ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS
MAESTRÍA EN SISTEMAS DE INFORMACIÓN, MENCIÓN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

IDENTIFICACION DE SUSCRIPTORES EN RIESGO DE ABANDONO UTILIZANDO MODELAMIENTO DE APRENDIZAJE DE MAQUINA PARA PREDICCION POR CLASIFICACION PARA UNA EMPRESA DE MEDICINA PREPAGADA DEL ECUADOR.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACIÓN MENCIÓN EN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

JORGE LUIS GUATO SANTAMARIA

Director: María Gabriela Pérez Hernández, PhD.
maria.perez@epn.edu.ec

Codirector: Iván Marcelo Carrera Izurieta, PhD.
ivan.carrera@epn.edu.ec

Quito, marzo 2024

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación “Identificación De Suscriptores En Riesgo De Abandono Utilizando Modelamiento de Aprendizaje De Maquina para Predicción Por Clasificación Para Una Empresa De Medicina Prepagada Del Ecuador” desarrollado por Jorge Luis Guato Santamaría estudiante de la Maestría de Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la defensa oral.

María Gabriela Pérez Hernández, PhD.
DIRECTOR

Iván Marcelo Carrera Izurieta, PhD.
CODIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, Jorge Luis Guato Santamaría declaro bajo juramento que el trabajo aquí descrito es de mi autoría; no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

Jorge Luis Guato Santamaría

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

La presente investigación aborda la problemática de la identificación de suscriptores en riesgo de abandono en una empresa de medicina prepagada con sede en Ecuador. El objetivo principal de este estudio es desarrollar e implementar un sistema de predicción por clasificación basado en técnicas de aprendizaje de máquina, que permita identificar a los suscriptores que tienen una alta probabilidad de abandonar los servicios ofrecidos por la empresa.

En la fase de revisión de literatura, se analizan investigaciones previas relacionadas con la retención de clientes en el contexto de la salud y el uso de algoritmos de aprendizaje de máquina para la predicción de abandono en diferentes industrias. Asimismo, se exploran conceptos teóricos clave, como algoritmos de clasificación, técnicas de modelamiento y peculiaridades del sistema de medicina prepagada en Ecuador.

La metodología propuesta involucra la recopilación y análisis de datos históricos de suscriptores, incluyendo información demográfica, patrones de uso de servicios médicos y comportamiento previo de abandono. Se emplearán diversas técnicas de aprendizaje de máquina, como regresión logística, máquinas de soporte vectorial y árboles de decisión, para construir y comparar modelos predictivos. La evaluación de los modelos se realizará mediante métricas de rendimiento, como precisión, sensibilidad y especificidad.

Se espera que los resultados de este estudio proporcionen a la empresa de medicina prepagada una herramienta eficaz para anticipar la posible deserción de sus suscriptores. Así, se podrán diseñar estrategias de retención personalizadas, mejorar la calidad de los servicios ofrecidos y fortalecer la relación con los clientes. Además, esta investigación contribuirá al campo del aprendizaje de máquina aplicado al sector de la salud en Ecuador, ofreciendo una perspectiva útil para otras organizaciones similares que busquen optimizar su retención de clientes.

Palabras clave: Identificación, suscriptores, abandono, modelado, clasificación, medicina prepagada

ABSTRACT

The present research addresses the issue of identifying subscribers at risk of churn in a prepaid medical company based in Ecuador. The main objective of this study is to develop and implement a classification prediction system based on machine learning techniques, allowing the identification of subscribers with a high probability of abandoning the services offered by the company.

In the literature review phase, previous research related to customer retention in the health context and the use of machine learning algorithms for churn prediction in different industries is analyzed. Key theoretical concepts such as classification algorithms, modeling techniques, and peculiarities of the prepaid medical system in Ecuador are also explored.

The proposed methodology involves the collection and analysis of historical subscriber data, including demographic information, patterns of medical service usage, and prior churn behavior. Various machine learning techniques, such as logistic regression, support vector machines, and decision trees, will be employed to build and compare predictive models. Model evaluation will be conducted using performance metrics such as accuracy, sensitivity, and specificity.

It is expected that the results of this study will provide the prepaid medical company with an effective tool to anticipate potential subscriber churn. Thus, personalized retention strategies can be designed, the quality of services offered can be improved, and customer relationships can be strengthened. Additionally, this research will contribute to the field of machine learning applied to the healthcare sector in Ecuador, offering a useful perspective for other similar organizations seeking to optimize customer retention.

Keywords: Identification, subscribers, churn, modeling, classification, prepaid medicine.

1. INTRODUCCIÓN

1.1 Planteamiento del problema

El problema que se aborda en este proyecto es la identificación de suscriptores en riesgo de abandono para una empresa de medicina prepaga en Ecuador utilizando técnicas de modelamiento de aprendizaje de máquina para la predicción por clasificación. La empresa enfrenta una alta tasa de abandono de suscriptores, lo que resulta en una pérdida de ingresos y una disminución en la satisfacción del cliente. Por lo tanto, la identificación temprana de los suscriptores en riesgo de abandonar el servicio y la implementación de medidas para retenerlos son cruciales para la sostenibilidad de la empresa. La empresa de medicina prepagada del Ecuador enfrenta un problema de abandono de sus suscriptores, lo que afecta su rentabilidad y sostenibilidad a largo plazo. La empresa tiene dificultades para identificar con precisión qué suscriptores son más propensos a abandonar la empresa y no ha implementado estrategias de retención efectivas para disminuir la tasa de abandono. Además, la empresa cuenta con una gran cantidad de datos de suscriptores, pero no tiene un método eficaz para analizar y utilizar estos datos para predecir el riesgo de abandono. Por lo tanto, el problema principal que se plantea es cómo desarrollar un modelo de análisis de datos avanzado utilizando técnicas de aprendizaje de máquina que permita a la empresa identificar patrones y tendencias en el comportamiento de sus suscriptores, predecir qué suscriptores tienen mayor riesgo de abandonar la empresa y diseñar estrategias de retención efectivas para reducir la tasa de abandono. Esto ayudará a la empresa a mejorar su rentabilidad y sostenibilidad a largo plazo, así como a mejorar la satisfacción del cliente al brindar un mejor servicio y atención personalizada.

El incremento de la tasa de abandono o deserción de contratos de medicina prepagada tiene un aumento, pueden estar involucrados muchos factores como el tiempo de servicio, el tipo de prestador, si se presenta un reclamo cuanto tiempo se demora en resolverla. Las técnicas tradicionales de retención presentan un retraso en el tiempo con respecto a la cancelación, es decir tienen un enfoque reactivo.

La modalidad de los contratos en su mayoría es anual, por lo tanto, al cumplir el año se procede al proceso de renovación, en la actualidad, los contratos conseguidos en una venta mensual o cosecha, tiene una deserción del 40% al cumplir el año de contrato, es decir el 60% tiende a renovarse y seguir usando el servicio. Por lo tanto, se busca identificar a los afiliados (contratos) con mayor riesgo de desafiliación, y así mejorar la estrategia de retención de cliente, puesto que es más difícil y costoso conseguir un cliente nuevo que retener a un cliente antiguo, focalización tanto la inversión económica, así como el tiempo de reacción a la intención de cancelación del contrato.

Una mala experiencia del cliente en efectivizar sus reembolsos, casos de servicios pagados por el afiliado el cual busca recuperar su deducible, mal información o entendimiento en la cobertura del plan contratado, conlleva a que el afiliado o cliente comience a tener una visión negativa de la aseguradora que contrató.

Por lo tanto, se generan planes estratégicos de retención los cuales tienen una inversión considerable, esta inversión o presupuesto de retención se basa en el análisis del tiempo de vida del contrato, es decir, cuánto dura una cosecha obtenida en un mes (número de contratos nuevos conseguidos en un mes), se analiza las características de las tasas de abandono más fuerte en periodos de un mes y se analiza las características de ese contrato sin embargo no se considera cuáles son las variables principales que influyen en el por qué abandona.

Por lo tanto, el presupuesto de retención no tiene un enfoque hacia las causas principales por las que motiva la desafiliación, sino en las características del contratante, es decir nos centramos en agrupar clientes, pero no en medir la calidad de servicio, por lo tanto, el presupuesto medido para retención se distorsiona o pierde objetividad en identificar los contratos que tienen una más alta posibilidad de abandono.

El identificar las variables que estén directamente correlacionas con los contratos que tienen más alta probabilidad de abandono dará una mejor pauta o guía para acciones más efectivas y mejoras en la inversión en la estrategia de retención, y tener mejor tiempo de reacción hacia un cliente con intención de cancelación. Es muy importante encontrar la causa principal o más representativa para el enfoque del presupuesto de retención sea más

sólido, y de la misma forma optimizado es decir que la inversión sea canalizada a los contratos que verdaderamente tienen alta probabilidad de deserción.

Para abordar este problema, se propone la utilización de técnicas de modelamiento de aprendizaje de máquina por clasificación, con el fin de desarrollar un modelo que permita identificar a los suscriptores en riesgo de abandono. De esta manera, la empresa podrá tomar medidas preventivas oportunas para retener a los clientes y evitar la pérdida económica que implica el abandono de la compañía.

1.2 Objetivos

1.2.1 Objetivo general.

Identificar los suscriptores en riesgo de abandono utilizando modelamiento de aprendizaje de máquina para predicción por clasificación para una empresa de medicina prepagada en Ecuador.

1.2.2 Objetivos específicos.

1. Entender el comportamiento del negocio de una empresa de medicina prepagada en el Ecuador, su entorno y relación con los clientes, así como de una revisión de la literatura enfocada en el estudio de contribuciones relacionadas con el uso de modelamiento de aprendizaje de máquina para el análisis de clientes en riesgo de abandono en empresas privadas.
2. Identificar las variables más importantes que pueden estar relacionadas con el abandono de los suscriptores, seleccionando y preprocesando los datos relevantes para el modelamiento de aprendizaje de máquina, incluyendo la eliminación de datos incompletos o irrelevantes y la transformación de los datos en un formato adecuado para aplicar técnicas de análisis exploratorio de datos.
3. Desarrollar un modelo de aprendizaje de máquina para predecir el riesgo de abandono de los suscriptores utilizando técnicas de clasificación, como la

regresión logística, el árbol de decisión o el algoritmo de máquinas de vectores de soporte.

4. Validar el modelo desarrollado utilizando técnicas de validación cruzada y curva ROC, y determinar su precisión, sensibilidad y especificidad.
5. Evaluar el rendimiento del modelo utilizando medidas de precisión y de error, como la matriz de confusión y la aplicación de los respectivos ajustes en el modelo para mejorar los resultados.

1.3 Alcance

El alcance de este estudio se centra en desarrollar un modelo de aprendizaje de máquina para la identificación de suscriptores en riesgo de abandonar su membresía en una empresa de medicina prepagada en Ecuador. La investigación abarcará la adquisición y el preprocesamiento de datos históricos relevantes, la selección y aplicación de algoritmos de aprendizaje de máquina para la predicción por clasificación, y la evaluación rigurosa del rendimiento del modelo. Además, se explorarán variables significativas que influyen en el abandono de suscriptores en el contexto de la medicina prepagada en Ecuador. Las recomendaciones prácticas derivadas de los resultados servirán como guía estratégica para la empresa, con el objetivo de mejorar la retención de sus suscriptores y, en última instancia, fortalecer su posición en el mercado de la medicina prepagada ecuatoriana.

1.4 Marco Teórico

1.4.5 Medicina Prepagada en Ecuador

La medicina prepagada en Ecuador se ha erigido como un sistema de atención médica de creciente relevancia, caracterizado por permitir a los individuos y familias acceder a una amplia gama de servicios de salud mediante el pago anticipado de una membresía. En este contexto, el análisis de la medicina prepagada adquiere una importancia fundamental, dado que constituye el trasfondo esencial para comprender los factores que inciden en la retención de suscriptores, un tema de indiscutible relevancia en la presente investigación. Este sistema de salud se distingue por ofrecer cobertura integral, abarcando desde consultas médicas de rutina hasta procedimientos especializados y hospitalización. Los suscriptores efectúan pagos anticipados periódicos, lo que garantiza su acceso a la atención médica durante el período de validez de su membresía.

Además, las empresas de medicina prepagada establecen alianzas estratégicas con una red de proveedores médicos, hospitales y clínicas, lo que facilita el acceso a servicios de atención médica de alta calidad. Sin embargo, la retención de suscriptores en este contexto presenta desafíos significativos, como la competencia en un mercado saturado, los cambios en las circunstancias personales de los suscriptores y la percepción del valor de

la membresía. La presente investigación se sumerge en este entorno dinámico para desarrollar un modelo de aprendizaje de máquina que permita identificar suscriptores en riesgo de abandono y proporcionar a la empresa de medicina prepagada las herramientas necesarias para tomar medidas proactivas con el fin de retener a estos suscriptores y mejorar la calidad de su servicio en el contexto ecuatoriano.

Figura 1. Entes de control de empresas de medicina prepagada.

1.4.5 Actores en la Medicina Prepagada en Ecuador

En el contexto de la medicina prepagada en Ecuador, varios actores y datos clave desempeñan un papel fundamental. Los actores principales incluyen las empresas de medicina prepagada, los suscriptores o beneficiarios, los proveedores de atención médica, y las entidades reguladoras gubernamentales. Las empresas de medicina prepagada son las protagonistas, ofreciendo una variedad de planes y servicios de atención médica. Los suscriptores, por su parte, son los clientes que adquieren membresías y utilizan los servicios de atención médica. Los proveedores de atención médica, como hospitales, médicos y clínicas, forman una red crucial para ofrecer servicios médicos a los suscriptores. Además, las entidades reguladoras gubernamentales supervisan y regulan la industria de la medicina prepagada en términos de calidad y cumplimiento normativo. En cuanto a los datos, se recopilan información detallada sobre los suscriptores, que abarca desde su historial médico hasta la frecuencia de uso de los servicios. Los datos financieros y de facturación también son esenciales para el funcionamiento de estas empresas. El análisis de estos datos y la interacción entre estos actores son aspectos críticos en la identificación de suscriptores en riesgo de abandono y la posterior aplicación de modelos de aprendizaje de máquina para mejorar la retención de suscriptores en el contexto específico de la medicina prepagada en Ecuador.

Dentro del ámbito de la medicina prepagada en Ecuador, los suscriptores representan el núcleo esencial de este sistema de atención médica. Estos suscriptores, también conocidos

como beneficiarios, son individuos o familias que han optado por adquirir membresías en empresas de medicina prepagada para acceder a una variedad de servicios de atención médica. La relación entre los suscriptores y las empresas de medicina prepagada es fundamental, ya que los primeros dependen de los servicios proporcionados por las segundas para satisfacer sus necesidades médicas. Los suscriptores varían en edad, estado de salud, necesidades médicas y preferencias, lo que añade complejidad a la gestión de estos clientes. Sus datos, que incluyen información personal, historiales médicos y patrones de uso de los servicios de salud, constituyen un recurso invaluable para la identificación de suscriptores en riesgo de abandono. En este contexto, la presente investigación se enfoca en el análisis de estos datos de los suscriptores para desarrollar un modelo de aprendizaje de máquina que permita prever y abordar de manera proactiva el abandono de membresías, mejorando así la retención y la satisfacción de los suscriptores en el contexto específico de la medicina prepagada en Ecuador.

Dentro del marco de la medicina prepagada en Ecuador, los prestadores de servicios de salud representan un componente esencial de este sistema. Estos prestadores son hospitales, clínicas, médicos y otros profesionales de la salud que forman parte de la red de proveedores afiliados a las empresas de medicina prepagada. La colaboración entre las empresas de medicina prepagada y los prestadores de servicios de salud es fundamental, ya que los suscriptores dependen de esta red para recibir atención médica oportuna y de calidad. Los prestadores de servicios de salud desempeñan un papel clave en la satisfacción de los suscriptores y en la calidad general de la atención médica brindada. Además, los datos relacionados con la utilización de servicios, la calidad de la atención y la retroalimentación de los suscriptores sobre sus experiencias con los prestadores son valiosos para comprender la dinámica de la medicina prepagada en Ecuador. En el contexto de esta investigación, se considerará la relación entre las empresas de medicina prepagada y los prestadores de servicios de salud, así como la disponibilidad y calidad de los datos relacionados con los prestadores para desarrollar un modelo de aprendizaje de máquina que contribuya a la identificación de suscriptores en riesgo de abandono y, en última instancia, a mejorar la retención de suscriptores en el sector de la medicina prepagada en Ecuador.

Además, la Agencia de Aseguramiento de la Calidad de los Servicios de Salud y Medicina Prepagada (ACESS) es la entidad encargada de regular y supervisar los servicios de

medicina prepagada en Ecuador. Es importante considerar las regulaciones y normativas establecidas por esta entidad en el marco teórico del tema, ya que estas pueden influir en la forma en que las empresas de medicina prepagada operan y en la calidad de los servicios que ofrecen.

1.4.3. Comportamiento de Suscriptores y Prestadores en la medicina prepagada en Ecuador

El comportamiento de los suscriptores en el contexto de la medicina prepagada en Ecuador se revela como un aspecto de suma importancia en la comprensión de la dinámica de este sistema de atención médica. Los suscriptores, como usuarios activos de los servicios de salud prepagados, exhiben una variedad de patrones de comportamiento que abarcan desde la frecuencia de utilización de servicios médicos hasta la interacción con proveedores de atención médica. Este comportamiento puede estar influenciado por factores individuales, como la edad, el estado de salud, las preferencias personales y las circunstancias cambiantes de la vida. Además, las experiencias previas de los suscriptores, su percepción de la calidad de la atención y su nivel de satisfacción con la empresa de medicina prepagada juegan un papel crucial en su decisión de mantener o abandonar su membresía. El análisis y la comprensión de estos patrones de comportamiento, respaldados por datos históricos y transaccionales, son fundamentales para el desarrollo de modelos de aprendizaje de máquina que puedan identificar de manera precisa a los suscriptores en riesgo de abandonar su membresía, lo que permitiría a la empresa de medicina prepagada tomar medidas proactivas y personalizadas para retener a estos suscriptores, mejorando así la retención y la satisfacción general de los usuarios en el contexto específico de la medicina prepagada en Ecuador.

El comportamiento de los prestadores de servicios de salud en el contexto de la medicina prepagada en Ecuador es un aspecto de crucial importancia para comprender el funcionamiento de este sistema. Los prestadores, que incluyen hospitales, clínicas, médicos y otros profesionales de la salud, desempeñan un papel vital en la prestación de servicios médicos a los suscriptores. Su comportamiento abarca desde la calidad y la eficiencia de la atención brindada hasta la comunicación y la colaboración con las empresas de medicina prepagada. La relación entre los prestadores y las empresas de

medicina prepagada es un factor determinante en la satisfacción y la retención de los suscriptores, ya que estos últimos confían en la red de proveedores para recibir atención médica oportuna y de calidad. La eficacia de los prestadores en la atención al cliente, la gestión de reclamos y la coordinación de servicios también influyen en la percepción general de la calidad de la atención médica.

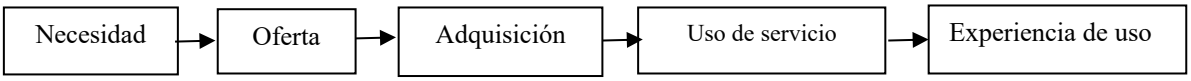


Figura 2. Ciclo suscripción de servicio de medicina prepagada.

El análisis y la comprensión del comportamiento de los prestadores, junto con la disponibilidad de datos relacionados con su desempeño y la satisfacción de los suscriptores, son fundamentales para desarrollar modelos de aprendizaje de máquina que contribuyan a la identificación de suscriptores en riesgo de abandono. Esto permitirá a la empresa de medicina prepagada tomar medidas específicas y colaborativas con los prestadores para mejorar la retención de suscriptores y elevar la calidad de la atención médica en el contexto único de la medicina prepagada en Ecuador.

1.4.4 Aprendizaje de Maquina

El aprendizaje de máquina es una rama de la inteligencia artificial que proporciona a los sistemas la capacidad de aprender y mejorar de manera automática a partir de la experiencia. Los modelos de aprendizaje de máquina se alimentan de datos para realizar predicciones de manera robusta. Cuantos más datos se tengan, mejor será el modelo.

