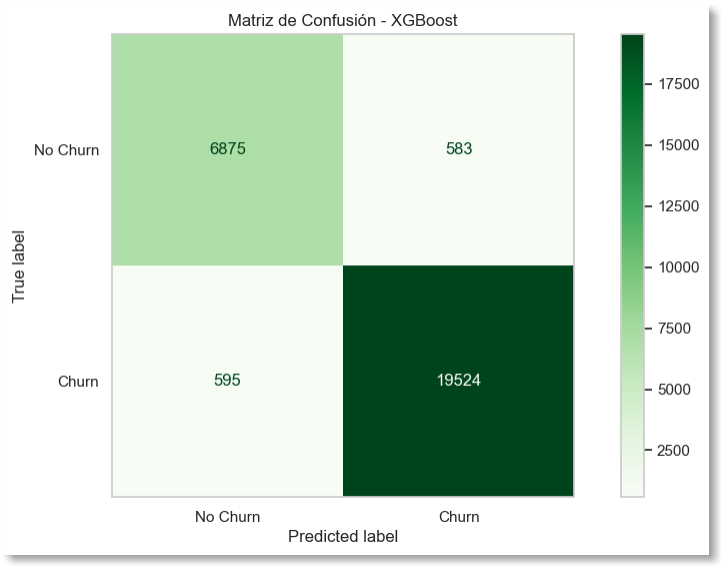
ROC-AUC Score: 0.9933

Nota: Autoría propia

La siguiente matriz de confusión representa los valores reales que los clientes que abandonaron o no el comercio para el algoritmo de XGBoost

Figura 13

Matriz de Confusión – XGBoost

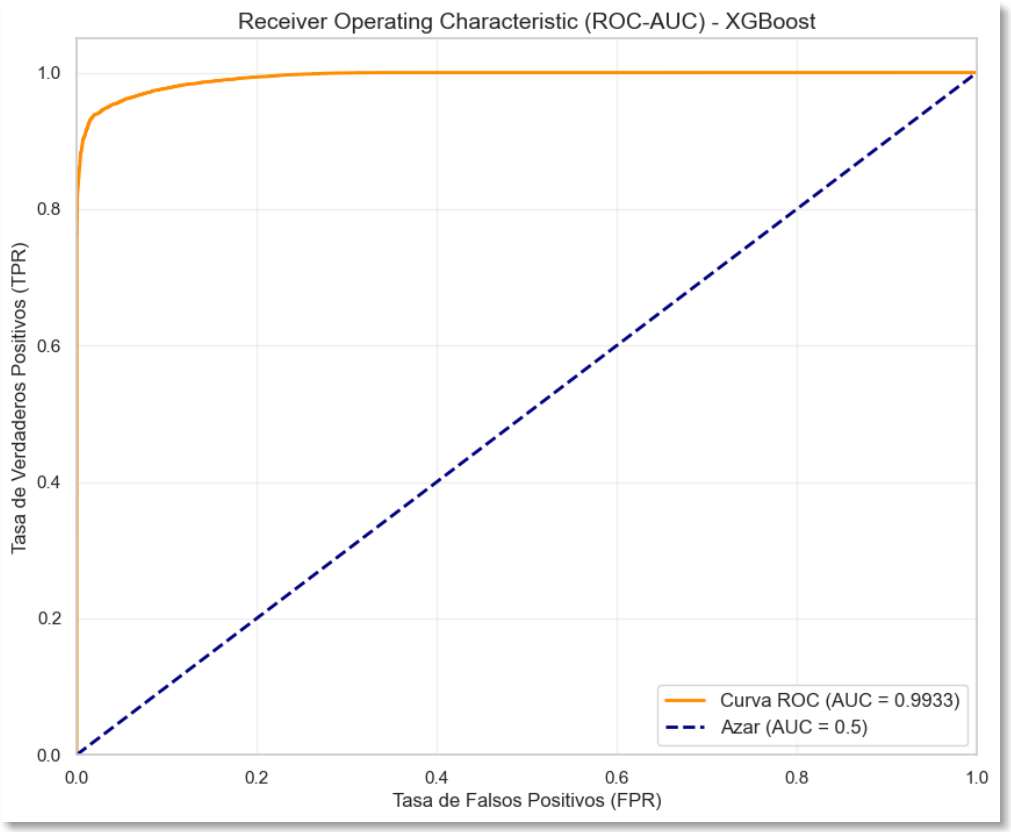


Nota: Autoría propia

El modelo XGBoost se destacó como el algoritmo de mejor desempeño global, con una exactitud del 96 % y un ROC-AUC de 0.9933, lo que refleja alta capacidad para distinguir entre clientes en riesgo y no riesgo de abandono. Este resultado confirma lo planteado en el estado del arte, donde los modelos de ensamblaje basados en boosting muestran una mayor robustez frente a relaciones no lineales y patrones complejos en datos transaccionales.