En el aprendizaje de máquina, existen dos tipos principales de algoritmos: el aprendizaje supervisado y el aprendizaje no supervisado. En el aprendizaje supervisado, se entrena el modelo con datos etiquetados que especifican tanto la entrada como la salida del algoritmo. Los algoritmos de clasificación y regresión son ejemplos de aprendizaje supervisado. En el aprendizaje no supervisado, se trabajan con datos que no han sido etiquetados. Estos algoritmos se utilizan principalmente en tareas donde es necesario analizar los datos para extraer nuevo conocimiento o agrupar entidades por afinidad.

Es importante tener en cuenta que la calidad de los datos utilizados para entrenar el modelo es fundamental para la precisión de las predicciones. Además, se debe considerar la selección del algoritmo de aprendizaje supervisado más adecuado para el problema en cuestión, así como la evaluación del rendimiento del modelo utilizando métricas de evaluación adecuadas.

El Aprendizaje de maquina tiene una amplia variedad de aplicaciones en diferentes campos, como la medicina, la agricultura, la industria, la banca, el comercio electrónico, entre otros. En el marco teórico del tema "Identificación de suscriptores en riesgo de abandono utilizando modelamiento de aprendizaje de máquina para predicción por clasificación para una empresa de medicina prepagada del Ecuador", se utiliza el AA para predecir el abandono de los suscriptores de una empresa de medicina prepagada en Ecuador. Para ello, se utiliza un modelo de aprendizaje supervisado que se entrena con datos etiquetados para realizar predicciones precisas.

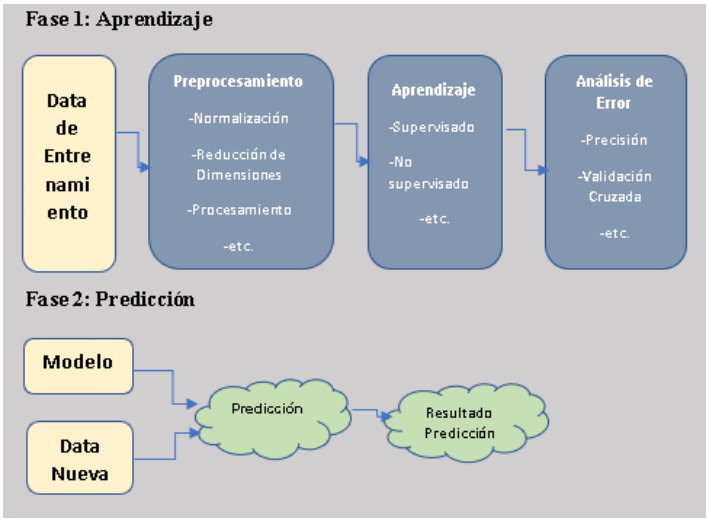


Figura 3. Diagrama de Proceso de Aprendizaje de Maquina

1.4.5 Tipos de Algoritmos de Aprendizaje de Maquina

Existen varios tipos de algoritmos de aprendizaje de máquina, y se pueden clasificar en función de la naturaleza de la tarea que realizan. Aquí hay algunas categorías generales de algoritmos de aprendizaje de máquina:

- Aprendizaje Supervisado:
  En el aprendizaje supervisado, el modelo se entrena utilizando un conjunto de datos que contiene ejemplos etiquetados, es decir, datos emparejados con las respuestas deseadas. Los algoritmos de aprendizaje supervisado incluyen:
  Regresión Lineal: Para problemas de predicción numérica.
  Regresión Logística: Para problemas de clasificación binaria.
  Máquinas de Soporte Vectorial (SVM): Para clasificación y regresión.
  Árboles de Decisión y Bosques Aleatorios: Para clasificación y regresión.
  Redes Neuronales: Modelos inspirados en la estructura del cerebro, utilizados para tareas complejas.
- Aprendizaje No Supervisado:
  En el aprendizaje no supervisado, el modelo se entrena en datos sin etiquetas. Los algoritmos de aprendizaje no supervisado incluyen:
  Agrupamiento (Clustering): K-Means, Hierarchical Clustering.
  Reducción de Dimensionalidad: Análisis de Componentes Principales (PCA), T-Distributed Stochastic Neighbor Embedding (t-SNE).
  Reglas de Asociación: Apriori, Eclat.
- Aprendizaje por Reforzamiento:
  En el aprendizaje por reforzamiento, el modelo toma decisiones secuenciales en un entorno y recibe retroalimentación en forma de recompensas o castigos. Algoritmos de aprendizaje por reforzamiento incluyen:
  Q-Learning: Utilizado en problemas de toma de decisiones secuenciales.
  Algoritmos de Política: Como los métodos de Monte Carlo.
- Aprendizaje Semisupervisado:
  Combina elementos del aprendizaje supervisado y no supervisado, donde el modelo se entrena con datos parcialmente etiquetados.
- Aprendizaje Profundo (Deep Learning):

Utiliza redes neuronales profundas para aprender representaciones jerárquicas de datos. Incluye arquitecturas como redes neuronales convolucionales (CNN) para imágenes, y redes neuronales recurrentes (RNN) para datos secuenciales.

- Aprendizaje por Transferencia:
  Utiliza conocimiento aprendido en una tarea para mejorar el rendimiento en otra tarea relacionada.
- Aprendizaje Ensemble:
  Combina múltiples modelos para mejorar el rendimiento. Ejemplos incluyen Bosques Aleatorios, Gradient Boosting, y Bagging.
- Aprendizaje No Convencional:
  Incluye enfoques novedosos como Aprendizaje Automático Explicable (XAI), Aprendizaje Federado (Federated Learning), y Aprendizaje por Máquinas de Soporte Vectorial Cuánticas.

1.4.6 Importancia del Aprendizaje de Máquina

El aprendizaje de máquina es un campo crucial en la actualidad debido a su capacidad para extraer conocimiento valioso de grandes conjuntos de datos y automatizar tareas complejas. Aquí se destacan algunas de las razones clave que explican la importancia del aprendizaje de máquina:

- Automatización y Eficiencia:
  El aprendizaje de máquina permite la automatización de tareas complejas, lo que mejora la eficiencia en diversas industrias. Automatizar procesos puede llevar a una reducción de costos y a una mayor productividad.
- Toma de Decisiones Basada en Datos:
  Algoritmos de aprendizaje de máquina permiten tomar decisiones basadas en datos en lugar de depender exclusivamente de intuiciones humanas. Esto puede llevar a decisiones más informadas y precisas.
- Análisis Predictivo:
  El aprendizaje de máquina permite prever patrones y tendencias en los datos, lo que facilita la toma de decisiones anticipadas. Esto es valioso en áreas como finanzas, salud, marketing y más.

- Personalización y Recomendaciones:

  Muchas plataformas en línea utilizan algoritmos de aprendizaje de máquina para personalizar experiencias de usuario y ofrecer recomendaciones personalizadas. Ejemplos incluyen recomendaciones de productos, contenido personalizado y servicios adaptados a las preferencias individuales.
- Avances en Medicina:

  En medicina, el aprendizaje de máquina se utiliza para diagnóstico médico, pronóstico de enfermedades, descubrimiento de medicamentos y personalización de tratamientos. Estas aplicaciones pueden mejorar la precisión y la eficacia de la atención médica.
- Automatización de Procesos Industriales:

  En la industria, el aprendizaje de máquina se aplica para el mantenimiento predictivo de maquinaria, la optimización de procesos de fabricación y la mejora de la cadena de suministro.
- Entendimiento de Datos Complejos:

  Ayuda a comprender y extraer información de conjuntos de datos grandes y complejos que serían difíciles de analizar manualmente. Esto es especialmente útil en campos como la investigación científica y la exploración de datos.
- Aprendizaje Continuo y Adaptabilidad:

  Los modelos de aprendizaje de máquina pueden adaptarse y mejorar con el tiempo a medida que se alimentan con más datos. Esto permite la creación de sistemas más robustos y adaptativos.
- Desarrollo de Tecnologías Emergentes:

  El aprendizaje de máquina es fundamental para el desarrollo de tecnologías emergentes como vehículos autónomos, asistentes virtuales, y la inteligencia artificial en general.
- Seguridad y Detección de Fraudes:

  Se utiliza para identificar patrones anómalos y detectar posibles fraudes en transacciones financieras, sistemas de seguridad y otras aplicaciones.
- Aprendizaje Federado y Privacidad:

  El aprendizaje federado permite entrenar modelos de manera colaborativa sin compartir datos sensibles, lo que aborda preocupaciones de privacidad.

En resumen, el aprendizaje de máquina ha demostrado ser esencial en una variedad de campos, impulsando la innovación, mejorando la eficiencia y permitiendo nuevas formas de abordar problemas complejos en la sociedad y la industria.

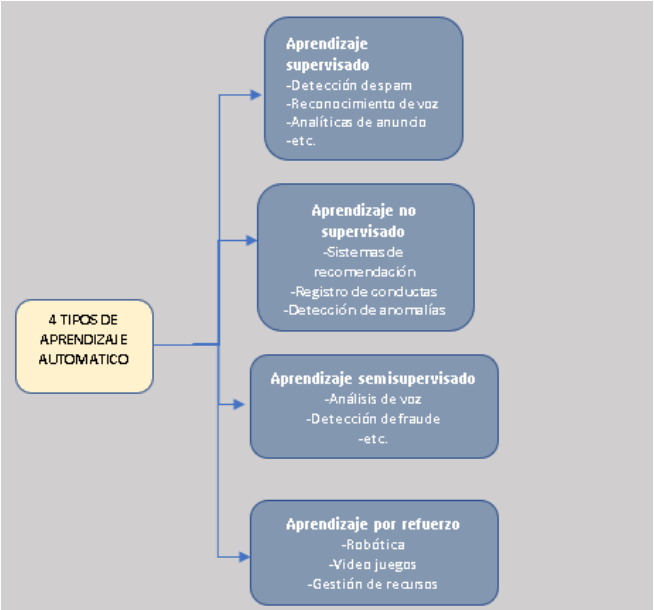


Figura 4. Resumen tipos aprendizaje automático

2. METODOLOGÍA

En el desarrollo del proyecto sobre la identificación de suscriptores en riesgo de abandono en una empresa de medicina prepagada en Ecuador, se sigue la metodología CRISP-DM. El primer paso implica comprender a fondo los objetivos del negocio, centrándose en la identificación proactiva de suscriptores que podrían abandonar el servicio. Se establecen requisitos y limitaciones del proyecto para guiar la implementación efectiva de la solución.
La segunda fase se concentra en la comprensión de los datos disponibles. Se recopilan datos históricos relevantes, incluyendo información demográfica, historial de uso y transacciones. La exploración de datos ayuda a entender la estructura y la calidad de la información, proporcionando una base sólida para el análisis subsiguiente.
Posteriormente, se lleva a cabo la preparación de datos, abordando cualquier problema como valores faltantes y realizando transformaciones necesarias para el modelado. Este paso es esencial para garantizar que los datos estén listos y sean adecuados para la aplicación de modelos de aprendizaje de máquina.
La fase de modelado implica la selección de algoritmos de clasificación específicos para prever la deserción de suscriptores. Aquí es donde se aplican técnicas de aprendizaje de máquina para construir un modelo predictivo basado en los datos disponibles.
La evaluación del modelo es crítica para medir su rendimiento. Se utilizan métricas de clasificación, como precisión y eficacia, para determinar la capacidad del modelo para prever con precisión la deserción de los suscriptores.
Una vez que el modelo ha sido evaluado y ajustado según sea necesario, se procede a su implementación en producción. Esto implica la integración del modelo en el entorno operativo de la empresa de medicina prepagada, permitiendo su aplicación práctica.
Finalmente, se establece un proceso de monitoreo continuo para seguir de cerca el rendimiento del modelo en situaciones reales. Se realizan ajustes según sea necesario para garantizar que el modelo siga siendo efectivo a lo largo del tiempo y en respuesta a cambios en los datos o en el entorno empresarial.

2.1 Metodología CRISP-DM.

La metodología CRISP-DM, que significa "Cross-Industry Standard Process for Data Mining" (Proceso Estándar Intersectorial para la Minería de Datos), es una metodología estándar utilizada en el campo de la minería de datos. Proporciona una estructura sistemática para guiar a los profesionales a través de las diversas fases de un proyecto de minería de datos. Estas fases incluyen la comprensión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación, el despliegue y el monitoreo continuo. CRISP-DM es un enfoque cíclico e iterativo, lo que significa que las fases pueden repetirse según las necesidades del proyecto. Es ampliamente utilizado para desarrollar modelos predictivos y descriptivos en diversos sectores y contextos [1] [4].

En la Figura 5 se detalla acorde a la metodología CRISP-DM las etapas según este estudio.

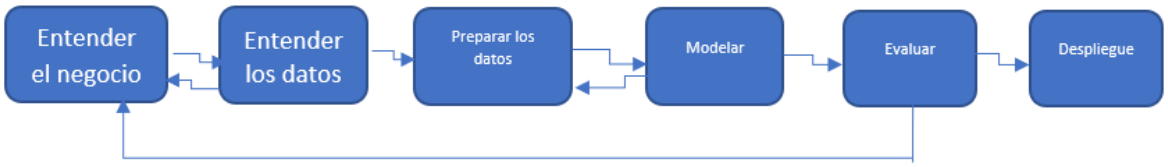


Figura 5. Metodología CRISP-DM

2.1.1 Entendimiento del Negocio.

El entendimiento del negocio es una fase crítica en la metodología CRISP-DM, especialmente en el contexto del proyecto. Esta etapa establece las bases para el éxito del proyecto al definir de manera clara y precisa los objetivos y requisitos del negocio [4].

En el contexto de la investigación, el entendimiento del negocio implica una inmersión profunda en la dinámica de la industria de la medicina prepagada en Ecuador. Algunos elementos clave a considerar en esta fase incluyen:

- Objetivos del Negocio:

  Identificar y comprender los objetivos específicos de la empresa de medicina prepagada en relación con la retención de suscriptores y la reducción de la deserción. Esto podría incluir metas financieras, de retención de clientes y mejoras en la calidad del servicio.

- Acceso a la Atención Médica:

  Garantizar que los afiliados tengan acceso oportuno y eficiente a servicios de atención médica de calidad en la red de proveedores asociados.
- Cobertura Integral:

  Ofrecer planes de medicina prepagada que cubran una amplia gama de servicios médicos, incluyendo consultas, exámenes, hospitalización, cirugías, entre otros.
- Calidad en la Atención:

  Mantener altos estándares de calidad en la atención médica, asegurándose de que los servicios ofrecidos cumplan con las normativas y expectativas de los afiliados.
- Prevención y Bienestar:

  Implementar programas de prevención y promoción de la salud para ayudar a los afiliados a mantener un estilo de vida saludable y reducir la necesidad de atención médica.
- Crecimiento de la Base de Afiliados:

  Atraer nuevos clientes y retener a los existentes para expandir la base de afiliados y aumentar la participación en los planes de medicina prepagada.
- Cobertura Geográfica:

  Ampliar la cobertura geográfica para llegar a más regiones y comunidades, brindando acceso a servicios de salud a un número mayor de personas.
- Satisfacción del Cliente:

  Mantener altos niveles de satisfacción del cliente mediante la mejora continua de servicios, la atención al cliente sea efectiva y la resolución rápida de problemas.
- Eficiencia Operativa:

  Optimizar los procesos internos para garantizar la eficiencia operativa y reducir los costos, lo que puede traducirse en beneficios para los afiliados.
- Innovación en Servicios:

  Introducir innovaciones en servicios médicos, como la implementación de tecnologías de la información en la gestión de la atención médica y la introducción de servicios especializados.
- Conformidad con la Legislación:

  Garantizar el cumplimiento de todas las regulaciones y requisitos legales en el ámbito de la medicina prepagada en Ecuador.
- Rentabilidad Financiera:

Mantener una gestión financiera saludable para garantizar la sostenibilidad a largo plazo y ofrecer beneficios económicos a la empresa y sus afiliados.

- Desarrollo de Alianzas Estratégicas:
  Establecer alianzas estratégicas con proveedores de servicios de salud, laboratorios, clínicas y otros actores relevantes para fortalecer la red de atención médica.
- Contexto Empresarial:
  Analizar el entorno empresarial de la medicina prepagada en Ecuador. Comprender la competencia, las tendencias del mercado, regulaciones y cualquier otro factor que pueda influir en la deserción de suscriptores.
- Partes Interesadas y Expectativas:
  Identificar y comunicarse con las partes interesadas clave dentro de la empresa. Entender sus expectativas, preocupaciones y necesidades relacionadas con la deserción de suscriptores.
- Requisitos y Limitaciones del Proyecto:
  Definir claramente los requisitos y las limitaciones del proyecto. Esto podría incluir restricciones presupuestarias, limitaciones de tiempo y disponibilidad de datos.
- Impacto de la Deserción:
  Evaluar el impacto financiero y operativo de la deserción de suscriptores en la empresa. Cuantificar las pérdidas y entender cómo la retención efectiva puede mejorar la rentabilidad.
- Datos Disponibles:
  Analizar la disponibilidad y la calidad de los datos relacionados con los suscriptores. Identificar las fuentes de datos relevantes y evaluar su idoneidad para el modelado de aprendizaje de máquina.
- Riesgos y Oportunidades:
  Identificar posibles riesgos y oportunidades asociados con el proyecto. Esto podría incluir riesgos técnicos, de implementación, así como oportunidades para mejorar otros aspectos del negocio.

  Esta fase sienta las bases para el desarrollo de un modelo efectivo de aprendizaje de máquina que aborde los desafíos específicos de la retención de suscriptores en este contexto particular.

2.1.2 Entendimiento de los Datos.

El entendimiento de los datos es una fase crucial en la metodología CRISP-DM, especialmente en el contexto de tu tesis de maestría sobre la identificación de suscriptores en riesgo de abandono en una empresa de medicina prepagada en Ecuador. Esta etapa se centra en adquirir un conocimiento profundo sobre los datos disponibles y establecer las bases para la preparación y el modelado efectivos [4].

- Recopilación de Datos:

  Se inicia identificando y recopilando todos los conjuntos de datos relevantes para la investigación. Incluirá datos demográficos de los suscriptores, historiales de transacciones, interacciones con el servicio y cualquier otra información que pueda tener impacto en la deserción.

  En este caso hay varias fuentes de varias áreas entre ellas las de financiero y de atención al cliente.

- Exploración de Datos:

  Se realiza un análisis exploratorio detallado para comprender la estructura de los datos. Examinar la distribución de variables, identifica posibles anomalías, y comprende la naturaleza de las relaciones entre diferentes atributos.

  Se identifica las bases de cobranzas, bases de reclamos, intenciones de desafiliación y movimientos de deserción las cuales se encuentran estructuradas en un datawarehouse y también en un datalake, dichas estructuras podemos utilizarlas sin problema ya que se encuentran identificadas sus primary keys y foreing keys.

  En la Tabla 1 se presenta un resumen exploratorio de los ingresos y usuarios, lo que proporciona una visión general de la situación financiera y del tamaño de la base de usuarios. Este análisis resumido es esencial para comprender la salud financiera de la empresa y para evaluar el alcance de su base de clientes.

  El resumen de ingresos en la Tabla 1 permite identificar la cantidad total de ingresos generados durante un período específico, así como el promedio mensual de ingresos. Esto proporciona una medida de la estabilidad y el crecimiento de los ingresos a lo largo del tiempo, lo que es fundamental para evaluar la viabilidad financiera de la empresa.

  Por otro lado, el resumen de usuarios en la Tabla 1 muestra la cantidad total de usuarios registrados en la plataforma, así como el promedio mensual de nuevos

usuarios. Este análisis proporciona información sobre el tamaño y el crecimiento de la base de usuarios, lo que es esencial para evaluar el potencial de mercado de la empresa y su capacidad para atraer y retener clientes.

Tabla 1. Resumen de ingresos y usuarios

Año 2022 – Sucursales	Prima	Expuestos	Contratos
Sucursal 1	$937,566.18	3,644	750
Sucursal 2	$19,104,782.58	67,744	7025
Sucursal 3	$68,663,331.16	183,372	50099
Totales	$88,705,679.92	252,774	57873

Es decir, la mayoría de data a utilizar se encuentra en repositorios específicos de base de datos.

- Calidad de los Datos:

  Evaluación de la calidad de los datos. Identificar y trata los posibles problemas, como valores atípicos, datos faltantes, o inconsistencias que podrían afectar la validez de tus resultados [1][4][6].

  Se encontraron varias inconsistencias sin embargo se lograron corregir trayendo la data directa del transaccional según las necesidades de análisis de variables.

  Dentro de las inconsistencias encontradas son valores de prima, muchas de estas inconsistencias las arreglamos tomando varios campos que están disponibles en el transaccional, recargarlos y almacenarlos en el datawarehouse o en este caso en el datalake.

  Debemos entender que la prima fijada o PVP variara si se aplicó alguna promoción en el momento de compra o si hubo inclusiones o exclusiones

  De igual manera se realizaron los tratamientos de duplicados, tratamiento de valores nulos

  Además, para validaciones comparamos los resultados de los exploratorios a la par con la herramienta de Inteligencia Empresarial, en la cual tenemos a disposición los resultados e indicadores

- Entendimiento de Variables Relevantes:

  Analiza la relevancia de cada variable en relación con el objetivo de tu tesis. Identifica las variables que probablemente influyan más en la deserción de suscriptores y comprende su impacto en el modelo.

Se realiza un análisis previo de la variable más importante en este caso la cancelación de contratos en un periodo de contrato.

Para comprender más a fondo los datos relacionados con la desafiliación de los clientes, se llevó a cabo un análisis exploratorio de los motivos de desafiliación, cuyos resultados se presentan en la Figura 6. Este análisis es fundamental para identificar las razones detrás de la decisión de los clientes de cancelar sus servicios, lo que proporciona información valiosa para la retención de clientes y la mejora de la experiencia del usuario.

En la Figura 6, se muestra una visualización de los motivos más comunes de desafiliación de los clientes, lo que permite identificar patrones o tendencias en las razones por las cuales los clientes deciden cancelar sus servicios. Estos motivos pueden incluir factores como problemas de servicio, insatisfacción con el producto, cambios en las circunstancias personales del cliente, entre otros.

El análisis de los motivos de desafiliación en la Figura 6 proporciona una comprensión más profunda de las necesidades y expectativas de los clientes, lo que puede utilizarse para implementar estrategias de retención específicas orientadas a abordar los problemas identificados y mejorar la satisfacción del cliente. Además, este análisis permite identificar áreas de mejora en los productos o servicios ofrecidos, lo que puede contribuir a la retención de clientes a largo plazo y al crecimiento del negocio.

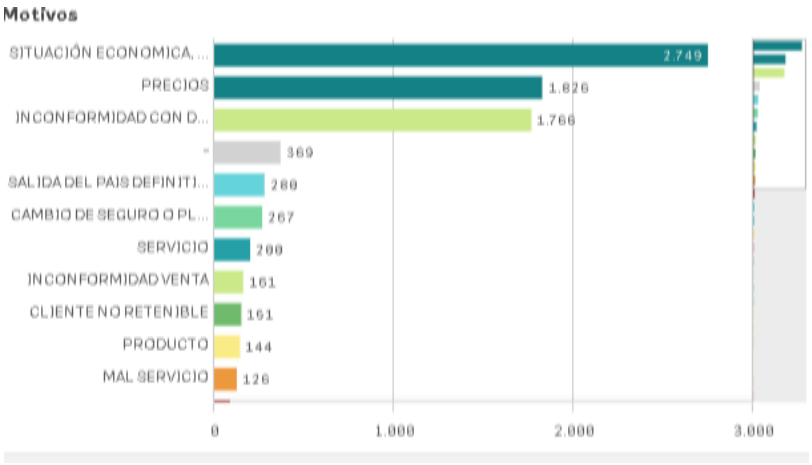


Figura 6. Exploración Motivos de Desafiliación

- Tipos de Datos:

  Se clasifica los tipos de datos presentes en los conjuntos de datos (categóricos, numéricos, etc.). Esto es crucial para determinar las técnicas de modelado apropiadas durante las fases posteriores del proyecto.

  Se identifica los tipos de datos se realiza el tratamiento respectivo.

- Exploración Temporal (si aplica):

  Se analiza datos temporales, explorar las tendencias a lo largo del tiempo. Comprender patrones estacionales o cambios a lo largo de diferentes periodos puede ser esencial para la predicción de la deserción.

  En la Figura 7 se lleva a cabo un análisis exploratorio de la desafiliación de contratos por mes, lo que proporciona una visión detallada de cómo varía el número de contratos cancelados o desafiliados a lo largo del tiempo. Este análisis es crucial para comprender la dinámica de la desafiliación y para identificar posibles patrones o tendencias en la fluctuación de la tasa de desafiliación a lo largo de los meses.

  El examen de la desafiliación de contratos por mes en la Figura 7 permite detectar períodos de aumento o disminución en la tasa de cancelación, así como identificar posibles factores o eventos que puedan estar influyendo en el comportamiento de los clientes. Este análisis exploratorio proporciona una base sólida para el diseño de estrategias de retención de clientes y para la implementación de acciones preventivas orientadas a reducir la desafiliación y mejorar la satisfacción del cliente.

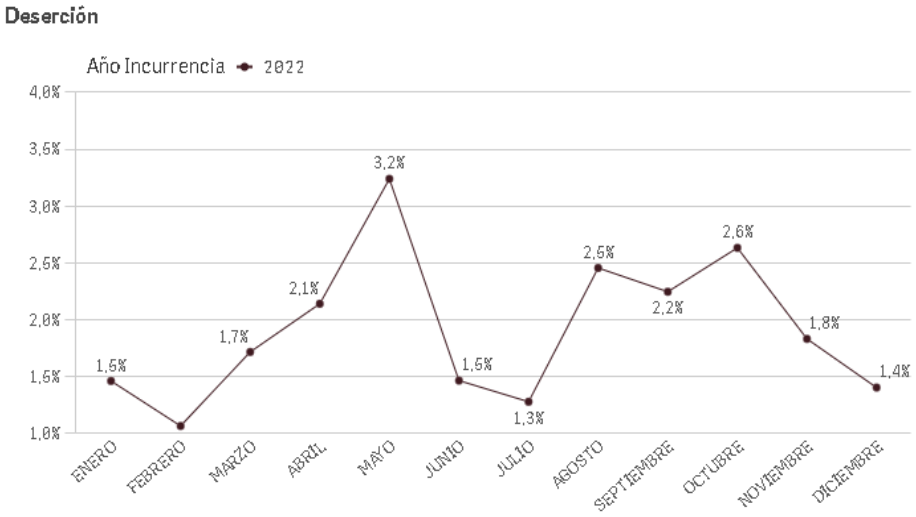


Figura 7. Exploración de desafiliación de contratos

Por otro lado, en la Tabla 2 se lleva a cabo un muestreo de la desafiliación por cada mes de venta, lo que implica la selección aleatoria de una muestra representativa de los contratos desafiliados para su análisis detallado. Este muestreo proporciona información específica sobre los contratos cancelados en cada mes, lo que puede ser útil para identificar patrones o tendencias en los motivos de la desafiliación, así como para comprender mejor las características de los clientes que deciden cancelar sus contratos en un momento determinado.

Tabla 2. Movimientos de contratos desde su inicio de suscripción

MES		1	2	3	4	5	6	7	8	9	10	11	12	13	14
202210	433	428	419	413	392	376	359	336	317	299	288	274	253	226	219
202211	375	374	366	357	343	330	313	291	279	268	259	251	231	201	198
202212	453	448	439	430	413	391	372	352	334	319	306	293	279	254	243
202301	481	481	476	460	446	426	395	376	369	351	342	335	320	283	272
202302	406	405	397	387	365	342	321	309	298	282	272	264	254	223	-
202303	468	467	459	454	438	414	393	372	351	339	325	310	289	-	-
202304	432	431	426	418	403	377	356	344	339	327	308	299	-	-	-
202305	496	496	486	476	462	444	418	401	387	374	363	-	-	-	-
202306	528	526	515	500	476	459	437	413	387	378	-	-	-	-	-
202307	539	538	527	517	499	476	454	437	423	-	-	-	-	-	-

2.1.3 Preparación de los datos

En la fase de preparación de datos en la metodología CRISP-DM es esencial para garantizar que los datos sean aptos para el modelado de aprendizaje de máquina. En el contexto de tu tesis de maestría sobre la identificación de suscriptores en riesgo de abandono en una empresa de medicina prepagada en Ecuador, esta etapa se centra en limpiar y transformar los datos de manera que puedan ser efectivamente utilizados en la construcción del modelo predictivo [6] [9].

- Limpieza de datos: Detección y eliminación de valores nulos: Se analizan los datos para identificar valores faltantes y se eliminan aquellos que no se pueden imputar con precisión. Se pueden utilizar diferentes técnicas para la detección de valores nulos, como:
- Análisis estadístico: Se calcula la frecuencia de valores nulos en cada variable y se eliminan las variables con un porcentaje elevado de valores nulos.
- Corrección de valores inconsistentes: Se revisan los datos para detectar valores inconsistentes o erróneos y se corrigen de acuerdo a la información disponible. Se pueden utilizar diferentes técnicas para la corrección de valores inconsistentes, como:
- Verificación de rangos: Se establecen rangos válidos para cada variable y se corrigen los valores que se encuentran fuera de estos rangos.
- Imputación por valores cercanos: Se reemplazan los valores inconsistentes por el valor más cercano dentro de un rango válido.
- Eliminación de duplicados: Se identifican y eliminan registros duplicados que puedan afectar la precisión del modelo. Se pueden utilizar diferentes técnicas para la detección de duplicados, como:
- Comparación de claves únicas: Se comparan las claves únicas de cada registro para identificar duplicados.

  Manejo de datos faltantes:
- Imputación por promedio: Se utiliza la media de la variable para reemplazar los valores faltantes cuando la distribución de la variable lo permite. Esta técnica es simple y rápida, pero puede ser poco precisa si la distribución de la variable no es normal.
- Imputación por regresión: Se utiliza un modelo de regresión para imputar los

valores faltantes a partir de otras variables. Esta técnica es más precisa que la imputación por promedio, pero requiere la selección de variables predictoras adecuadas.

- Imputación por k-nearest neighbors: Se busca la k observaciones más similares al registro con valores faltantes y se utiliza la media de las k variables para imputar los valores faltantes. Esta técnica es más robusta que la imputación por promedio o por regresión, pero puede ser más computacionalmente costosa.
- Detección y tratamiento de valores atípicos:
- Detección por métodos estadísticos: Se utilizan métodos estadísticos como la prueba de Grubbs y la prueba de Dixon para identificar valores atípicos. Estos métodos se basan en la comparación de los valores de cada variable con la distribución de la variable.
- Detección por visualización: Se utilizan gráficos como boxplots y histogramas para identificar valores atípicos. Estos gráficos permiten identificar valores que se encuentran significativamente alejados del resto de la distribución.
- Eliminación o transformación de valores atípicos: Se eliminan los valores atípicos que puedan afectar significativamente el modelo o se transforman utilizando técnicas como la winsorización. La eliminación de valores atípicos puede ser una solución simple, pero puede afectar la precisión del modelo si se eliminan demasiados datos. La winsorización es una técnica que transforma los valores atípicos a un valor límite dentro de la distribución.
- Transformación de variables:
- Normalización: Se normalizan las variables numéricas para que tengan una media de 0 y una desviación estándar de 1. La normalización es una técnica que permite comparar variables con diferentes unidades de medida.
- Estandarización: Se estandarizan las variables numéricas para que tengan una media de 0 y una desviación estándar de 1. La estandarización es una técnica similar a la normalización, pero se utiliza para variables que no tienen una distribución normal.
- Codificación de variables categóricas: Se codifican las variables categóricas utilizando técnicas como la codificación one-hot o la codificación ordinal. La codificación de variables categóricas es necesaria para que los algoritmos de

aprendizaje automático puedan procesar estas variables.

- Creación de nuevas variables: Se crean nuevas variables a partir de las variables existentes para mejorar la capacidad predictiva del modelo. La creación de nuevas variables puede ser una forma de obtener información adicional a partir de los datos disponibles.

La preparación de datos es un proceso iterativo y requiere un equilibrio entre la eficacia y la preservación de la integridad de los datos. Una vez completada esta fase, los datos están listos para ser utilizados en la construcción y evaluación de modelos de aprendizaje de máquina en la siguiente etapa del ciclo CRISP-DM.

2.1.4 Modelamiento

Esta fase implica seleccionar y entrenar modelos que puedan generalizar patrones a partir de los datos disponibles [12] [15].

- Selección de algoritmos:

  Análisis de las características de los datos: Se analiza la distribución de las variables, la presencia de valores atípicos y la correlación entre variables.

  Se seleccionan algoritmos que sean apropiados para un problema de clasificación binaria.

  Algunos algoritmos que se pueden considerar:

  Regresión logística: Un algoritmo clásico para la clasificación binaria, simple de interpretar y con buen rendimiento en general.

  Máquinas de Soporte Vectorial: Un algoritmo que busca encontrar el hiperplano que mejor separa las dos clases, con capacidad para manejar problemas no lineales.

  Bosques Aleatorios: Un conjunto de árboles de decisión que ofrece robustez y flexibilidad, con capacidad para manejar variables categóricas y numéricas.

  Redes Neuronales Artificiales: Un modelo inspirado en el funcionamiento del cerebro humano, capaz de aprender patrones complejos y no lineales.

- Configuración de parámetros:

  Ajuste manual de parámetros: Se ajustan manualmente los parámetros de cada algoritmo, como la tasa de aprendizaje, el número de iteraciones, el tamaño del lote y la regularización.

Se comienza con valores predeterminados para los parámetros y luego se ajustan de forma incremental para observar su impacto en el rendimiento del modelo.

Se utilizan métricas de evaluación como la precisión, la sensibilidad, la especificidad y el AUC-ROC para evaluar el rendimiento del modelo con diferentes configuraciones de parámetros.

Es importante realizar un ajuste manual de los parámetros para comprender mejor su impacto en el modelo y encontrar una configuración que funcione bien para el conjunto de datos específicos.

- Búsqueda de hiperparámetros:

  Se utilizan técnicas de búsqueda automatizada para encontrar la configuración de parámetros óptima.

  La búsqueda en rejilla evalúa todas las combinaciones posibles de valores de parámetros dentro de un rango predefinido.

  La búsqueda aleatoria explora el espacio de parámetros de forma aleatoria y utiliza técnicas como el algoritmo de Metropolis-Hastings para seleccionar la siguiente configuración de parámetros.

  Existen bibliotecas de software como Scikit-optimize y Optuna que facilitan la implementación de técnicas de búsqueda de hiperparámetros.

  La búsqueda de hiperparámetros puede ser computacionalmente costosa, especialmente para conjuntos de datos grandes o algoritmos complejos.

- Entrenamiento del modelo:

  División de los datos:

  Se divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.

  El conjunto de entrenamiento se utiliza para entrenar el modelo y ajustar los parámetros. El conjunto de prueba se utiliza para evaluar el rendimiento del modelo final, evitando el sobreajuste.

  La proporción de datos asignada a cada conjunto puede variar, pero una división común es 80% para entrenamiento y 20% para prueba.

  Entrenamiento del modelo en el conjunto de entrenamiento:

  Se utiliza el conjunto de entrenamiento para ajustar los parámetros del modelo y aprender los patrones de los datos.

El proceso de entrenamiento puede ser iterativo, con ajustes en la configuración de parámetros y la arquitectura del modelo para optimizar el rendimiento. Se utilizan técnicas como el descenso del gradiente para minimizar la función de pérdida y optimizar los parámetros del modelo. El tiempo de entrenamiento puede variar dependiendo del tamaño del conjunto de datos, la complejidad del modelo y la potencia de procesamiento disponible.

- Validación cruzada:

  Implementación de técnicas de validación cruzada:

  Se utilizan técnicas como la validación cruzada k-fold o la validación cruzada estratificada para evaluar la capacidad de generalización del modelo.

  La validación cruzada k-fold divide el conjunto de entrenamiento en k subconjuntos y entrena el modelo k veces, utilizando cada subconjunto como conjunto de prueba una vez.

  La validación cruzada estratificada preserva la distribución de las clases en cada subconjunto para evitar sesgos en la evaluación del modelo.

- Selección del mejor modelo:

  Se selecciona el modelo con el mejor rendimiento en la validación cruzada.

  Se pueden comparar diferentes algoritmos, configuraciones de parámetros y arquitecturas de modelo para seleccionar la mejor opción para el problema específico.

  Es importante considerar no solo la precisión del modelo, sino también su capacidad de generalizar a datos no vistos. En la investigación las prueba la resultante es el modelo XGBoost, ampliando un poco más esta opción se detalla lo siguiente:

  XGBoost es una herramienta poderosa para el modelado de aprendizaje automático que ofrece una serie de ventajas sobre otros algoritmos. Su precisión, velocidad, escalabilidad y flexibilidad lo convierten en una opción atractiva para una amplia gama de problemas. Sin embargo, es importante tener en cuenta su complejidad y sensibilidad a los parámetros para ajustar el modelo correctamente y obtener un buen rendimiento.

- Combina los siguientes enfoques:

  Árboles de decisión: Un árbol de decisión es un modelo de aprendizaje automático que utiliza una estructura jerárquica para clasificar o predecir valores. El árbol se compone de nodos internos, que representan preguntas o decisiones, y nodos terminales, que representan las diferentes clases o valores de la variable objetivo.

  Boosting: El boosting es un conjunto de algoritmos que combinan múltiples modelos débiles para crear un modelo fuerte con mejor rendimiento. XGBoost utiliza un tipo de boosting llamado "gradient boosting", que se basa en la idea de agregar árboles de decisión de forma iterativa.

  Función de pérdida: La función de pérdida es una medida del error del modelo. XGBoost utiliza la función de pérdida de regresión logística para la clasificación y la función de pérdida de error cuadrático medio para la regresión.

  Regularización:La regularización es una técnica que se utiliza para evitar el sobreajuste del modelo. XGBoost utiliza dos tipos de regularización:

  Regularización L1: Penaliza la suma de los valores absolutos de los coeficientes del modelo.

  Regularización L2: Penaliza la suma de los cuadrados de los coeficientes del modelo.

- Ecuaciones de XGBoost:

  La ecuación principal de XGBoost es la siguiente:

$$Fm(x) = Fm - 1(x) + \eta \sum j = 1 Jmhj(x)$$

Ecuacion 1[14]

  Donde:

  - Fm(x): Función de predicción del modelo en la iteración m.
  - Fm-1(x): Función de predicción del modelo en la iteración m-1.
  - η: Tasa de aprendizaje.
  - Jm: Número de árboles en la iteración m.
  - hj(x): Función de predicción del árbol j en la iteración m.
  - La función de predicción de cada árbol se calcula de la siguiente manera:

$$hj(x) = wj \sum i = 1 njI(xi \in Rij)$$

Ecuacion 2[14]

Donde:

- wj: Peso del árbol j.
- nj: Número de nodos terminales en el árbol j.
- Rij: Región del espacio de características que corresponde al nodo terminal i del árbol j.
- I($xi \in Rij$): Indicador que toma el valor 1 si la instancia x_i pertenece a la región Rij y 0 si no 1

Error de propagación en XGBoost:

En XGBoost, el error de propagación se refiere a la diferencia entre la predicción del modelo en una iteración y la predicción del modelo en la iteración anterior. Se utiliza para calcular la ganancia de cada árbol en el proceso de boosting.

Cálculo del error de propagación:

El error de propagación para una instancia i en la iteración m se calcula de la siguiente manera:

$$ei(m) = yi - Fm - 1(xi)$$

Ecuacion 3[14]

Donde:

- ei(m): Error de propagación para la instancia i en la iteración m.
- yi: Valor objetivo de la instancia i.
- Fm-1(xi): Función de predicción del modelo en la iteración m-1.

Ganancia de un árbol:

La ganancia de un árbol se define como la reducción del error de propagación que se logra al agregar el árbol al modelo. Se calcula de la siguiente manera:

$$Gain = \sum i = 1nei(m)hj(xi) - 21\lambda\sum j = 1nwj2$$

Ecuacion 4[14]

Donde:

- Gain: Ganancia del árbol j.
- n: Número de instancias en el conjunto de entrenamiento.
- λ: Parámetro de regularización L2.

- Selección del árbol:

  El árbol con la mayor ganancia se selecciona para agregarse al modelo en cada iteración.

  El error de propagación se calcula de forma eficiente en XGBoost utilizando técnicas como la suma de cuadrados parciales.
- Optimización:

  El objetivo de XGBoost es encontrar los valores de los parámetros que minimicen la función de pérdida regularizada. Esto se realiza mediante un algoritmo de descenso del gradiente.
- Aplicaciones de XGBoost:

  Recomendación de productos: XGBoost se puede usar para recomendar productos a los usuarios en función de su comportamiento de compra.

  Detección de fraude: XGBoost se puede usar para detectar transacciones fraudulentas en sistemas financieros.