Figura 14

Gráfica Curva ROC-AUC

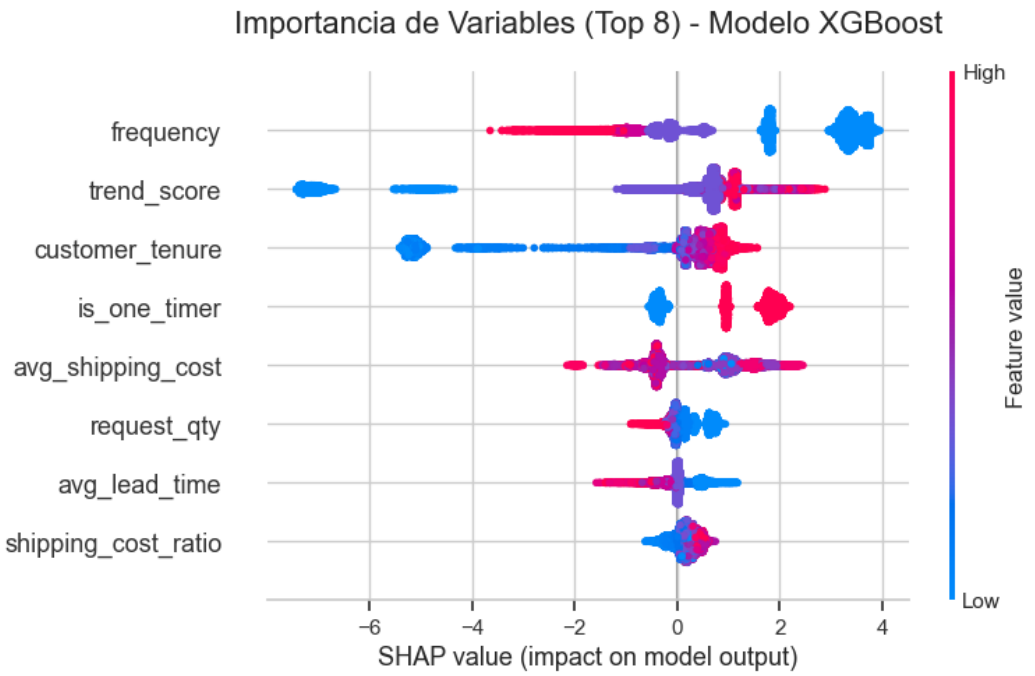


Nota: Autoría propia

La siguiente figura relaciona las variables más importantes para la predicción del abandono de clientes según el modelo entrenado XGBoost, este resultado sirvió como base para formular estrategias de retención para la prevención temprana del churn.

Figura 15

Top 8 Importancia de Variables Algoritmo XGBoost



Nota: Autoría propia

Recomendaciones Estratégicas

En concordancia con el cuarto objetivo específico, el estudio permite formular recomendaciones estratégicas de retención basadas al evaluar el análisis de variables y las predicciones de los modelos. Dichos modelos demuestran que la analítica predictiva, especialmente el algoritmo XGBoost, permite anticipar el abandono del cliente, facilitando diseñar campañas más personalizadas que se adapten a la necesidad de cada cliente. El uso de SHAP permite entender lo que los datos dicen sobre el cliente y como esta transformación de datos se convierte en una herramienta humana donde no se actúa en reacción, si no como un mecanismo de gestión preventiva y de innovación de procesos alimentando el ciclo de mejora continua PHVA.

Planear: Identificar clientes en riesgo

Hacer: Ejecutar estrategias de retención

Verificar: Evaluar métricas y resultados

Actuar: Ajustar procesos y campañas

Tabla 8.

Formulación de Estrategias de Retención

Recomendaciones Estratégicas de Retención		
Variable Relevante (SHAP)	Acción Estratégica	Impacto en la Calidad E Innovación
Frecuencia de compra	Diseñar iniciativas premiando la lealtad o fidelidad por la frecuencia de compra de cada usuario.	Esta estrategia no solo fortalece la permanencia del cliente, si no que reduce la tasa de abandono garantizando la mejora continua dentro del ciclo PHVA.
Trend score (tendencia de actividad)	Poner en marcha indicadores de alerta y desarrollar estrategias de recuperación de clientes ajustadas al perfil del usuario.	Permite una atención anticipada que detecta y corrige las malas experiencias del cliente antes que se conviertan en un gran problema.
Costos de envío	Fortalecer alianzas, coordinar con aliados estratégicos para agilizar tiempos de entrega y reducir costos de envíos.	Permite un posicionamiento estratégico, ventajas competitivas y solidez en el mercado.
Experiencia en el primer pedido	Priorizar las entregas puntuales cumpliendo con plazos de envío y mantener flujo de comunicación constante con el cliente	Garantiza una experiencia inicial satisfactoria asegurando un cliente satisfecho que busque regresar, buscando asegurar que el ciclo PHVA se nutra de experiencias positivas desde la primera compra.
Sensibilidad al descuento	Crear ofertas y promociones que concuerden con lo que el cliente suele comprar.	Esta propuesta permite que la ejecución sea más acorde a las necesidades del cliente lo que lo conecta aún más con la empresa, logrando una gestión más eficiente.
Retrasos en la entrega	Monitorear tiempos de entrega y aplicar mejoras continuas en la cadena logística.	Reduce fallas críticas en la experiencia, alineando procesos con estándares de calidad.