  Análisis de riesgo crediticio: XGBoost se puede usar para predecir la probabilidad de que un cliente incumpla un préstamo.

  Predicción de mantenimiento: XGBoost se puede usar para predecir cuándo fallará un componente de una máquina para programar el mantenimiento preventivo.
- Ventajas de XGBoost:

  Precisión: XGBoost generalmente ofrece una mayor precisión que otros algoritmos como Random Forest o Support Vector Machines, especialmente en conjuntos de datos complejos.

  Velocidad: XGBoost es un algoritmo rápido, tanto en entrenamiento como en predicción, lo que lo hace ideal para aplicaciones donde la velocidad es importante.

  Escalabilidad: XGBoost puede manejar conjuntos de datos grandes con millones de ejemplos sin problemas de rendimiento.

  Flexibilidad: XGBoost ofrece una variedad de parámetros que pueden ajustarse para optimizar el rendimiento para diferentes tipos de problemas.

  Regularización: XGBoost incluye técnicas de regularización para evitar el sobreajuste y mejorar la capacidad de generalización del modelo.

Interpretabilidad: XGBoost ofrece una buena interpretabilidad, permitiendo visualizar la importancia de las variables y cómo el modelo toma decisiones.

- Desventajas de XGBoost:

  Complejidad: XGBoost tiene una configuración más compleja que otros algoritmos, lo que puede requerir más tiempo y esfuerzo para encontrar la configuración óptima.

  Sensibilidad a los parámetros: XGBoost puede ser sensible a la configuración de los parámetros, lo que puede afectar significativamente el rendimiento del modelo.

  Consumo de memoria: XGBoost puede consumir mucha memoria durante el entrenamiento, especialmente con conjuntos de datos grandes.

- Ajuste del modelo:

  Optimización del rendimiento: Se realizan ajustes al modelo, como la selección de variables, la configuración de parámetros o la prueba de diferentes algoritmos, para mejorar su rendimiento.

- Evaluación del modelo:

  Cálculo de métricas de evaluación: Se calculan métricas como la precisión, la sensibilidad, la especificidad y el AUC-ROC para evaluar el rendimiento del modelo en el conjunto de prueba.

  Análisis de la matriz de confusión: Se analiza la matriz de confusión para comprender mejor los errores del modelo.

- Interpretación de resultados:

  Análisis de las características más influyentes: Se identifican las variables que tienen mayor impacto en las predicciones del modelo.

- Optimización del modelo:

  Refinamiento del modelo: Se realizan ajustes adicionales al modelo, como la optimización de hiperparámetros o la exploración de diferentes enfoques de modelado, para obtener el mejor rendimiento posible.

- La fase de modelado es iterativa y puede requerir ajustes para lograr un equilibrio óptimo entre la complejidad del modelo y su capacidad de generalización. Una vez completada esta etapa, estarás listo para pasar a la fase de evaluación y tomar decisiones basadas en los resultados obtenidos.

2.1.5 Evaluación

La evaluación del modelo se llevará a cabo en dos aspectos principales:

Impacto en la retención de suscriptores:
Se compararán los resultados obtenidos por el modelo de predicción con la situación previa a su implementación. Se calcularán métricas de retención, como la tasa de cancelación de suscripciones y la duración promedio de las suscripciones. Estas métricas se analizarán en un período de tiempo apropiado, lo que permitirá evaluar el impacto del modelo en la capacidad de la empresa para retener a sus suscriptores.

Rentabilidad de la empresa:
Se calculará el retorno de inversión (ROI) del modelo, considerando los costos asociados con su desarrollo, implementación y mantenimiento, frente a los beneficios derivados de la reducción de la cancelación de suscripciones y el aumento de la retención de clientes. Además, se evaluará si el modelo contribuye a la generación de ingresos adicionales mediante la identificación de oportunidades de retención y upselling.

2.1.6 Despliegue

El despliegue del modelo se realizará en un entorno de producción para su uso real. A continuación, se detallan los pasos específicos para su implementación:

Implementación en entorno de producción:
Se integrará el modelo de predicción en el sistema de gestión de la empresa de medicina prepaga. Para ello, se desarrollarán scripts o servicios que permitan ejecutar el modelo de forma automatizada en intervalos regulares, de acuerdo con el procesamiento batch mensual mencionado. Este proceso garantizará que el modelo esté siempre actualizado con los datos más recientes disponibles.

Inserción de resultados en la base de datos:
Los resultados de las predicciones se insertarán en la base de datos SQL de la empresa, asociados a cada contrato de suscriptor mediante su identificador único (ID contrato). Estos resultados incluirán la probabilidad de abandono calculada por el modelo, así como cualquier otra información relevante para la toma de decisiones posteriores.

Seguimiento y ajuste:
Se establecerá un proceso de monitoreo continuo para evaluar el rendimiento del modelo en producción. Esto implicará la revisión periódica de las métricas de retención y rentabilidad, así como el análisis de posibles desviaciones en el comportamiento de los suscriptores. En función de estos resultados, se realizarán ajustes al modelo, como la inclusión de nuevas variables predictoras o la modificación de los umbrales de clasificación, para mejorar su precisión y efectividad en la identificación de suscriptores en riesgo de abandono.

Con estas acciones de evaluación y despliegue, se garantizará que el modelo de predicción de abandono de suscriptores contribuya de manera efectiva a la retención y rentabilidad de la empresa de medicina prepaga del Ecuador, permitiendo una gestión proactiva de la deserción de clientes y una optimización de los recursos disponibles.

3. RESULTADOS

Este proceso de desarrollo se caracteriza por su profundidad y amplitud, proporcionando una descripción detallada de cada etapa del proceso, desde la carga y preprocesamiento de datos hasta la interpretación de los resultados. En este caso tenemos disponibilidad de información desde el 2017 al 2022 dado que esa es la disponibilidad de la estructura de datos principal en este caso nos apoyaremos de las transacciones realizadas de cobros y la de siniestros.

En el ámbito empresarial, la gestión de datos ha experimentado una evolución significativa en los últimos años. Antes del año 2017, la falta de un registro adecuado de datos dificultaba la comprensión y el análisis de diversas variables clave en el negocio. Para abordar esta problemática, nos hemos propuesto recopilar datos demográficos y características relevantes que nos permitan definir las variables principales para la construcción de un modelo sólido.

Nos apoyaremos en sistemas de registro de quejas y atención al cliente más modernos, los cuales proporcionarán un historial detallado de suscripciones y comportamientos de los clientes. Este historial incluirá información sobre el costo inicial de suscripción, así como las variaciones de precios en cada renovación anual. Este enfoque nos brindará una visión integral del comportamiento del cliente a lo largo del tiempo y nos ayudará a identificar patrones y tendencias significativas. Además, nos enfrentamos al desafío de comprender el comportamiento de uso de un servicio de medicina prepagada, especialmente en lo que respecta a las condiciones de uso y limitaciones del mismo. La información recopilada incluirá datos extensos sobre siniestros, tales como pagos realizados y conceptos de servicios cubiertos. Este conjunto de datos complejo y detallado nos permitirá analizar el comportamiento de los usuarios en términos de transacciones y uso del servicio, así como identificar tendencias emergentes.

Es importante destacar que esta abundante cantidad de datos nos proporcionará insights valiosos sobre diversos aspectos, como el número de afiliados, su distribución por sexo y edad, los diagnósticos más comunes, así como los diferentes tipos de segmentaciones de mercado. Este análisis nos permitirá detectar comportamientos favorables y oportunidades de mejora, lo que nos ayudará a tomar decisiones más informadas y estratégicas para el crecimiento y desarrollo del negocio.

3.1 Proceso de Carga de Datos

La carga de datos es un paso fundamental en el proceso de análisis y modelado de información en cualquier proyecto empresarial, y en particular, en el contexto de un servicio de medicina prepagada. En este caso, se manejan dos conjuntos de datos principales: el conjunto de entrenamiento y el conjunto de prueba, cada uno de los cuales proporciona una perspectiva única pero complementaria sobre el comportamiento de los suscriptores.

Ambos conjuntos de datos están meticulosamente diseñados para capturar una amplia gama de información relevante sobre los suscriptores del servicio de medicina prepagada. Esto incluye datos demográficos esenciales como la edad y el género, que nos permiten comprender mejor el perfil de la población de suscriptores y adaptar nuestras estrategias en consecuencia. Por ejemplo, la distribución por edad y género puede influir en la demanda de servicios de salud específicos, lo que a su vez impacta en la planificación de recursos y la segmentación del mercado.

Además de los datos demográficos, los conjuntos de datos también contienen información detallada sobre el plan de medicina prepagada suscrito por cada individuo. Esto puede incluir detalles sobre la cobertura de servicios médicos, los límites de gastos, las exclusiones y otras condiciones específicas del plan. Entender las características de los diferentes planes disponibles es crucial para evaluar la satisfacción del cliente, identificar áreas de mejora en la oferta de servicios y optimizar la rentabilidad del negocio.

El historial de pagos y la utilización de servicios son aspectos críticos que se registran en los conjuntos de datos. Estos datos proporcionan una visión en profundidad del comportamiento financiero y de uso de los suscriptores a lo largo del tiempo. Por ejemplo, el análisis de los patrones de pago puede revelar tendencias de morosidad o fluctuaciones en la capacidad de pago de los suscriptores, lo que puede requerir intervenciones específicas para mitigar el riesgo de cancelación de la suscripción.

Además, la información detallada sobre la utilización de servicios médicos, como las visitas al médico, los procedimientos realizados y los medicamentos recetados, ofrece una perspectiva valiosa sobre la salud y el bienestar de los suscriptores. Esto puede ayudar a identificar áreas de atención prioritaria, detectar patrones de enfermedades crónicas o emergentes, y diseñar programas de promoción de la salud y prevención de enfermedades personalizados.

Finalmente, el motivo de cancelación para los suscriptores inactivos se registra meticulosamente en los conjuntos de datos. Comprender las razones por las cuales los suscriptores eligen cancelar su membresía es esencial para abordar las preocupaciones y necesidades subyacentes de los clientes, así como para implementar estrategias efectivas de retención y fidelización.

Descripción Detallada de los Conjuntos de Datos

1. Datos Demográficos:

Edad: Se registra la edad de cada suscriptor en años. Esta información es crucial para comprender la etapa de vida del cliente y su potencial demanda de servicios médicos.
Género: Se indica el género del suscriptor (masculino, femenino o no especificado). Esta variable puede influir en las preferencias de planes de salud y el comportamiento de uso de servicios.

2. Información del Plan de Medicina Prepagada:

Tipo de plan: Se identifica el tipo de plan de salud al que está afiliado el suscriptor (individual, familiar, empresarial, etc.). Esta variable es fundamental para comprender el nivel de cobertura, los beneficios y los costos asociados al plan.
Fecha de afiliación: Se registra la fecha en la que el suscriptor se afilió al plan de salud. Esta información permite analizar la antigüedad del cliente y su historial de relación con la empresa.
Estado del plan: Se indica si el plan está activo o inactivo. En caso de estar inactivo, se registra la fecha de cancelación.

3. Historial de Pagos:

Puntualidad en los pagos: Se registra si el suscriptor ha realizado sus pagos de manera puntual o si ha presentado atrasos. Esta información es un indicador importante de la satisfacción del cliente y su compromiso con el plan.
Historial de morosidad: Se registra si el suscriptor ha tenido episodios de morosidad en el pago de sus cuotas. Esta variable puede estar relacionada con problemas financieros o insatisfacción con el servicio.
Métodos de pago utilizados: Se indica el método de pago preferido por el suscriptor (efectivo, tarjeta de débito, tarjeta de crédito, débito automático, etc.). Esta información puede ser útil para comprender las preferencias del cliente y optimizar los procesos de cobro.

4. Utilización de Servicios:

Frecuencia de uso de servicios médicos: Se registra la frecuencia con la que el suscriptor utiliza servicios médicos, como consultas médicas, hospitalizaciones, exámenes de laboratorio, etc. Esta información es un indicador importante de la necesidad de atención médica y el valor percibido del plan.
Tipo de servicios utilizados: Se detalla el tipo de servicios médicos que ha utilizado el suscriptor, como consultas generales, especialidades médicas, procedimientos ambulatorios, hospitalizaciones, etc. Esta información permite identificar patrones de uso y necesidades específicas.
Costo asociado a cada servicio: Se registra el costo de cada servicio médico utilizado por el suscriptor. Esta información permite analizar la rentabilidad del cliente y su potencial de consumo de servicios.

5. Motivo de Cancelación (solo para suscriptores inactivos):

En el caso de los suscriptores que han cancelado su plan de salud, se registra la razón principal por la que decidieron cancelar el servicio. Esta información puede ser muy valiosa para identificar áreas de mejora y desarrollar estrategias de retención efectivas. Algunos ejemplos de motivos de cancelación podrían ser:

Costo del plan: El costo del plan excede la capacidad de pago del suscriptor.
Insatisfacción con la cobertura del plan: El plan no cubre las necesidades específicas del suscriptor.
Mala calidad del servicio: El suscriptor ha tenido experiencias negativas con la atención médica o el servicio al cliente.
Falta de acceso a los servicios: El suscriptor no tiene acceso fácil a los centros de atención médica o los servicios que necesita.
Cambio de compañía de seguros: El suscriptor ha encontrado un plan de salud que mejor se adapta a sus necesidades y presupuesto.

Análisis Combinado de Datos: La combinación de estos datos demográficos, información del plan, historial de pagos, utilización de servicios y motivos de cancelación permite obtener una visión integral del comportamiento del suscriptor y los factores que influyen en su satisfacción y lealtad a la empresa. Esta información es fundamental para desarrollar estrategias de retención efectivas, optimizar la rentabilidad del negocio y mejorar la calidad del servicio de salud.

Consideraciones Importantes: La calidad y precisión de los datos son cruciales para obtener resultados confiables en el análisis y la construcción de modelos predictivos.

Es importante proteger la privacidad de los datos de los suscriptores y cumplir con las regulaciones vigentes sobre protección de datos. La interpretación de los datos debe realizarse con cautela, considerando el contexto y las limitaciones de la información disponible.

En la Tabla 3 se presentan las variables relacionadas con los ingresos, lo que proporciona una visión detallada de los diferentes aspectos financieros que pueden influir en la situación económica de la empresa. Estas variables son fundamentales para comprender la salud financiera de la organización y para realizar análisis financieros más profundos.
Las variables de ingreso incluidas en la Tabla 3 pueden abarcar una variedad de aspectos, como los ingresos totales generados por la empresa durante un período específico, los ingresos por cliente o usuario, los ingresos por producto o servicio, los ingresos recurrentes o periódicos, entre otros. Estas variables pueden proporcionar información

valiosa sobre la estabilidad, el crecimiento y la rentabilidad de la empresa, lo que es esencial para la toma de decisiones financieras estratégicas.

El análisis de las variables de ingreso en la Tabla 3 permite identificar tendencias, patrones y relaciones entre diferentes aspectos financieros, lo que puede utilizarse para mejorar la eficiencia operativa, optimizar la asignación de recursos y desarrollar estrategias de crecimiento sostenible. Además, esta información puede ser útil para comunicar el desempeño financiero de la empresa a inversores, accionistas y otras partes interesadas clave.

La disponibilidad de datos provenientes de sistemas de registro recientes y detallados proporciona una oportunidad invaluable para comprender mejor las dinámicas del negocio de medicina prepagada. Estos datos se dividen en dos categorías principales: datos de quejas de clientes y datos de suscripciones y comportamiento. Cada categoría ofrece una perspectiva única que contribuye significativamente a la toma de decisiones estratégicas y a la mejora continua de los servicios ofrecidos.

Datos de sistemas de registro recientes: Quejas de clientes: Este conjunto de datos brinda una visión detallada de las áreas de insatisfacción de los clientes y los posibles motivos de deserción. Al analizar las quejas de los clientes, se pueden identificar patrones recurrentes, áreas de mejora y puntos de dolor críticos que requieren atención inmediata. Además, estas quejas sirven como una alerta temprana sobre problemas potenciales que podrían afectar la retención de clientes y la reputación de la empresa.

Tabla 3. Principales Variables de Ingresos.

Negocio	Producto Principal	Codigo Broker	Contrato	Estado Contrato	Fecha Cancelacion	Tipo Facturacion Contrato	Fecha Generacion	Forma Pago Contrato	Año Generacion	Mes Generacion	Expuestos	Primas
NEGOCIO INDIVIDUAL	MPI	20	305760	CANCELADO	7/2/2023	MENSUAL	1/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	1	$62.30
NEGOCIO INDIVIDUAL	MPI	966714	303651	RENOVADO	-	MENSUAL	22/12/2022	TARJETA CREDITO	2022	DICIEMBRE	1	$62.29
NEGOCIO INDIVIDUAL	MPI	1360867	291588	RENOVADO	-	MENSUAL	22/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	3	$239.46
NEGOCIO INDIVIDUAL	MPI	20	292706	RENOVADO	-	MENSUAL	22/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	1	$116.09
NEGOCIO INDIVIDUAL	MPI	20	291864	RENOVADO	-	MENSUAL	1/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	3	$140.30
NEGOCIO INDIVIDUAL	MPI	20	305152	RENOVADO	-	MENSUAL	14/12/2022	TARJETA CREDITO	2022	DICIEMBRE	1	$65.58
NEGOCIO INDIVIDUAL	B	1303779	281912	CANCELADO	10/1/2023	MENSUAL	1/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	2	$126.05
NEGOCIO INDIVIDUAL	MPI	20	305648	CANCELADO	1/7/2023	MENSUAL	1/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	2	$78.95
NEGOCIO INDIVIDUAL	MPI	20	307414	RENOVADO	-	MENSUAL	30/12/2022	TARJETA CREDITO	2022	DICIEMBRE	1	$71.53
NEGOCIO INDIVIDUAL	MPI	20	307414	RENOVADO	-	MENSUAL	30/12/2022	TARJETA CREDITO	2022	DICIEMBRE	1	$-3.58
NEGOCIO INDIVIDUAL	MPI	20	305095	CANCELADO	12/7/2023	MENSUAL	1/12/2022	DEBITO BANCARIO	2022	DICIEMBRE	1	$72.03

b. Interacciones con atención al cliente: La información recopilada sobre las interacciones con el servicio de atención al cliente proporciona una comprensión profunda del comportamiento de los clientes. Estos datos permiten identificar tendencias, patrones de comportamiento y predicciones sobre la probabilidad de abandono. Además, ayudan a personalizar la experiencia del cliente y a mejorar la calidad del servicio al proporcionar una respuesta rápida y efectiva a las consultas y preocupaciones de los clientes.

Datos de suscripciones y comportamiento:

a. Historial de suscripciones: Este conjunto de datos ofrece una visión longitudinal del tiempo que los clientes han estado afiliados a la empresa. Al analizar el historial de suscripciones, se pueden identificar tendencias de retención, ciclos de vida del cliente y oportunidades para mejorar la lealtad del cliente a largo plazo. Además, estos datos son fundamentales para evaluar el impacto de las estrategias de retención de clientes y para ajustar las políticas de suscripción según sea necesario.

b. Comportamiento de uso de servicios: La frecuencia y el tipo de servicios utilizados por cada suscriptor son indicadores críticos de su nivel de compromiso y satisfacción con el servicio de medicina prepagada. Al analizar el comportamiento de uso de servicios, se pueden identificar patrones de consumo, áreas de interés y oportunidades para personalizar las ofertas de servicios según las necesidades individuales de los clientes. Esto no solo aumenta la satisfacción del cliente, sino que también optimiza la utilización de recursos y mejora la rentabilidad del negocio.

En la Tabla 4 se presentan las variables relacionadas con el uso de servicios y siniestros, lo que ofrece una visión detallada de cómo los clientes interactúan con los servicios ofrecidos y de los incidentes o reclamaciones asociados con dichos servicios. Estas variables son fundamentales para comprender el nivel de actividad y la calidad del servicio proporcionado por la empresa, así como para evaluar los riesgos y las necesidades de los clientes.

Las variables de uso de servicios pueden incluir información sobre la frecuencia y la duración de la utilización de los servicios por parte de los clientes, así como detalles

específicos sobre los tipos de servicios utilizados y los patrones de uso a lo largo del tiempo. Por otro lado, las variables de siniestros pueden abarcar datos sobre reclamaciones de seguros, eventos adversos o incidentes que hayan ocurrido durante la prestación de servicios.

El análisis de estas variables en la Tabla 4 permite identificar tendencias, patrones y relaciones entre el uso de servicios y la ocurrencia de siniestros, lo que puede utilizarse para mejorar la calidad y la eficiencia de los servicios ofrecidos, así como para desarrollar estrategias de gestión de riesgos más efectivas. Además, esta información puede ser útil para identificar áreas de mejora en los procesos operativos y para proporcionar una experiencia más satisfactoria a los clientes.

En conjunto, estos conjuntos de datos proporcionan una visión integral del negocio de medicina prepagada, permitiendo a las empresas tomar decisiones informadas y estratégicas para mejorar la experiencia del cliente, aumentar la retención y maximizar el valor del servicio ofrecido. La capacidad de aprovechar estos datos de manera efectiva es fundamental para mantener la competitividad en un mercado en constante evolución y satisfacer las crecientes expectativas de los clientes.

Tabla 4. Variables de Uso de Servicios y Siniestros

Contrato	Fecha Gene	Transito	Causa Externa	Código Diagnostico	Genero Afiliado	Tipo Cuadro	Vlr. Recorte	Vlr. Deducible	Vlr. Copago	Vlr. Pagado Neto	Costo por Caso (Pagado)	Vlr. Presentado
307080	27/12/2022	9443948	ENFERMEDAD GENERAL	J00X	FEMENINO	CERRADO	$0.00	$0.00	$0.00	$9.02	9.016	$12.88
307080	29/12/2022	9448073	ENFERMEDAD GENERAL	J00X	FEMENINO	CERRADO	$0.00	$0.00	$0.00	$19.64	19.6396	$26.20
307080	29/12/2022	9448105	ENFERMEDAD GENERAL	J00X	FEMENINO	CERRADO	$0.00	$0.00	$0.00	$19.28	19.2766	$26.54
306751	11/12/2022	9399443	ENFERMEDAD GENERAL	J042	FEMENINO	CERRADO	$0.00	$30.00	$0.00	$0.00	0	$30.00
306751	11/12/2022	9399443	ENFERMEDAD GENERAL	J042	FEMENINO	LIBRE ELECCION	$0.00	$0.00	$0.00	$0.00	0	$65.00
306751	15/12/2022	9417749	ENFERMEDAD GENERAL	J042	FEMENINO	LIBRE ELECCION	$0.00	$10.00	$0.00	$4.00	4	$15.00
306751	15/12/2022	9417749	ENFERMEDAD GENERAL	J042	FEMENINO	LIBRE ELECCION	$0.00	$0.00	$0.00	$0.00	0	$50.00
306751	11/12/2022	9399443	ENFERMEDAD GENERAL	J042	FEMENINO	LIBRE ELECCION	$0.00	$40.00	$0.00	$0.00	0	$40.00

La inclusión de datos de siniestros en el análisis de medicina prepagada representa una oportunidad significativa para comprender mejor las dinámicas del servicio y mejorar la calidad de la atención. Sin embargo, esta incorporación conlleva una serie de consideraciones importantes que deben abordarse para aprovechar al máximo el potencial de estos datos.

Volumen y complejidad de los datos:

El primer desafío al incorporar datos de siniestros radica en el volumen y la complejidad de la información. Los registros de siniestros pueden incluir una gran cantidad de variables, como pagos, conceptos de servicio, diagnósticos y procedimientos médicos. Además, estos datos pueden presentarse en diversos formatos y sistemas, lo que dificulta su interpretación y análisis. Para abordar este desafío, es necesario implementar técnicas de limpieza, integración y normalización de datos para garantizar la coherencia y la calidad de los datos.

Variabilidad en el uso de servicios:

Otra consideración importante es la variabilidad en el uso de servicios médicos entre los suscriptores. Los patrones de utilización pueden variar considerablemente según factores como la edad, el estado de salud, los hábitos de vida y las necesidades individuales. Algunos suscriptores pueden requerir servicios médicos de forma regular, mientras que otros pueden utilizarlos con menos frecuencia. Esta variabilidad en el uso de servicios debe tenerse en cuenta al analizar los datos de siniestros y al diseñar estrategias de atención al cliente y gestión de riesgos.

Influencia de factores externos:

Además, es importante reconocer la influencia de factores externos en el comportamiento de uso de servicios médicos. Factores socioeconómicos, culturales, geográficos y la coyuntura económica pueden afectar las decisiones de los suscriptores en cuanto al acceso y la utilización de servicios de salud. Por ejemplo, los niveles de ingresos pueden influir en la capacidad de pago de los servicios médicos, mientras que las diferencias culturales pueden afectar las preferencias de tratamiento y la disposición a buscar atención médica.

Tabla 5. Variables acumuladas de comportamiento de ingresos vs uso de servicio

Fecha Mes Generación	Presentado	Costo por Caso Presentado	Pagado Neto	Prima	Beneficiarios	Siniestralidad	Prima por Expuesto	Frecuencia	Expuesto
ene-2022	$6,966,628.33	$198.89	$4,719,081.76	$7,315,954.40	35,028	64.50%	$37.96	18.17%	192,727
feb-2022	$7,769,944.31	$220.47	$5,187,484.19	$7,254,506.25	35,243	71.51%	$37.72	18.33%	192,310
mar-2022	$7,743,360.94	$276.04	$5,164,249.58	$7,521,258.85	28,052	68.66%	$37.08	13.83%	202,830
abr-2022	$7,702,738.18	$234.95	$5,177,847.48	$7,557,770.86	32,784	68.51%	$39.47	17.12%	191,459
may-2022	$7,927,439.97	$240.40	$5,379,820.93	$7,234,305.68	32,976	74.37%	$38.69	17.64%	186,983
jun-2022	$8,006,080.39	$236.31	$5,263,940.37	$7,335,108.76	33,880	71.76%	$39.17	18.09%	187,270
jul-2022	$7,904,760.31	$235.15	$5,331,042.94	$7,426,664.86	33,616	71.78%	$39.19	17.74%	189,508
ago-2022	$8,593,242.68	$225.51	$5,747,589.68	$7,434,229.33	38,106	77.31%	$38.87	19.92%	191,259
sep-2022	$8,327,560.69	$246.47	$5,646,779.26	$7,562,855.67	33,788	74.66%	$38.57	17.23%	196,105
oct-2022	$8,251,211.42	$245.17	$5,702,033.24	$7,534,959.55	33,655	75.67%	$40.73	18.19%	184,988
nov-2022	$8,184,338.37	$238.82	$5,629,403.42	$7,377,185.65	34,270	76.31%	$40.19	18.67%	183,538
dic-2022	$8,058,743.35	$229.40	$5,531,928.23	$7,150,879.98	35,130	77.36%	$39.79	19.55%	179,709

Fortalezas de la incorporación de datos de siniestros:

La incorporación de datos de siniestros en el análisis de la medicina prepagada ofrece una serie de fortalezas significativas que pueden tener un impacto positivo en la gestión y la calidad del servicio. Al aprovechar estos datos de manera efectiva, las empresas pueden mejorar su comprensión del comportamiento del cliente, identificar patrones de riesgo y desarrollar estrategias de retención personalizadas. A continuación, se detallan estas fortalezas en mayor profundidad:

Mejora en la comprensión del comportamiento del cliente:
Los datos de siniestros proporcionan una ventana única para comprender el comportamiento del cliente en relación con el uso de servicios de medicina prepagada. Estos datos permiten a las empresas analizar y comprender mejor las necesidades, preferencias y patrones de uso de los suscriptores. Al examinar los registros de siniestros, las empresas pueden identificar los servicios más utilizados, los momentos en que los suscriptores requieren atención médica y las áreas donde se pueden mejorar los servicios.

Identificación de patrones de riesgo:
La incorporación de datos de siniestros facilita la detección de patrones de comportamiento que podrían indicar un mayor riesgo de abandono por parte de los suscriptores. Por ejemplo, ciertos patrones de uso de servicios médicos o ciertos tipos de reclamaciones recurrentes pueden indicar insatisfacción o problemas subyacentes con el servicio. Al identificar estos patrones de riesgo temprano, las empresas pueden intervenir de manera proactiva para abordar las preocupaciones de los clientes y mejorar su retención. Los datos de siniestros permiten a las empresas diseñar estrategias de retención personalizadas y dirigidas a segmentos específicos de suscriptores. Al comprender las necesidades individuales de los clientes y sus patrones de comportamiento, las empresas pueden desarrollar mensajes, ofertas y servicios personalizados que se ajusten a las preferencias de cada cliente.
En la Tabla 6 se identifican las variables relacionadas con la desafiliación de contratos, lo que proporciona una visión detallada de los factores que pueden influir en la decisión de los clientes de cancelar sus contratos.

Tabla 6. Variables de desafiliación de contratos

Contrato	Parámetro Desafiliación	Motivo Desafiliación	Fecha Generación Movimiento	Segmentación Cartera	Clasificación Motivo	Desafiliados	Prima Desafiliado
306859	SERVICIO	SERVICIO	12/12/2022	VENTAS NUEVAS	VOLUNTARIO	1	$20.36
305702	SITUACIÓN ECONOMICA	SALIDA DEL PAIS DEFINITIVA	23/11/2022	V1	VOLUNTARIO	1	$92.50
307868	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	4	$127.89
307831	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	2	$88.20
307840	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	1	$49.00
307837	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	1	$49.00
307838	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	1	$49.00
306858	VENTAS	INCONFORMIDAD VENTA	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	1	$39.00
307841	PRODUCCION	MIGRACION DE CONTRATO	31/1/2023	VENTAS NUEVAS	VOLUNTARIO	1	$39.00

Las variables de desafiliación de contratos pueden incluir una variedad de aspectos, como la duración del contrato, el tipo de servicio o producto contratado, el historial de pagos, la satisfacción del cliente, entre otros. Estas variables pueden proporcionar información valiosa sobre los factores que contribuyen a la desafiliación de los clientes y pueden utilizarse para desarrollar estrategias de retención de clientes más efectivas.

El análisis de las variables de desafiliación de contratos en la Tabla 6 permite identificar relaciones y patrones entre diferentes factores y la desafiliación de los clientes, lo que puede ayudar a predecir y prevenir la desafiliación futura, así como a mejorar la satisfacción del cliente y la retención de clientes a largo plazo.

En la Tabla 7 se identifican las variables relacionadas con las intenciones de desafiliación y las quejas de los clientes, lo que proporciona una visión detallada de los aspectos que pueden influir en la decisión de los clientes de cancelar sus servicios o expresar su insatisfacción. Estas variables son fundamentales para comprender las señales tempranas de desafiliación y para abordar las preocupaciones de los clientes antes de que se conviertan en problemas mayores.

Las variables de intenciones de desafiliación y quejas pueden incluir una variedad de aspectos, como la frecuencia y la gravedad de las quejas de los clientes, la satisfacción general del cliente, la disposición a renovar los servicios, entre otros. Estas variables proporcionan información valiosa sobre el nivel de compromiso y satisfacción de los clientes, así como sobre los factores que pueden estar contribuyendo a su insatisfacción o deseo de cancelar sus servicios.

En la Tabla 8 se presenta un resumen del uso por prestaciones médicas utilizadas, lo que ofrece una visión general de cómo los clientes hacen uso de los servicios médicos ofrecidos. Esta tabla proporciona información detallada sobre las prestaciones médicas más utilizadas por los clientes, así como la frecuencia o la cantidad de uso de cada prestación.

Tabla 7. Variables de Intenciones y Quejas.

Ticket	Estado	Fecha Solicitud	Fecha Resuelto	Contrato Intensión	Grupo Negocio	Resultado Gestión	Arquetipo Afiliado	Retenido	Tipo Facturación Contrato	Prima Intensión
187501	Resuelto	18/11/2021	11/1/2022	-	SIN DENIFIR	cliente_no_retenido	-	SI	SIN DEFINIR	0
203813	Resuelto	3/12/2021	6/1/2022	273497	INDIVIDUAL	cliente_retenido	-	SI	MENSUAL	189,96
206475	Resuelto	7/12/2021	3/1/2022	260982	INDIVIDUAL	cliente_pensará_oferta	-	SI	MENSUAL	384,6
207541	Resuelto	8/12/2021	14/2/2022	278112	INDIVIDUAL	cliente_no_retenido	PROTECTORES	NO	MENSUAL	95,49
210611	Resuelto	10/12/2021	12/1/2022	275775	INDIVIDUAL	cliente_retenido	CONSTRUCTORES	NO	MENSUAL	55,44
210642	Resuelto	10/12/2021	7/1/2022	278077	INDIVIDUAL	cliente_no_retenido	-	NO	MENSUAL	187,26
212584	Resuelto	13/12/2021	19/1/2022	265631	INDIVIDUAL	cliente_retenido	-	SI	MENSUAL	102,92
213288	Resuelto	13/12/2021	10/1/2022	271248	INDIVIDUAL	cliente_no_retenido	-	NO	MENSUAL	81,79
213532	Resuelto	14/12/2021	17/1/2022	267016	INDIVIDUAL	cliente_pensará_oferta	-	SI	MENSUAL	156,57
215337	Resuelto	15/12/2021	13/1/2022	0	SIN DENIFIR	cliente_retenido	-	SI	SIN DEFINIR	0
216045	Resuelto	15/12/2021	10/1/2022	-	SIN DENIFIR	cliente_no_retenido	-	SI	SIN DEFINIR	0
218066	Resuelto	17/12/2021	3/1/2022	269883	INDIVIDUAL	cliente_no_retenido	PROTECTORES	NO	MENSUAL	125,4
218202	Resuelto	17/12/2021	3/1/2022	275353	INDIVIDUAL	cliente_no_retenido	-	NO	MENSUAL	75,55
220236	Resuelto	20/12/2021	12/1/2022	93546	RENACER	cliente_no_retenido	-	NO	MENSUAL	9,46
220038	Resuelto	20/12/2021	3/1/2022	273429	PLAN ODONTOLOGICO	cliente_no_retenido	ACOGIDOS	NO	MENSUAL	18,08
219781	Resuelto	20/12/2021	3/1/2022	275278	PLAN ODONTOLOGICO	cliente_no_retenido	CONSTRUCTORES	NO	MENSUAL	36,16

Las prestaciones médicas pueden incluir una variedad de servicios, como consultas médicas, procedimientos diagnósticos, tratamientos especializados, medicamentos recetados, entre otros. Conocer el uso de estas prestaciones por parte de los clientes es fundamental para comprender las necesidades de atención médica de la población y para adaptar los servicios ofrecidos en consecuencia.

Tabla 8. Resumen movimientos por prestaciones.

Prestación	Tránsitos	Vlr. Presentado	Vlr. Recorte	Vlr. Deducible	Vlr. Copago	Vlr. Pagado Neto
VISITA EN LA OFICINA DE UN NUEVO PACIENTE	297,620	$9,224,045.1	$1,303,803.9	$1,144,152.6	$484.8	$5,055,416.9
MEDICAMENTOS	237,376	$13,479,666.9	$1,403.2	$714,323.3	$319,485.9	$9,858,407.3
MEDICAMENTOS VADEMECUM B AL 70%	103,969	$2,373,982.3	$0.0	$0.0	$0.0	$1,670,561.8
MEDICAMENTOS VADEMECUM A AL 90%	91,890	$2,028,626.4	$65.4	$0.0	$0.0	$1,452,745.2
CONSULTA MEDICA ESPECIALIDAD	69,844	$1,712,379.5	$281,291.2	$0.0	$0.0	$1,064,874.3
GASTOS NO CUBIERTOS	55,111	$7,936,609.2	$0.0	$0.0	$819,372.7	$45,958.8
LABORATORIO	51,917	$4,067,253.5	$54,376.8	$451,750.0	$179,793.3	$2,962,772.7
TERAPIA FISICA INTEGRAL	46,961	$3,304,754.3	$121,612.7	$107,147.4	$172.8	$2,515,611.2
BIOMETRIA HEMATICA	34,639	$251,465.6	$4,012.2	$152.1	$0.0	$204,338.9
IMAGEN.	32,397	$4,207,432.6	$21.8	$498,068.7	$167,117.8	$3,060,182.9
REMOCIÓN DE CALCULO SUPRAGINGIVAL	27,063	$170,496.9	$94,329.9	$0.0	$0.0	$136,287.9
TELECONSULTA	20,643	$243,127.7	$102.0	$45.0	$0.0	$236,775.3
SUMINISTROS	19,467	$5,285,975.2	$58.8	$349,489.8	$611,388.5	$4,020,189.1
GLUCOSA	19,392	$63,679.4	$86.5	$19.4	$0.0	$52,992.4
EMO (UROANALISIS DE RUTINA)	16,366	$64,932.5	$55.6	$19.4	$0.0	$53,334.9
COLESTEROL	16,008	$71,598.2	$0.0	$1.8	$0.0	$59,769.3
TRIGLICERIDOS	15,937	$58,619.0	$85.3	$1.8	$0.0	$49,102.5
ABONO	15,169	$270.5	$0.0	$0.0	$0.0	$-97,220.6

3.2 Preprocesamiento de Datos.

El preprocesamiento de datos es una etapa crucial en cualquier proyecto de análisis o modelado predictivo. Comprende una serie de pasos diseñados para garantizar que los datos estén limpios, estructurados y listos para su análisis. A continuación, se detallan los principales aspectos del preprocesamiento de datos:

1. Librerías:

El proceso de preprocesamiento de datos hace uso de varias librerías que proporcionan herramientas y funciones específicas para manipular y analizar datos de manera eficiente. Entre las librerías más utilizadas se encuentran:
Pandas: Esta librería es ampliamente utilizada para la carga, manipulación y limpieza de datos. Proporciona estructuras de datos flexibles y potentes, como el DataFrame, que facilita la manipulación de conjuntos de datos tabulares.
NumPy: NumPy es una librería fundamental para realizar operaciones matemáticas y numéricas en Python. Proporciona soporte para matrices multidimensionales y funciones matemáticas de alto rendimiento, lo que la hace ideal para el procesamiento de datos numéricos.

2. Limpieza de datos:

La limpieza de datos es un paso crucial en el preprocesamiento, ya que garantiza la calidad y consistencia de los datos utilizados en el análisis. Algunas técnicas comunes de limpieza de datos incluyen:

Eliminación de valores nulos o inconsistentes: Los valores nulos o faltantes pueden afectar la precisión de los análisis. La función dropna() de la librería pandas se utiliza para eliminar filas que contienen valores nulos en cualquier columna, asegurando así que los datos estén completos.

Eliminación de variables irrelevantes: En ocasiones, algunas variables pueden no ser relevantes para el análisis o la predicción de interés. Estas variables pueden eliminarse del conjunto de datos utilizando técnicas como la selección de características.

Corrección de valores atípicos (outliers): Los valores atípicos pueden distorsionar los resultados de los análisis estadísticos y los modelos predictivos. La técnica de winsorizing, implementada en Python mediante la función winsorize() de la librería scipy.stats, se utiliza para corregir valores atípicos mediante la limitación de los valores extremos a un rango predefinido.

Estos son solo algunos de los pasos básicos involucrados en el preprocesamiento de datos. Dependiendo de la naturaleza de los datos y los objetivos del análisis, pueden requerirse técnicas adicionales para garantizar la calidad y validez de los resultados obtenidos. El preprocesamiento de datos es un proceso iterativo y continuo que requiere atención y cuidado para garantizar la integridad de los datos y la robustez de los análisis realizados. En la figura 8, se muestra el resultado de la carga de datos y cómo se comportan estos datos.

	CONTRATO	NUMRENOV	VIGDESDE	VIGHASTA	ESTADO_CONTRATO	BROKER	PROVINCIA_CONTRATO	PLANCORTO	ID_TSB	prima	prima_real	afiliados
0	1284	19	2020-04-14	2021-04-13	2	HUMANA S.A.	PICHINCHA	MH 30.000 CLASICO	0	557.42	388.19	2
1	1284	20	2021-04-14	2022-04-13	2	HUMANA S.A.	PICHINCHA	MH 30.000 - COD. 53064	0	579.16	485.24	2

Figura 8. Análisis de Variables y comportamiento inicial.

En la Figura 9 se presenta el resultado del conteo de suscripciones vigentes para cada mes, lo que proporciona una visualización de la evolución temporal en el tiempo. Este análisis temporal es fundamental para comprender cómo varía el número de suscripciones a lo largo del período considerado y para identificar posibles tendencias o patrones emergentes.

El seguimiento de la evolución mensual de las suscripciones vigentes es esencial para evaluar el crecimiento o declive del servicio a lo largo del tiempo y para identificar posibles estacionalidades o cambios estacionales en la demanda de los clientes. Además, este análisis

puede ayudar a identificar períodos de mayor o menor actividad en el servicio, lo que puede ser útil para la planificación de recursos y estrategias de marketing.