Nota: Autoría propia

Conclusiones

- De acuerdo con el primer objetivo específico se logró estructurar una base de datos de comportamiento de compra y logística que permitió transformar los registros brutos en indicadores de negocio representados en 28 variables que fueron relevantes para entrenar el modelo.
- Respecto al segundo objetivo específico, basado en la literatura, se escogieron 3 algoritmos de aprendizaje automático ampliamente empleados: regresión logística, random forest y XGBoost, a fin de entrenarlos a partir de las variables obtenidas en el conjunto de datos estructurado en el anterior objetivo, para la selección de las variables necesarias para el correspondiente entrenamiento se aplicó el cálculo de la correlación y la técnica VIF (multicolinealidad) logrando seleccionar las 15 variables más influyentes para el modelo.
- Frente al tercer objetivo específico se evaluó el rendimiento de los modelos con las métricas: precision, recall, F1-Score, AUC-ROC, y accuracy, determinando que el modelo con mayor eficacia para la predicción de abandono de clientes fue el XGBoost que obtuvo: precisión 97 %, recall 97 %, F1-Score 97 %, accuracy 96 %, y AUC-ROC 99,33 %.
- Conforme al cuarto objetivo específico, el estudio formuló recomendaciones estratégicas de retención basadas en el análisis de variables y las predicciones del modelo. Los resultados muestran que la analítica predictiva, especialmente con XGBoost, permite anticipar el abandono y diseñar campañas más personalizadas y efectivas, adaptadas al perfil de cada cliente. El uso de SHAP confirmó la relevancia de variables críticas como frecuencia de compra, costos de envío, experiencia en el primer pedido y sensibilidad al descuento, asegurando que las estrategias se fundamenten en causas reales del churn. Así, la analítica se consolida como un mecanismo de gestión

preventiva e innovación de procesos, integrándose al ciclo PHVA y fortaleciendo la competitividad del comercio electrónico.

- El proyecto cumplió con el objetivo general al lograr entrenar un modelo de aprendizaje automático basado en el algoritmo XGBoost, mediante el análisis de patrones de comportamiento de clientes que permitirá identificar de forma temprana y precisa a los usuarios con alta probabilidad de abandono en la plataforma de comercio electrónico sector retail-moda. El modelo XGBoost obtuvo la más alta exactitud de 96 % y un ROC-AUC de 99,33 % lo que permite identificar una excelente capacidad discriminativa, la matriz de confusión y el reporte de métricas confirman la efectividad del modelo para detectar clientes en riesgo.

Referencias

AbdelAziz, N. M., Bekheet, M., Salah, A., El-Saber, N., & AbdelMoneim, W. T. (2025). A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction. *Information (Switzerland)*, *16*(7). https://doi.org/10.3390/info16070537

Aguado López, I. (2020). *Deep Learning: Redes Neuronales Convolucionales en R*.

Ali, N., & Shaban, O. (2024). Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance. *Cogent Business & Management*, *11*. https://doi.org/10.1080/23311975.2024.2361321

Asfe, A. M., Rahman, M. R., & Hossain, M. S. (2025). MNeuralTab: Integrating meta-modeling and neural networks for customer churn prediction in e-commerce. *Discover Applied Sciences*, *7*(6). https://doi.org/10.1007/s42452-025-07157-0

Batta, A., Kar, A. K., & Satpathy, S. (2023). Cross-Platform Analysis of Seller Performance and Churn for Ecommerce Using Artificial Intelligence. *Journal of Global Information Management*, *31*(1). https://doi.org/https://doi.org/10.4018/JGIM.322439

Boozary, P., Sheykhan, S., GhorbanTanhaei, H., & Magazzino, C. (2025). Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. *International Journal of Information Management Data Insights*, *5*(1). https://doi.org/10.1016/j.jjimei.2025.100331

Casanova-Villalba, C. I., Herrera-Sánchez, M. J., & Proaño-González, E. A. (2023). Impacto de la analítica predictiva en la toma de decisiones gerenciales. *Revista Científica Ciencia y Método*, *1*(3), 16–30. https://doi.org/10.55813/gaea/rcym/v1/n3/17

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. P. (2022). SMOTE- synthetic minority over-sampling technique. *Scispace*.

De, C., Fijos En, S., Yesid, A., Pinto, M., Gutiérrez, M. N., Universidad, R., Francisco, D., & De Caldas, J. (2023). *Algoritmo De Predicción De Abandono De Clientes De Servicios Fijos En Telecomunicaciones Basado En Un Conjunto De Datos De La Empresa IBM*.

Fader, P. S., & Hardie, B. G. S. (2007). How to project customer retention. *Journal of Interactive Marketing*, *21*(1), 76–90. https://doi.org/10.1002/dir.20074

Hassan, S. U., Abdulkadir, S. J., Zahid, M. S. M., & Al-Selwi, S. M. (2025). Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review. In *Computers in Biology and Medicine* (Vol. 185). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2024.109569

HKTDC RESEARCH. (2025, January 23). *América Latina: evolución del panorama del mercado del comercio electrónico*. Análisis y Noticias - Mercados de Consumo - Mundo.

Höppner, S., Stripling, E., Baesens, B., Broucke, S. vanden, & Verdonck, T. (2017). *Profit Driven Decision Trees for Churn Prediction*. http://arxiv.org/abs/1712.08101

Kasemrat, R., Kraiwanit, T., Khaothawirat, A., Chinnapha, S., & Luo, Q. (2025). BENCHMARKING MACHINE LEARNING MODELS FOR PREDICTIVE ANALYTICS IN E-COMMERCE STRATEGY. *Corporate and Business Strategy Review*, *6*(2), 146–155. https://doi.org/10.22495/cbsrv6i2art15

LEY ESTATUTARIA 1581 DE 2012, Pub. L. 1581, 2012 (2012).

Liu, Z., Jiang, P., De Bock, K. W., Wang, J., Zhang, L., & Niu, X. (2024). Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. *Technological Forecasting and Social Change*, *198*, 122945. https://doi.org/https://doi.org/10.1016/j.techfore.2023.122945

Margalina, V. (2021). Factores que afectan la intención de compra de los consumidores de moda en el comercio electrónico Un modelo teórico para américa latina. *Sigma*.

Martínez, A., & Segarra Marlon. (2024). *Caracterización del abandono de clientes mediante el uso de algoritmos de aprendizaje automático para empresas de firma electrónica en Ecuador*.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

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

Ministerio de Tecnologías de la Información y las Comunicaciones (MINTIC). (2024, April 23). *Observatorio eCommerce*.

Molnar, C. (2025). *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. http://leanpub.com/interpretable-machine-learning

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. *Procedia Computer Science*, *161*, 834–840. https://doi.org/10.1016/j.procs.2019.11.190

OECD handbook for internationally comparative education statistics 2018 : concepts, standards, definitions and classifications. (2018). OECD Publishing.

Petrescu, M., & Krishen, A. S. (2023). A decade of marketing analytics and more to come: JMA insights. In *Journal of Marketing Analytics* (Vol. 11, Number 2, pp. 117–129). Palgrave Macmillan. https://doi.org/10.1057/s41270-023-00226-6

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. *Business Information Systems*, *1*, 3–12. https://doi.org/10.52825/bis.v1i.42

Régimen Común Sobre Propiedad Industrial, Pub. L. 486, Ministerio de Comercio, Industria y Turismo (2000).

Rojas, J. (2024). *Propuesta para la gestión de la fidelización de clientes para la compañía Hashicorp a tra vés de Machine Learning*.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). *Practical Bayesian Optimization of Machine Learning Algorithms*.

Tam, L. T., Vi, L. G., & Tuan, N. M. (2025). Comparison of Methods for Handling Imbalanced Data in Customer Churn Prediction with Feature Selection Using SHAP and mRMR Frameworks. *Cybernetics and Information Technologies*, *25*(3), 68–87. https://doi.org/10.2478/cait-2025-0023

Tariq, M. U., Babar, M., Poulin, M., & Khattak, A. S. (2022). Distributed model for customer churn prediction using convolutional neural network. *Journal of Modelling in Management*, *17*(3), 853–863. https://doi.org/10.1108/JM2-01-2021-0032

Torres-Campana, M. I., Álvarez-Gavilanes, J. E., & Murillo-Párraga, D. Y. (2025). E-commerce y análisis predictivo mediante big data para anticipar tendencias y comportamientos de compra. *Gestio et Productio. Revista Electrónica de Ciencias Gerenciales*, *7*(2), 235–258. https://doi.org/10.35381/gep.v7i2.314

Velasco Rebolledo, J. (2024). *Machine Learning Fundamentos, algoritmos y aplicaciones para los negocios, industria y finanzas* (1st ed.).