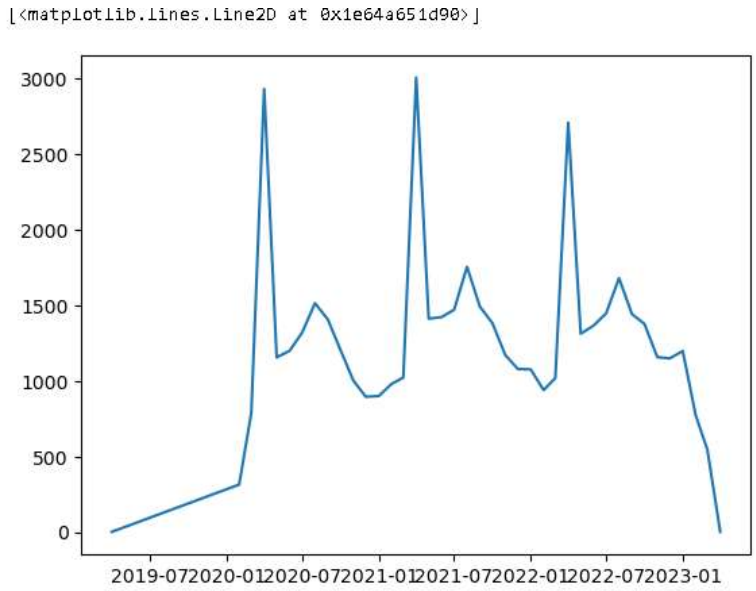


Figura 9. Conteo de suscripciones por meses.

En la Figura 10 se lleva a cabo un análisis del uso de servicios por mes, lo que proporciona una perspectiva detallada sobre cómo varía la utilización de los servicios a lo largo del tiempo. Este análisis es crucial para comprender la dinámica de la demanda de los servicios ofrecidos y para identificar posibles patrones estacionales, tendencias de uso o cambios en los hábitos de los usuarios a lo largo de los meses.

El seguimiento del uso mensual de los servicios permite detectar períodos de mayor o menor actividad, identificar picos de demanda o estacionalidades que puedan influir en la planificación de recursos y estrategias operativas. Además, este análisis puede proporcionar información valiosa para la optimización de la oferta de servicios, la personalización de las estrategias de marketing y la mejora continua de la experiencia del cliente.

Figura 10. Frecuencia de usos de Servicio.

En la Figura 11 se lleva a cabo una verificación exhaustiva de datos nulos o vacíos en el conjunto de datos, lo que es un paso crítico en el proceso de preprocesamiento de datos. Esta etapa es fundamental para garantizar la integridad y la calidad de los datos antes de realizar cualquier análisis o modelado. La identificación y el manejo adecuado de los datos faltantes son esenciales para evitar sesgos o imprecisiones en los resultados finales.

Una vez identificados los datos nulos o vacíos en la Figura 11, se procede a realizar la limpieza de datos correspondiente, lo cual se presenta en la Figura 12. Durante este proceso, se aplican diversas técnicas para tratar con los datos faltantes, como la eliminación de registros incompletos, la imputación de valores utilizando métodos estadísticos o la extrapolación de información de registros similares.

La limpieza de datos en la Figura 12 se lleva a cabo de manera sistemática y rigurosa, asegurando que el conjunto de datos final sea coherente, completo y apto para su posterior análisis. Esta etapa es crucial para garantizar la confiabilidad y la validez de los resultados obtenidos a partir de los datos, así como para maximizar la eficacia de cualquier modelo predictivo o análisis estadístico que se realice.

```
df_4.isnull().sum()
```

```
ontrato                                  0
ndex                                     0
ONTRATO                                  0
UMRENOV                                  0
IGDESDE                                  0
IGHASTA                                  0
STADO_CONTRATO                           0
ROKER                                    0
ROVINCIA_CONTRATO                        0
LANCORTO                                 0
D_TSB                                    0
rima                                     0
rima_real                                0
filiados                                 0
an_mujeres                               0
an_hombres                               0
ES_VIGDESDE                              0
d_contrato                               0
bjeto                                    0
antidad_renov_base                       0
ncremento_Renov                          0
ncremento_Renov_real                     0
edia_prima                               0
edia_prima_real                          0
td_prima                                 0
td_prima_real                            0
v                                        0
v_real                                   0
rimas                                    0
rimas_real                               0
ncremento_PrimaRenov                     0
ncremento_PrimaRenov_real                0
umero_renovacion_utltima                 0
orma_pago                                0
ucursal                                  0
ltima_fecha_d_facturacion                0
ltima_cuota                              0
dad_titular                              0
enero_titular                            0
agado_neto                            6085
ransito                               6085
RONICA_pagado_neto                   10629
OTENCIALMENTE_CRONICA_pagado_neto     9754
```

Figura 11. Verificación de nulos o vacíos

	contrato_limpio	promedio_dias	cantidad_dias	promedio_iteraciones_areas	cantidad_iteraciones_areas	incidensias
556	10008	4.000000	16.0	5.000000	20	4
593	10027	4.894737	93.0	4.421053	84	19
600	100292	5.000000	5.0	4.000000	4	1
620	10055	4.611111	83.0	3.444444	62	18
624	100632	5.317073	218.0	3.463415	142	41
...	...	...	...	...	...	...
40263	99494	4.294118	73.0	2.823529	48	17
40266	99514	7.857143	55.0	4.000000	28	7
40277	99894	4.129032	128.0	3.645161	113	31
40280	99900	9.333333	28.0	3.333333	10	3
40281	9992	4.666667	112.0	4.958333	119	24

Figura 12. Resultado limpieza datos

3.3 Transformación de Variables.

La transformación de variables es un paso esencial en el preprocesamiento de datos que tiene como objetivo preparar los datos para su posterior análisis o modelado. Este proceso implica convertir las variables en un formato adecuado para su uso en algoritmos de aprendizaje automático o análisis estadístico. Aquí se detalla ampliamente cómo se llevan a cabo dos técnicas comunes de transformación de variables:

Transformación de variables categóricas a variables numéricas:

Las variables categóricas representan categorías o grupos discretos y no numéricos, como el color, el tipo de vehículo o la región geográfica. Sin embargo, muchos algoritmos de aprendizaje automático requieren que todas las variables de entrada sean numéricas. Para convertir las variables categóricas en un formato numérico adecuado, se emplea la técnica de codificación one-hot.

La codificación one-hot crea nuevas columnas binarias (dummy variables) para cada categoría única en la variable original. Estas columnas tienen un valor de 1 si la observación pertenece a esa categoría y 0 en caso contrario. La función get_dummies() de la librería pandas es una herramienta conveniente para realizar esta codificación de manera eficiente.

Por ejemplo, si tenemos una variable categórica "Color" con tres categorías: Rojo, Verde y Azul, la codificación one-hot crearía tres nuevas columnas: "Rojo", "Verde" y "Azul". Cada fila tendría un 1 en la columna correspondiente al color de esa observación y 0 en las demás columnas.

Normalización de variables numéricas:

Las variables numéricas pueden tener diferentes escalas y rangos, lo que puede dificultar la comparación entre ellas o afectar el rendimiento de ciertos algoritmos de aprendizaje automático. La normalización es un proceso que ajusta los valores de las variables para que tengan una escala uniforme.

Una técnica común de normalización es la estandarización, que transforma los datos de manera que tengan una media de 0 y una desviación estándar de 1. Esto se logra restando la media de cada valor y dividiendo por la desviación estándar.

La función StandardScaler() de la librería sklearn.preprocessing proporciona una forma sencilla de realizar la estandarización en Python. Esta función calcula la media y la desviación estándar de cada característica y luego ajusta y transforma los datos según esta información.

La normalización de variables numéricas garantiza que todas las características contribuyan de manera equitativa al modelo final, independientemente de sus escalas originales. Esto puede mejorar la convergencia del algoritmo y ayudar a obtener resultados más estables y consistentes.
En la Figura 13 se presenta el resultado de la recategorización y normalización de variables, lo que representa un paso crucial en el proceso de preprocesamiento de datos. Este proceso tiene como objetivo transformar y estandarizar las variables en el conjunto de datos para que sean más adecuadas para su análisis posterior y para la construcción de modelos predictivos.

La recategorización de variables puede implicar la agrupación de categorías similares, la creación de nuevas categorías o la eliminación de categorías poco relevantes o redundantes. Esto ayuda a simplificar y estructurar los datos de una manera más significativa y fácilmente interpretable.

Por otro lado, la normalización de variables es importante para asegurar que todas las variables tengan la misma escala o rango de valores, lo que facilita la comparación y el análisis de las diferentes variables. Esto es especialmente importante en modelos de aprendizaje de máquina, donde las diferencias en la escala de las variables pueden afectar negativamente el rendimiento del modelo.

Out[71]:

_mujeres	can_hombres	Incremento_Renov	Incremento_Renov_real	media_prima	media_prima_real	std_prima	std_prima_real	
1	1	1.038001	1.250006	568.280000	436.715000	15.372501	68.624713	0
1	1	1.000000	1.075715	571.913333	465.136667	12.551595	69.123433	0
1	0	1.020782	1.010039	94.410000	113.125000	1.371787	0.799031	0
1	0	1.312749	1.080042	104.676667	116.346667	17.808825	5.608621	0
1	0	1.035673	1.100012	184.035000	281.840000	4.560839	17.635243	0

Figura 13. Resultados de recategorización y normalización de variable

En la Figura 14 se detalla el proceso de recategorización de variables cualitativas, lo que implica la transformación de variables categóricas en un formato más adecuado para el análisis y modelado posterior. Esta etapa es esencial para garantizar que las variables cualitativas sean tratadas de manera adecuada y significativa en el contexto del análisis de datos.

La recategorización de variables cualitativas puede involucrar varias técnicas, como la agrupación de categorías similares, la combinación de categorías poco frecuentes o la creación de nuevas categorías que sean más representativas o significativas para el análisis. Este proceso ayuda a simplificar la estructura de los datos y a reducir la complejidad, lo que facilita la interpretación y el análisis de las variables cualitativas.

Además, la recategorización de variables cualitativas puede ayudar a mejorar la calidad de los modelos predictivos al reducir el ruido o la variabilidad en los datos y al permitir una representación más precisa de las relaciones entre las variables y la variable objetivo.

```
Out[73]: {'directo': 0,
          'broker': 1,
          'EFECTIVO-DEBITO': 0,
          'TARJETA CREDITO': 1,
          'PAGO WEB': 2,
          'PICHINCHA': 0,
          'GUAYAS': 1,
          'OTRO': 2,
          'AZUAY': 3,
          'MEDIECUADOR HUMANA S A MATRIZ': 0,
          'HUMANA S A GUAYAQUIL': 1,
          'HUMANA S A CUENCA': 2,
          'MH 30.000': 0,
          'MH 50.000': 1,
          'PH 15.000': 2,
          'MH 150.000': 3,
          'PH 50.000': 4,
          'MH 80.000': 5,
          'PH 30.000': 6}
```

Figura 14. Validación de recategorización de variables cualitativas

3.4 Selección de características.

La selección de características es un proceso crucial en el análisis de datos y el modelado predictivo, que implica identificar las variables más relevantes para predecir o explicar el fenómeno de interés. Este proceso no solo mejora la precisión y eficiencia de los modelos, sino que también ayuda a reducir el sobreajuste y el tiempo de entrenamiento. A continuación, se detalla ampliamente cómo se lleva a cabo la selección de características utilizando las pruebas de t-student para variables cuantitativas y pruebas de Chi cuadrado para analizar las variables cualitativas más relevantes:

Determinación de las Variables Cualitativas Relevantes:

Las variables categóricas desempeñan un papel crucial en el modelado predictivo, ya que pueden proporcionar información valiosa sobre el comportamiento de los suscriptores y sus decisiones de abandono. En este análisis, se investiga la relación entre las variables categóricas y la variable objetivo ("Y"), que indica la probabilidad de abandono de un suscriptor.

Se implementa un algoritmo en Python utilizando la biblioteca scipy.stats para realizar la prueba de chi-cuadrado en todas las variables categóricas presentes en el conjunto de datos. Este algoritmo identifica las variables categóricas significativas en relación con la variable objetivo, utilizando un nivel de significancia del 0.20.
El algoritmo produce un listado de las variables categóricas significativas junto con sus estadísticas de chi-cuadrado y valores p correspondientes. Además, se proporcionan tablas de contingencia que muestran la distribución de la variable objetivo en cada categoría de las variables categóricas seleccionadas.

Se observa que algunas variables categóricas tienen una relación significativa con la variable objetivo, lo que sugiere que estas variables pueden influir en la probabilidad de abandono de los suscriptores. Estos hallazgos son fundamentales para comprender los factores que afectan la retención de los clientes en la empresa de medicina prepaga.

El análisis de variables categóricas proporciona información valiosa para la construcción del modelo de predicción de abandono de suscriptores. Las variables categóricas significativas identificadas pueden integrarse en el modelo para mejorar su capacidad predictiva y proporcionar recomendaciones específicas para la retención de clientes.

Además, este análisis destaca la importancia de considerar las características individuales de los suscriptores, como su historial de uso y preferencias, al diseñar estrategias de retención personalizadas.

En la Figura 15 se lleva a cabo el proceso de validación y estandarización de variables cualitativas, un paso crítico en el preprocesamiento de datos que garantiza la consistencia y la calidad de las variables antes de su análisis y modelado.
Por ejemplo, en este caso, "forma_pago_X", junto con el indicador "STR", sugiriendo que se trata de una variable categórica.

Luego, se presenta el valor de la prueba de chi-cuadrado, que es 781.02, seguido del valor p asociado, que es 0.0. Esto indica que hay una asociación significativa entre la forma de pago y la variable de respuesta.

A continuación, se muestran las proporciones de la forma de pago para cada nivel de la variable de respuesta. Por ejemplo, para el nivel "EFECTIVO-DEBITO", el 84.24% de las observaciones corresponden al grupo Y=0, mientras que solo el 15.76% corresponden al grupo Y=1. Este patrón se repite para los otros niveles de la variable

```
################################
STR
forma_pago_X
################################
Chi-2... 781.0229606650252
P-value... 0.0
Y                          0         1
forma_pago_X
EFECTIVO-DEBITO     0.842377  0.157623
PAGO WEB            0.615258  0.384742
TARJETA CREDITO     0.880664  0.119336
################################
Y                       0     1
forma_pago_X
EFECTIVO-DEBITO      9454  1769
PAGO WEB             1121   701
TARJETA CREDITO      7424  1006


################################
STR
PROVINCIA_CONTRATO
################################
Chi-2... 22.371709573206225
P-value... 5e-05
Y                              0         1
PROVINCIA_CONTRATO
AZUAY                   0.848649  0.151351
GUAYAS                  0.815496  0.184504
OTRO                    0.795687  0.204313
PICHINCHA               0.842619  0.157381
################################
Y                          0     1
PROVINCIA_CONTRATO
AZUAY                    157    28
GUAYAS                  1684   381
OTRO                     701   180
PICHINCHA              15457  2887


################################
STR
sucursal
################################
Chi-2... 131.08591172060967
P-value... 0.0
```

Figura 15.Resultados de la determinación de variables cualitativas relevantes

Determinación de las Variables Cuantitativas Relevantes:

Las variables cuantitativas son importantes indicadores de comportamiento y características de los suscriptores. En este análisis, se investiga cómo estas variables contribuyen a la predicción del abandono de suscriptores y su impacto en la rentabilidad de la empresa.

Se implementa un algoritmo en Python utilizando la biblioteca pingouin para realizar pruebas de t de Student independientes en todas las variables cuantitativas presentes en el conjunto de datos. Este algoritmo identifica las variables cuantitativas significativas en relación con la variable objetivo ("Y"), utilizando un nivel de significancia del 0.05.

El algoritmo produce un listado de las variables cuantitativas significativas junto con sus estadísticas de t de Student independiente. Además, se proporcionan las medias y desviaciones estándar de estas variables para cada grupo de la variable objetivo ("Y").

Se observa que algunas variables cuantitativas muestran diferencias significativas entre los grupos de abandono y retención de suscriptores. Estos hallazgos sugieren que estas variables pueden influir en la probabilidad de abandono de los suscriptores y son relevantes para la toma de decisiones estratégicas.

En la Figura 16 se lleva a cabo el proceso de validación y estandarización de variables cuantitativas, un paso esencial en el preprocesamiento de datos que asegura la coherencia y la calidad de las variables numéricas antes de su análisis y modelado.

Durante la validación, se examinan las variables cuantitativas para identificar posibles errores, valores atípicos o datos faltantes que puedan afectar la integridad de los datos. Este proceso es fundamental para garantizar la fiabilidad de las variables y la precisión de los resultados derivados de ellas.
Como por ejemplo la variable "NUMRENOV" exhibió una diferencia considerable entre los grupos, con un valor absoluto de prueba t de 22.20. Esto sugiere que la cantidad de renovaciones tiene un impacto significativo en la distinción entre los grupos. Además, la media de "NUMRENOV" fue notablemente más alta en el grupo "Y=0" en comparación con el grupo "Y=1", con valores promedio de 5.04 y 3.67 respectivamente.

Otra variable que podemos analizar es la "prima" mostró una diferencia estadísticamente significativa entre los grupos, con un valor absoluto de prueba t de 14.25. Las estadísticas descriptivas revelaron que la media de la prima fue más alta en el grupo "Y=0" en comparación con el grupo "Y=1", con valores promedio de 177.40 y 143.13 respectivamente.

Este hallazgo sugiere que el monto de la prima puede ser un factor crucial en la distinción entre los grupos definidos por "Y".

Por último, la variable "prima_real" también exhibió una diferencia significativa entre los grupos, con un valor absoluto de prueba t de 12.95. La media de la prima real fue más alta en el grupo "Y=0" en comparación con el grupo "Y=1", con valores promedio de 169.85 y 139.79 respectivamente

```
################################
INT
NUMRENOV
T-test    22.199659
Name: T, dtype: float64
------------------------------
      media     var_s
Y
0  5.043280  3.469785
1  3.673188  2.493083
********************************
********************************


################################
INT
prima
T-test    14.251084
Name: T, dtype: float64
```

Figura 16. Resultados de las variables cuantitativas resultantes

En las Figuras 17 y 18 se detallan todas las variables resultantes que serán utilizadas en el entrenamiento del modelo. Estas figuras representan un paso crucial en el proceso de preparación de datos, donde se han aplicado diversas técnicas de preprocesamiento para asegurar la calidad y la relevancia de las variables utilizadas en el análisis y modelado subsiguiente.

.

```
Out[77]: Index(['NUMRENOV', 'CANAL', 'forma_pago_X', 'PROVINCIA_CONTRATO', 'sucursal',
                'PLANCORTO_x', 'prima', 'prima_real', 'MES_VIGd', 'afiliados',
                'can_mujeres', 'can_hombres', 'Incremento_Renov',
                'Incremento_Renov_real', 'media_prima', 'media_prima_real', 'std_prima',
                'std_prima_real', 'cv', 'cv_real', 'Incremento_PrimaRenov',
                'Incremento_PrimaRenov_real', 'nuemro_renovacion_utltima',
                'edad_titular', 'pagado_neto', 'transito', 'CRONICA_pagado_neto',
                'POTENCIALMENTE CRONICA_pagado_neto', 'AGUDA_pagado_neto',
                'ODONTOLOGIA_pagado_neto', 'COVID-19_pagado_neto',
                'NO APLICA_pagado_neto', 'MATERNIDAD_pagado_neto', 'CANCER_pagado_neto',
                'AMBULATORIO_pagado_neto', 'HOSPITALARIO_pagado_neto',
                'EMERGENCIA_pagado_neto', 'HOSPITAL DEL DIA_pagado_neto',
                'SIN DEFINIR_pagado_neto', 'PAGO AFILIADO_pagado_neto',
                'PAGO PRESTADOR_pagado_neto', 'REEMBOLSO LIQUIDADO_pagado_neto',
                'REEMBOLSO NEGADO_pagado_neto', 'REEMBOLSO DEVUELTO_pagado_neto',
                'CRONICA_transito', 'POTENCIALMENTE CRONICA_transito', 'AGUDA_transito',
                'ODONTOLOGIA_transito', 'COVID-19_transito', 'NO APLICA_transito',
                'MATERNIDAD_transito', 'CANCER_transito', 'AMBULATORIO_transito',
                'HOSPITALARIO_transito', 'EMERGENCIA_transito',
                'HOSPITAL DEL DIA_transito', 'SIN DEFINIR_transito',
                'PAGO AFILIADO_transito', 'PAGO PRESTADOR_transito',
                'REEMBOLSO LIQUIDADO_transito', 'REEMBOLSO NEGADO_transito',
                'REEMBOLSO DEVUELTO_transito', 'promedio_dias', 'cantidad_dias',
                'promedio_iteraciones_areas', 'cantidad_iteraciones_areas',
                'incidensias', 'Y'],
               dtype='object')
```

Figura 17. Variables resultantes Y

```
Out[81]: Index(['AGUDA_transito', 'POTENCIALMENTE CRONICA_transito',
                'AMBULATORIO_transito', 'HOSPITALARIO_pagado_neto',
                'CRONICA_pagado_neto', 'media_prima', 'cantidad_iteraciones_areas',
                'sucursal', 'MATERNIDAD_pagado_neto', 'PAGO PRESTADOR_pagado_neto',
                'REEMBOLSO LIQUIDADO_transito', 'Incremento_Renov',
                'ODONTOLOGIA_pagado_neto', 'POTENCIALMENTE CRONICA_pagado_neto',
                'AGUDA_pagado_neto', 'prima_real', 'AMBULATORIO_pagado_neto',
                'PLANCORTO_x', 'std_prima', 'CRONICA_transito',
                'PAGO AFILIADO_transito', 'NUMRENOV', 'edad_titular',
                'media_prima_real', 'incidensias', 'std_prima_real', 'cantidad_dias',
                'PAGO AFILIADO_pagado_neto', 'forma_pago_X', 'PAGO PRESTADOR_transito',
                'Incremento_PrimaRenov', 'promedio_iteraciones_areas', 'cv',
                'nuemro_renovacion_utltima', 'promedio_dias', 'Incremento_Renov_real',
                'transito', 'MATERNIDAD_transito', 'PROVINCIA_CONTRATO', 'prima',
                'pagado_neto', 'REEMBOLSO LIQUIDADO_pagado_neto'],
               dtype='object')
```

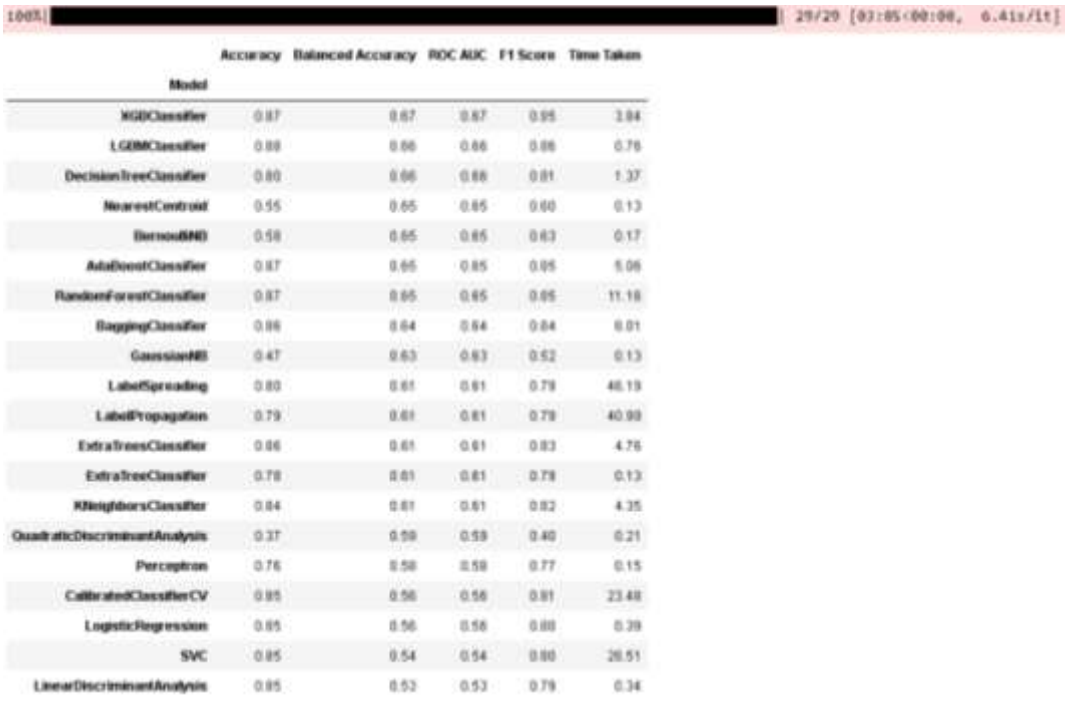
Figura 18. Variables resultantes X

3.5 Entrenamiento del Modelo

El proceso de entrenamiento del modelo es fundamental para desarrollar un sistema de aprendizaje automático eficiente y preciso. Aquí se detalla exhaustivamente cada etapa, desde la selección de librerías hasta la validación cruzada:

En este contexto, se utilizó la biblioteca LazyPredict para realizar un análisis exhaustivo de múltiples modelos de aprendizaje automático supervisado. En primer lugar, se realizó una división de los datos en conjuntos de entrenamiento y prueba mediante la función train_test_split de la biblioteca sklearn.model_selection, con el fin de garantizar una evaluación precisa del rendimiento de los modelos. Luego, se instanció un clasificador LazyClassifier.Este clasificador fue entrenado utilizando los conjuntos de datos de entrenamiento y prueba mediante el método fit. En la Figura 19 como resultado, se obtuvieron una variedad de modelos de aprendizaje automático, junto con las predicciones correspondientes para el conjunto de prueba. Este análisis de LazyPredict constituye una etapa crucial en la exploración de diferentes enfoques de modelado y proporciona una base sólida para la selección del modelo óptimo en función de criterios específicos de rendimiento y precisión en la predicción de datos.

El análisis exhaustivo realizado mediante LazyPredict reveló que el modelo XGBoost Classifier sobresale en términos de métricas de rendimiento clave para nuestro conjunto de datos específico. Con una precisión (accuracy) de 0.87, un área bajo la curva ROC (ROC AUC) de 0.85 y un puntaje F1 de 0.85, el XGBoost Classifier demostró consistentemente un alto nivel de exactitud y capacidad de generalización en la predicción de los datos. Además, vale la pena destacar que este modelo logró este impresionante rendimiento en un tiempo de ejecución eficiente. Estos hallazgos respaldan la elección del XGBoost Classifier como el modelo más adecuado para nuestro conjunto de datos, ya que ofrece un equilibrio óptimo entre precisión y eficiencia computacional. Este resultado tiene importantes implicaciones para la toma de decisiones en futuros proyectos de análisis de datos, donde la selección de un modelo de aprendizaje automático efectivo es fundamental para obtener resultados precisos y confiables.



Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
XGBClassifier	0.67	0.67	0.67	0.65	3.84
LGBMClassifier	0.69	0.66	0.66	0.66	0.76
DecisionTreeClassifier	0.60	0.66	0.66	0.61	1.37
NearestCentroid	0.55	0.65	0.65	0.60	0.13
BernoulliNB	0.58	0.65	0.65	0.63	0.17
AdaBoostClassifier	0.67	0.65	0.65	0.65	5.06
RandomForestClassifier	0.67	0.65	0.65	0.65	11.18
BaggingClassifier	0.66	0.64	0.64	0.64	0.81
GaussianNB	0.47	0.63	0.63	0.52	0.13
LabelSpreading	0.80	0.61	0.61	0.79	46.19
LabelPropagation	0.79	0.61	0.61	0.79	40.99
ExtraTreesClassifier	0.66	0.61	0.61	0.63	4.76
ExtraTreeClassifier	0.78	0.61	0.61	0.78	0.13
KNeighborsClassifier	0.64	0.61	0.61	0.62	4.35
QuadraticDiscriminantAnalysis	0.37	0.59	0.59	0.40	0.21
Perceptron	0.76	0.58	0.58	0.77	0.15
CalibratedClassifierCV	0.85	0.56	0.56	0.81	23.48
LogisticRegression	0.85	0.56	0.56	0.80	0.39
SVC	0.85	0.54	0.54	0.80	26.51
LinearDiscriminantAnalysis	0.85	0.53	0.53	0.79	0.34

Figura 19. Exploración y comparación resultados algoritmos

Después de haber identificado que el XGBoost Classifier es el modelo con mejor rendimiento para nuestros datos, procedemos a importar la biblioteca XGBoost y a entrenar el modelo XGBoost Classifier con los datos de entrenamiento.

Em este caso para el modelo realizamos um analisis de características de las cuales la característica más significativa en términos de puntaje de importancia es "transito", con un puntaje de 0.25, seguida de "forma_pago_X" y "promedio_iteraciones_areas" con puntajes de 0.07 y 0.06 respectivamente. Estas características son identificadas como elementos clave que influyen en la capacidad del modelo para realizar predicciones precisas. Al examinar el acumulado de puntajes, vemos que estas tres características en conjunto representan aproximadamente el 38% de la importancia total, destacando su relevancia en la predicción del resultado deseado. Por otro lado, las características como "nuemro_renovacion_utltima" y "Incremento_PrimaRenov" tienen puntajes de importancia más bajos, lo que sugiere una contribución relativamente mínima al proceso de predicción. Estos hallazgos proporcionan una guía crucial para la selección y optimización de características en futuros análisis, así como para la comprensión de los factores subyacentes que influyen en las predicciones del modelo.

Tabla 9. Resultado variables más relevantes top 20.

Score	Acu Score	columna	
0	0.25	0.25	transito
1	0.07	0.32	forma_pago_X
2	0.06	0.38	promedio_iteraciones_areas
3	0.04	0.42	pagado_neto
4	0.03	0.45	NUMRENOV
5	0.03	0.48	REEMBOLSO LIQUIDADO_transito
6	0.03	0.51	MATERNIDAD_pagado_neto
7	0.02	0.53	cantidad_dias
8	0.02	0.56	Incremento_Renov
9	0.02	0.58	promedio_dias
10	0.02	0.6	PROVINCIA_CONTRATO
11	0.02	0.62	cv
12	0.02	0.64	Incremento_Renov_real
13	0.02	0.66	AMBULATORIO_pagado_neto
14	0.02	0.68	CRONICA_transito
15	0.02	0.7	AGUDA_pagado_neto
16	0.02	0.72	edad_titular
17	0.02	0.74	POTENCIALMENTE CRONICA_pagado_neto
18	0.02	0.76	PAGO AFILIADO_pagado_neto
19	0.02	0.77	media_prima_real
20	0.02	0.79	std_prima

3.6 Evaluación del Modelo

En esta sección, profundizaremos en el análisis de la matriz de confusión generada por el modelo XGBoost, así como en las implicaciones de los resultados obtenidos.La biblioteca XGBoost es conocida por su eficacia y rendimiento en una variedad de problemas de aprendizaje automático, lo que la convierte en una elección adecuada para este análisis.

El parámetro "objective" se configuró en "reg:linear", lo que indica que el modelo se entrenará para realizar regresión lineal, es decir, predecir valores numéricos continuos. Este enfoque fue elegido en función de la naturaleza de nuestros datos y el objetivo específico de nuestro análisis. Además, se estableció el parámetro "random_state" en 42 para asegurar la reproducibilidad de los resultados, lo que garantiza que los resultados del modelo sean consistentes entre diferentes ejecuciones del código. Finalmente, el parámetro "base_score" se definió en 0.5, proporcionando un punto de partida neutral para las predicciones del modelo. Estos parámetros fueron seleccionados después de una cuidadosa consideración de las características de los datos y con el objetivo de maximizar el rendimiento del modelo en términos de precisión predictiva y estabilidad.

Matriz de Confusión:

La matriz de confusión es una herramienta esencial que proporciona una visión detallada de cómo el modelo clasifica las instancias en función de su verdadero estado. En el contexto del problema de predicción de abandono de suscriptores en un servicio de medicina prepagada, la matriz de confusión del modelo XGBoost se presenta de la siguiente manera:

Análisis de la Matriz de Confusión:
El análisis detallado de la matriz de confusión revela varios aspectos importantes sobre el desempeño del modelo XGBoost:

Precisión en la Predicción de Suscriptores en Riesgo de Abandono:

La precisión en la predicción de suscriptores en riesgo de abandono, representada en la Figura 21, es una métrica fundamental para evaluar el desempeño del modelo. En esta figura se observa una precisión general del 86%, lo que significa que el 86% de las predicciones realizadas por el modelo son correctas en todas las categorías. Esta métrica, también conocida como sensibilidad o tasa de verdaderos positivos, adquiere una relevancia crucial, ya que indica la capacidad del modelo para identificar de manera correcta a aquellos suscriptores que realmente están en riesgo y requieren intervención.

Este resultado destaca la eficacia del modelo en la identificación precisa de los suscriptores en riesgo de abandono, lo cual es esencial para implementar acciones preventivas o estrategias de retención de clientes de manera oportuna. Una alta precisión en la predicción de los suscriptores en riesgo de abandono proporciona a las empresas una valiosa herramienta para gestionar proactivamente la satisfacción y la lealtad de los clientes, lo que puede tener un impacto significativo en la retención y el crecimiento del negocio.

```
In [85]:  1 xgb_model.score(X_test, y_test)
Out[85]: 0.8699944123672937
```

Figura 20. Resultados Score

Precisión en la Predicción de Suscriptores que no están en Riesgo de Abandono:

El modelo exhibe una alta precisión del 90% en la predicción de suscriptores que no están en riesgo de abandono. Esta métrica, conocida como especificidad o tasa de verdaderos negativos, es igualmente importante ya que indica la capacidad del modelo para identificar correctamente a aquellos suscriptores que no están en riesgo y no necesitan intervención.

```
plot_confusion_matrix(xgb_model,X_test,y_test, normalize =
plt.show()
```

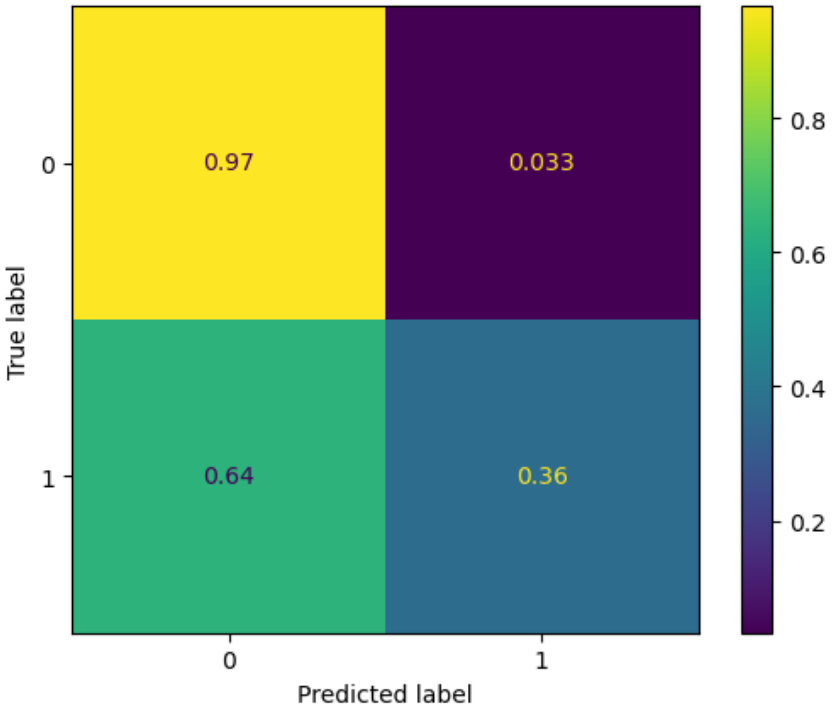


Figura 21. Resultados Matriz Confusión

En la Figura 22 y Figura 23, se presentan los resultados obtenidos del entrenamiento del modelo, los cuales revelan la presencia de falsos positivos y falsos negativos, aspectos cruciales a considerar en el análisis de su desempeño.

Falsos Positivos:

El modelo logra predecir correctamente 4,358 de los 4,507 contratos que están en riesgo de abandono, lo que indica una precisión del 86% en esta categoría. No obstante, se observa que el modelo también clasifica de manera errónea 149 contratos que no están en riesgo de abandono como pertenecientes a esta categoría.

Falsos Negativos:

Por otro lado, el modelo muestra una precisión del 87% al predecir correctamente 313 de los 362 contratos que no están en riesgo de abandono. Sin embargo, se identifica una cantidad

considerable de 549 contratos en riesgo de abandono que son clasificados incorrectamente como no pertenecientes a esta categoría.

Estos hallazgos resaltan la importancia de considerar los falsos positivos y falsos negativos en la evaluación del rendimiento del modelo, ya que pueden tener implicaciones significativas en la toma de decisiones y estrategias relacionadas con la retención de clientes.

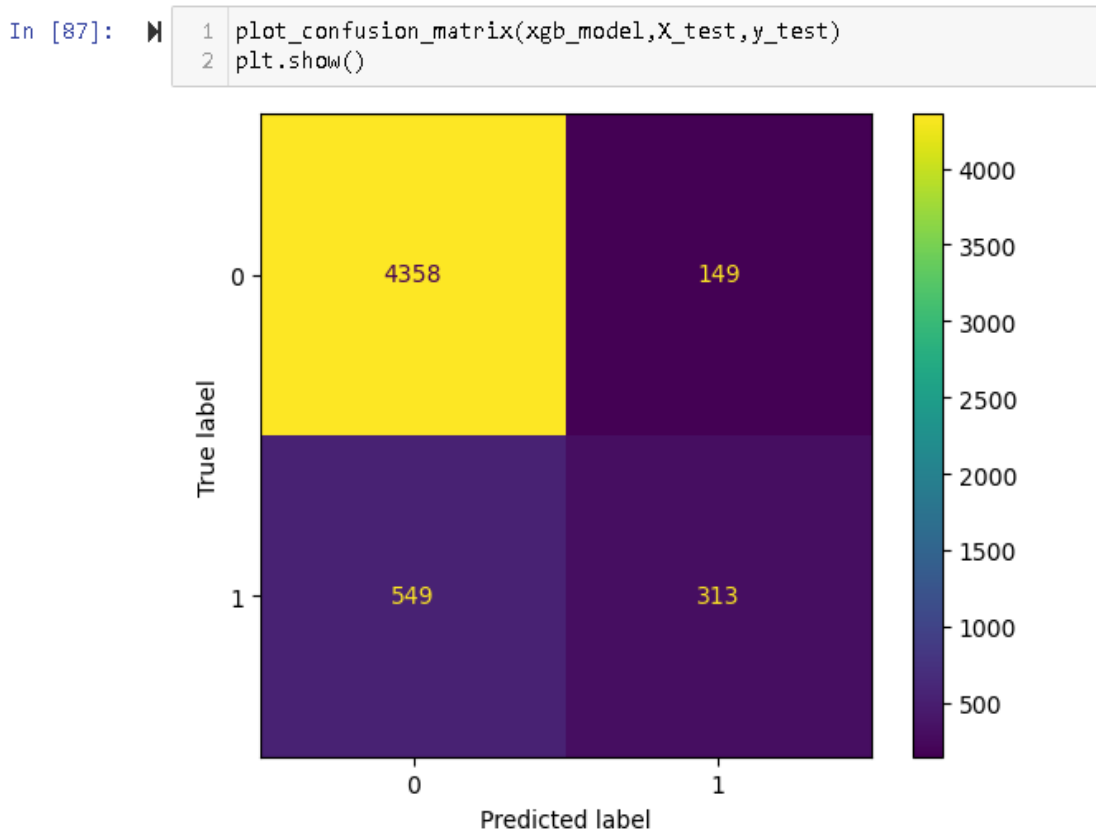


Figura 22. Resultados Matriz Confusión

```
from sklearn.svm import SVC
from sklearn import  metrics
xgb_CURVE = metrics.plot_roc_curve(xgb_model, X_test, y_test)

```

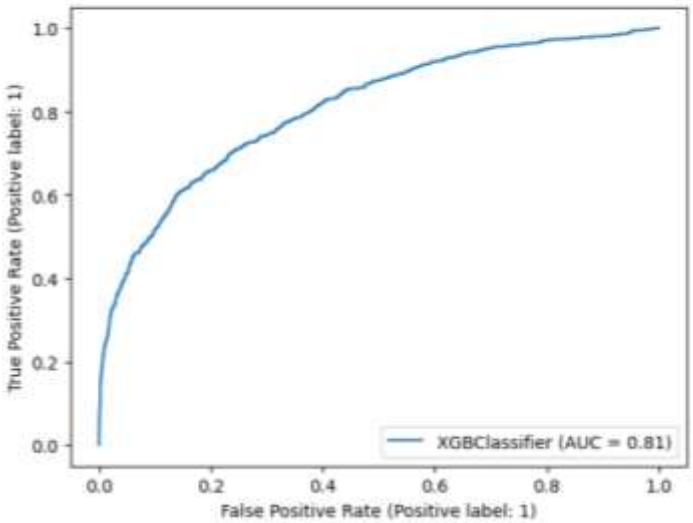


Figura 23. Resultados Curva ROC

4. CONCLUSIONES

- Se han destacado las consideraciones sobre la construcción del modelo, tomando en cuenta la disponibilidad de datos y los desafíos asociados a su análisis.

- Se identificaron las variables más relevantes para predecir el abandono de los suscriptores, incluyendo datos demográficos, información del plan de salud, historial de pagos, utilización de servicios y motivos de cancelación.

- La combinación estratégica de datos demográficos, características clave, registros recientes de interacciones con clientes y datos de siniestros, junto con la aplicación de técnicas adecuadas de limpieza, preparación y modelado, permitirá construir un modelo robusto y efectivo para la predicción de la deserción.

- Con los resultados obtenidos de precisión, sensibilidad, especificidad con resultados que tienen un buen grado de confiabilidad, por lo que se puede decir que la biblioteca XGBOOST es una buena herramienta para este tipo de modelamientos.

- Con los resultados del modelo establecer estrategias enfocadas hacia la atención del cliente, aumentar la rentabilidad y fortalecer la posición competitiva de la empresa en el mercado ecuatoriano.

5. RECOMENDACIONES

- Implementar técnicas de aprendizaje continuo para adaptar el modelo a cambios en el comportamiento de los clientes y del mercado, así como a las nuevas estrategias empresariales. Esto implica mantener el modelo actualizado y receptivo a las dinámicas cambiantes del entorno comercial y las preferencias de los clientes.

- Realizar reentrenamientos periódicos del modelo con datos actualizados con el fin de detectar posibles nuevos comportamientos. Esto garantizará que el modelo esté

siempre al tanto de las tendencias emergentes y pueda ajustarse en consecuencia para mantener su precisión y relevancia.

- Evaluar constantemente la necesidad de incorporar nuevas variables o modificar las existentes. Este análisis debe realizarse en función de la evolución de los sistemas de información y de la disponibilidad de datos relevantes. Es importante que estas nuevas variables agreguen valor significativo al modelo y contribuyan a una mejor predicción del abandono de suscriptores.

- Realizar estudios adicionales para comprender y ampliar el conocimiento sobre los factores que influyen en el abandono de los suscriptores. Esto podría implicar investigaciones cualitativas para obtener insights más profundos sobre las motivaciones y comportamientos de los clientes, así como el análisis de datos externos para identificar posibles variables adicionales que podrían mejorar la capacidad predictiva del modelo. Estos estudios adicionales pueden ayudar a perfeccionar el modelo y a desarrollar estrategias de retención más efectivas a largo plazo.

- Implementar técnicas de aprendizaje continuo para adaptar el modelo a cambios en el comportamiento de los clientes y del mercado.

- Evaluar la necesidad de incorporar nuevas variables o modificar las existentes, claro está dependiendo si hay nuevos sistemas incorporados y estas nuevas variables aporten significativamente.

- Realizar estudios adicionales para comprender y ampliar mejor el conocimiento sobre los factores que influyen en el abandono de los suscriptores así como la afectación por tomas de decisiones estratégicas.

6. BIBLIOGRAFÍA

[1] W. Y. Ayele, "Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset," International Journal of Advanced Computer Science and Applications, vol. 11, no. 6, pp. 20-32, 2020.

[2] Totango, "The Complete Guide to Customer Retention," 2017. [Online]. Available: https://www.totango.com/whitepapers/the-complete-guide-to-customer-retention. [Accessed: Apr. 01, 2023].

[3] A. Decker, "Organizational Customers' Retention Strategies on Customer Satisfaction: Case of Equity Bank Thika Branch, Kenya," Mar. 1, 2001.

[4] E S. Nasir, "Customer Retention Strategies and Customer Loyalty," Jan. 1, 2016.

[5] C. L. Corso, "Aplicación de algoritmos de clasificación supervisada usando Weka," Universidad Tecnológica Nacional, Facultad Regional Córdoba, Córdoba, 2009

[6] F. Herrera, "Big Data: Preprocesamiento y calidad de datos," novática, vol. 237, pp. 17, 2016.

[7] J. M. Marín, "Introducción a las redes neuronales aplicadas," Universidad Carlos III de Madrid, Madrid, 2012.

[8] S. L. Qian, J. He, and C. L. Wang, "Telecom Customer Churn Prediction Model Based on Improved SVM," Journal of Management Sciences, vol. 20, no. 1, pp. 54-58, 2007.

[9] R. Wang and C. Chen, "The application of attribute selection based on data mining to churn predictive model," Computer Applications and Software, vol. 24, no. 11, pp. 98-113, 2007.

[10] The Chartered Institute of Marketing, "Cost of Customer Acquisition versus Customer Retention," 2010.

[11] A. Amin, S. Shehzad, C. Khan, I. Ali, and S. Anwar, "Churn prediction in telecommunication industry using rough set approach," in New Trends in Computational Collective Intelligence, Springer, 2015, pp. 83-95.

[12] G. Klepac, Developing Churn Models Using Data Mining Techniques and

Social Network Analysis, IGI Global, 2014.

[13] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 2094-2097, 2016.

[14] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[15] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," in Eighth International Conference on Digital Information Management (ICDIM 2013), pp. 131-136, IEEE, 2013.

[16] A. Scherer, N. V. Wünderlich, and F. Von Wangenheim, "The value of self-service: Long-term effects of technology-based self-service usage on customer retention," MIS Quarterly, vol. 39, no. 1, pp. 177-200, 2015.

[17] C. K. Manner, "Who posts online customer reviews? The role of sociodemographics and personality traits," Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, vol. 26, pp. 125-142, 2013.

[18] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[19] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.

[20] K. Coussement and D. Van den Poel, "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers," Expert Systems with Applications, vol. 36, no. 3, pp. 6127-6134, 2009.

7. ANEXOS

Anexo I. Código

El código fue construido como base en Jupyter y con llamadas a consultas en SQL el cual está disponible en el siguiente link
https://github.com/jorgeluisguatosantamaria/Modelo_Desercion

Anexo II Resultados de Procesamiento y Modelamiento

Resultados Variables cualitativas.

##################################

STR

forma_pago_X

##################################

Chi-2... 781.0229606650252

P-value... 0.0

Y	0	1
forma_pago_X		
EFECTIVO-DEBITO	0.842377	0.157623
PAGO WEB	0.615258	0.384742
TARJETA CREDITO	0.880664	0.119336

##################################

Y	0	1
forma_pago_X		
EFECTIVO-DEBITO	9454	1769
PAGO WEB	1121	701
TARJETA CREDITO	7424	1006

##################################

STR

PROVINCIA_CONTRATO

##################################

Chi-2... 22.371709573206225

P-value... 5e-05

Y	0	1
PROVINCIA_CONTRATO		
AZUAY	0.848649	0.151351
GUAYAS	0.815496	0.184504
OTRO	0.795687	0.204313
PICHINCHA	0.842619	0.157381

##################################

Y	0	1
PROVINCIA_CONTRATO		
AZUAY	157	28
GUAYAS	1684	381

OTRO 701 180
PICHINCHA 15457 2887

##################################
STR
sucursal
##################################
Chi-2... 131.88591172060967
P-value... 0.0
Y 0 1
sucursal
HUMANA S A CUENCA 0.727848 0.272152
HUMANA S A GUAYAQUIL 0.778302 0.221698
MEDIECUADOR HUMANA S A MATRIZ 0.851296 0.148704
##################################
Y 0 1
sucursal
HUMANA S A CUENCA 115 43
HUMANA S A GUAYAQUIL 2805 799
MEDIECUADOR HUMANA S A MATRIZ 15079 2634

##################################
STR
PLANCORTO_x
##################################
Chi-2... 181.03903644415112
P-value... 0.0
Y 0 1
PLANCORTO_x
MH 150.000 0.878581 0.121419
MH 30.000 0.905825 0.094175
MH 50.000 0.856780 0.143220
MH 80.000 0.863079 0.136921
PH 15.000 0.780561 0.219439
PH 30.000 0.816536 0.183464
PH 50.000 0.853867 0.146133
##################################

Y	0	1
PLANCORTO_x		
MH 150.000	644	89
MH 30.000	933	97
MH 50.000	2022	338
MH 80.000	4507	715
PH 15.000	2618	736
PH 30.000	4780	1074
PH 50.000	2495	427

Resultados Variables Cuantitativas.

##################################
INT
NUMRENOV
T-test 22.199659
Name: T, dtype: float64
---
media var_s
Y
0 5.043280 3.469785
1 3.673188 2.493083
****
****

##################################
INT
prima
T-test 14.251084
Name: T, dtype: float64
---
media var_s
Y
0 177.399504 134.91097
1 143.131775 99.10480
****
****

```
################################
INT
prima_real
T-test    12.95215
Name: T, dtype: float64
--------------------------------
       media       var_s
Y
0  169.846099  129.438403
1  139.790273  100.824356
********************************
********************************


################################
INT
Incremento_Renov
T-test   -2.270177
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  1.179818  0.280617
1  1.191667  0.287358
********************************
********************************


################################
INT
Incremento_Renov_real
T-test   -5.786075
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  1.120019  0.185601
```

1 1.141295 0.254926
****
****

##########
INT
media_prima
T-test   13.942355
Name: T, dtype: float64
---
       media   var_s
Y
0 163.669137 125.356605
1 132.505233   92.486281
****
****

##########
INT
media_prima_real
T-test   13.160372
Name: T, dtype: float64
---
       media   var_s
Y
0 159.669774 120.486201
1 131.233539   94.145416
****
****

##########
INT
std_prima
T-test   8.210476
Name: T, dtype: float64

```
      media      var_s
Y
0  22.592325  29.416499
1  18.260841  22.990848
********************************
********************************


################################
INT
std_prima_real
T-test    5.759089
Name: T, dtype: float64
--------------------------------
      media      var_s
Y
0  17.978028  27.323590
1  15.142233  22.319623
********************************
********************************


################################
INT
cv
T-test    2.140504
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  0.151512  0.138442
1  0.146018  0.139158
********************************
********************************


################################
INT
Incremento_PrimaRenov
```

T-test -2.270177

Name: T, dtype: float64

---

    media   var_s

Y

0 1.179818 0.280617

1 1.191667 0.287358

****

****

##########

INT

nuemro_renovacion_utltima

T-test   22.199659

Name: T, dtype: float64

---

    media   var_s

Y

0 5.043280 3.469785

1 3.673188 2.493083

****

****

##########

INT

edad_titular

T-test   18.318215

Name: T, dtype: float64

---

     media   var_s

Y

0 46.188788 14.856420

1 41.248274 12.898862

****

****

```
##################################
INT
pagado_neto
T-test    7.878193
Name: T, dtype: float64
--------------------------------
        media        var_s
Y
0  1342.842062  3131.880847
1   712.402193  2491.587717
**********************************
**********************************


##################################
INT
transito
T-test    17.455164
Name: T, dtype: float64
--------------------------------
      media      var_s
Y
0  16.933052  19.008427
1   8.612864  10.451409
**********************************
**********************************


##################################
INT
CRONICA_pagado_neto
T-test    5.500629
Name: T, dtype: float64
--------------------------------
       media        var_s
Y
0  712.696181  1812.013340
1  376.735227  1186.683066
**********************************
```

****

##########
INT
POTENCIALMENTE CRONICA_pagado_neto
T-test 4.99856
Name: T, dtype: float64
---
      media   var_s
Y
0 513.327216 1420.468723
1 287.512939 1313.852089
****
****

##########
INT
AGUDA_pagado_neto
T-test 4.946077
Name: T, dtype: float64
---
      media   var_s
Y
0 256.194175 756.507926
1 136.762486 424.201912
****
****

##########
INT
ODONTOLOGIA_pagado_neto
T-test -2.341549
Name: T, dtype: float64
---
      media   var_s
Y

0 53.628036 254.032225
1 268.254211 1288.279432
****
****

##########
INT
MATERNIDAD_pagado_neto
T-test -6.322708
Name: T, dtype: float64
---
       media   var_s
Y
0 1091.025848 1233.790308
1 1950.150901 1419.624280
****
****

##########
INT
AMBULATORIO_pagado_neto
T-test 12.048867
Name: T, dtype: float64
---
      media   var_s
Y
0 643.077110 1259.261250
1 264.252914 415.232036
****
****

##########
INT
HOSPITALARIO_pagado_neto
T-test 2.04537
Name: T, dtype: float64

```
--------------------------------
        media        var_s
Y
0  3856.167236  4596.565674
1  3145.896302  5914.194005
********************************
********************************


FALLO..... SIN DEFINIR_pagado_neto
##################################
INT
PAGO AFILIADO_pagado_neto
T-test    3.164413
Name: T, dtype: float64
--------------------------------
       media        var_s
Y
0  538.458121  1609.747572
1  343.608159   769.168618
********************************
********************************


##################################
INT
PAGO PRESTADOR_pagado_neto
T-test    6.668652
Name: T, dtype: float64
--------------------------------
        media        var_s
Y
0  1120.100799  2736.337868
1   626.139625  2454.048827
********************************
********************************


##################################
```

INT

REEMBOLSO LIQUIDADO_pagado_neto

T-test 7.795541

Name: T, dtype: float64

---

	media	var_s
Y		
0	1347.844249	3136.634455
1	719.827600	2503.479013

****

****

##########

INT

CRONICA_transito

T-test 12.169882

Name: T, dtype: float64

---

	media	var_s
Y		
0	8.161414	9.812154
1	4.170492	4.791655

****

****

##########

INT

POTENCIALMENTE CRONICA_transito

T-test 11.340299

Name: T, dtype: float64

---

	media	var_s
Y		
0	6.891801	7.667335
1	4.190875	4.597283

****

****

```
################################
INT
AGUDA_transito
T-test    9.13771
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  5.664460  6.155488
1  3.849802  4.528093
********************************
********************************


################################
INT
MATERNIDAD_transito
T-test   -2.039837
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  3.437895  3.135114
1  4.140187  3.561981
********************************
********************************


################################
INT
AMBULATORIO_transito
T-test    17.376252
Name: T, dtype: float64
--------------------------------
      media      var_s
Y
0  16.666545  18.749703
```

1 8.440615 10.294689
****
****

FALLO..... SIN DEFINIR_transito
##########
INT
PAGO AFILIADO_transito
T-test 5.649494
Name: T, dtype: float64
---
    media   var_s
Y
0 5.836906 7.963141
1 4.088825 6.060727
****
****

##########
INT
PAGO PRESTADOR_transito
T-test 15.711186
Name: T, dtype: float64
---
     media   var_s
Y
0 14.680737 17.123892
1   7.600536   9.172902
****
****

##########
INT
REEMBOLSO LIQUIDADO_transito
T-test 17.484853
Name: T, dtype: float64

```
--------------------------------
      media      var_s
Y
0  16.541670  18.655445
1   8.326793  10.030041
**********
**********


##########
INT
promedio_dias
T-test   -12.396247
Name: T, dtype: float64
--------------------------------
     media     var_s
Y
0  5.670108  2.446647
1  6.240286  2.161727
**********
**********


##########
INT
cantidad_dias
T-test    5.447839
Name: T, dtype: float64
--------------------------------
     media      var_s
Y
0  84.90835  77.379183
1  76.94898  70.719576
**********
**********


##########
INT
```

promedio_iteraciones_areas

T-test 16.062622

Name: T, dtype: float64

---

  media var_s

Y

0 3.720949 1.328485

1 3.287811 1.790271

****

****

##########

INT

cantidad_iteraciones_areas

T-test 10.192277

Name: T, dtype: float64

---

  media var_s

Y

0 63.192358 67.128143

1 50.211194 63.566782

****

****

##########

INT

incidensias

T-test 9.834171

Name: T, dtype: float64

---

  media var_s

Y

0 16.535126 15.842030

1 13.611936 13.781804

****

****

Resultado Exploracion de Algoritmos.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
XGBClassifier	0.87	0.67	0.67	0.85	3.84
LGBMClassifier	0.88	0.66	0.66	0.86	0.76
DecisionTreeClassifier	0.80	0.66	0.66	0.81	1.37
NearestCentroid	0.55	0.65	0.65	0.60	0.13
BernoulliNB	0.58	0.65	0.65	0.63	0.17
AdaBoostClassifier	0.87	0.65	0.65	0.85	5.06
RandomForestClassifier	0.87	0.65	0.65	0.85	11.16
BaggingClassifier	0.86	0.64	0.64	0.84	8.01
GaussianNB	0.47	0.63	0.63	0.52	0.13
LabelSpreading	0.80	0.61	0.61	0.79	46.19
LabelPropagation	0.79	0.61	0.61	0.79	40.99
ExtraTreesClassifier	0.86	0.61	0.61	0.83	4.76
ExtraTreeClassifier	0.78	0.61	0.61	0.79	0.13
KNeighborsClassifier	0.84	0.61	0.61	0.82	4.35
QuadraticDiscriminantAnalysis	0.37	0.59	0.59	0.40	0.21
Perceptron	0.76	0.58	0.58	0.77	0.15
CalibratedClassifierCV	0.85	0.56	0.56	0.81	23.48
LogisticRegression	0.85	0.56	0.56	0.80	0.39
SVC	0.85	0.54	0.54	0.80	26.51
LinearDiscriminantAnalysis	0.85	0.53	0.53	0.79	0.34
LinearSVC	0.84	0.52	0.52	0.78	6.33
PassiveAggressiveClassifier	0.75	0.50	0.50	0.74	0.18
RidgeClassifierCV	0.84	0.50	0.50	0.77	0.25
RidgeClassifier	0.84	0.50	0.50	0.77	0.14
SGDClassifier	0.84	0.50	0.50	0.77	0.55
DummyClassifier	0.84	0.50	0.50	0.77	0.08

Resultados Variables Significativas

	Score	Acu Score	columna
0	0.25	0.25	transfo
1	0.07	0.32	forma_pago_X
2	0.05	0.38	promedio_iteraciones_areas
3	0.04	0.42	pagado_neto
4	0.03	0.45	NUMRENOV
5	0.03	0.48	REEMBOLSO LIQUIDADO_transfo
6	0.03	0.51	MATERNIDAD_pagado_neto
7	0.02	0.53	cantidad_dias
8	0.02	0.56	Incremento_Renov
9	0.02	0.58	promedio_dias
10	0.02	0.60	PROVINCIA_CONTRATO
11	0.02	0.62	ca
12	0.02	0.64	Incremento_Renov_rea
13	0.02	0.66	AMBULATORIO_pagado_neto
14	0.02	0.68	CRONICA_transfo
15	0.02	0.70	AGUDA_pagado_neto
16	0.02	0.72	edad_titular
17	0.02	0.74	POTENCIALMENTECRONICA_pagado_neto
18	0.02	0.76	PAGO AFILIADO_pagado_neto
19	0.02	0.77	media_prima_rea
20	0.02	0.79	std_prima
21	0.02	0.80	Incidencias
22	0.02	0.82	std_prima_rea
23	0.02	0.84	MATERNIDAD_transfo
24	0.01	0.85	PAGO PRESTADOR_pagado_neto
25	0.01	0.86	prima_rea
26	0.01	0.88	prima
27	0.01	0.89	PAGO PRESTADOR_transfo
28	0.01	0.90	cantidad_iteraciones_areas
29	0.01	0.92	PLANCORTO_X
30	0.01	0.93	sucursal
31	0.01	0.94	CRONICA_pagado_neto
32	0.01	0.95	PAGO AFILIADO_transfo
33	0.01	0.96	media_prima
34	0.01	0.97	AMBULATORIO_transfo
35	0.01	0.98	ODONTOLOGIA_pagado_neto
36	0.01	0.99	POTENCIALMENTECRONICA_transfo
37	0.01	1.00	HOSPITALARIO_pagado_neto
38	0.00	1.00	AGUDA_transfo
39	0.00	1.00	numero_renovacion_ultima
40	0.00	1.00	Incremento_PrimaRenov
41	0.00	1.00	REEMBOLSO LIQUIDADO_pagado_neto