PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE

MAGISTER EN SISTEMAS DE INFORMACIÓN CON MENCIÓN EN DATA SCIENCE

TEMA:

ANÁLISIS PREDICTIVO DEL CHURN DE CLIENTES PARA UNA EMPRESA PROVEEDORA DEL SERVICIO DE INTERNET PARA HOGARES EN EL ECUADOR

AUTOR:

WILLIAM HERNÁN CHUQUER ERAZO

DIRECTOR:

Ph.D. JHONNY VLADIMIR PINCAY NIEVES

Quito, Agosto – 2024

DEDICATORIA

Dedico infinitamente este proyecto a Dios, por darme la energía, voluntad y fuerzas necesarias para alcanzar este objetivo. A mi esposa, a quien amo con todo mi ser, y quien ha estado a mi lado en las buenas y en las malas, apoyándome incansablemente en la realización de este proyecto. En este mismo año, ella me dará el mayor y más hermoso regalo, el nacimiento de nuestra hija, a quien espero con ansias, y que ha sido una gran fuente de inspiración para alcanzar y cumplir con esta meta.

También dedico este proyecto a mis padres, hermanos y amigos, quienes han estado pendientes de mi proceso durante todo este tiempo, confiando en mi capacidad para alcanzar este logro de superación personal y profesional.

AGRADECIMIENTO

A Dios, por ser mi fuerza y motor de vida, quien bajo su manto protector, he podido alcanzar muchos de los objetivos que me he propuesto a lo largo de este tiempo.

A mi esposa, por ser un pilar fundamental en mi vida, quien con su constante apoyo y palabras oportunas ha sido clave para alcanzar esta meta personal como profesional.

A mis padres y hermanos, por su incondicional apoyo y por estar siempre pendientes de mí.

A Ángel Haro y Jacqueline Villamarín, por su guía y apoyo, y por proporcionarme valiosa información que ha sido crucial para alcanzar los resultados deseados en este proyecto.

Finalmente, al Dr. Jhonny Pincay por su valiosa contribución, compromiso y experiencia, que fueron esenciales para lograr cumplir con este proyecto.

RESUMEN

El presente estudio se enfoca en el desarrollo de un modelo predictivo para identificar el churn de clientes en una empresa proveedora del servicio de internet para hogares en el Ecuador. Utilizando la metodología CRISP-DM, en la fase de preparación de los datos se realizó un análisis exploratorio utilizando Python, lo que permitió identificar patrones preliminares en el comportamiento de los clientes. En las fases posteriores, se utilizaron herramientas como Alteryx, que facilitó la creación de flujos de trabajo para lograr un análisis exhaustivo de los datos de clientes, aplicando técnicas de limpieza, segmentación y parametrización para garantizar la calidad y relevancia de la información. Qlik Sense fue empleada para lograr una visualización clara y efectiva de los resultados. La técnica de WOE + IV fue crucial para la segmentación y parametrización de variables, mejorando la precisión predictiva.

Se entrenaron y compararon varios modelos de machine learning, entre ellos, Regresión Logística, Árboles de Decisión y Random Forest, siendo este último el modelo con el mejor desempeño. Random Forest alcanzó una precisión del 99.5% para la predicción de cancelaciones administrativas y del 81% para cancelaciones voluntarias. Estos resultados proporcionan a la empresa herramientas sólidas para la implementación de estrategias de retención de clientes, permitiendo una intervención proactiva en la gestión de riesgos y mejorando la satisfacción del cliente. El estudio también destaca la importancia de actualizar periódicamente los modelos para adaptarse a cambios en los patrones de comportamiento de los clientes.

ABSTRACT

This study focuses on developing a predictive model to identify customer churn in a home internet service provider in Ecuador. Using the CRISP-DM methodology, the data preparation phase included an exploratory analysis conducted with Python, which helped identify preliminary patterns in customer behavior. In subsequent phases, tools like Alteryx were employed to create workflows for a comprehensive analysis of customer data, applying techniques such as data cleansing, segmentation, and parameterization to ensure the quality and relevance of the information. Qlik Sense was used to achieve clear and effective visualization of the results. The WOE + IV technique was crucial for variable segmentation and parameterization, enhancing predictive accuracy.

Several machine learning models, including Logistic Regression, Decision Trees, and Random Forest, were trained and compared, with Random Forest emerging as the best-performing model. Random Forest achieved an accuracy of 99.5% for predicting administrative cancellations and 81% for voluntary cancellations. These results provide the company with robust tools to implement effective customer retention strategies, enabling proactive risk management and improving customer satisfaction. The study also emphasizes the importance of periodically updating the models to adapt to changes in customer behavior patterns.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO I: INTRODUCCIÓN

1. Introducción

En los últimos años, el uso generalizado del Internet se ha consolidado como una herramienta indispensable en el desarrollo social y económico, generando un impacto significativo en la vida diaria de las personas. En Ecuador, el uso de Internet ha experimentado un crecimiento notable, transformando los hábitos de consumo y la manera en que los ciudadanos interactúan con los servicios en línea, destacando la importancia estratégica de la industria de las telecomunicaciones.

La Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador, reconocida por su liderazgo en el mercado nacional mediante la tecnología FTTH (Fiber To The Home), enfrenta un gran desafío para retener a sus clientes debido al alto índice de churn en los últimos años, lo cual afecta su rentabilidad y sostenibilidad a largo plazo. Atraer nuevos clientes requiere inversiones significativas, por lo que la pérdida de clientes antes de un período crítico es financieramente perjudicial para la empresa.

El churn o tasa de abandono de clientes, se refiere al porcentaje de usuarios que dejan de utilizar los servicios de una empresa en un periodo de tiempo determinado (Salesforce, 2023). En la industria de las telecomunicaciones, el churn es un indicador crucial, ya que una alta tasa de deserción puede impactar significativamente los ingresos y la estabilidad financiera de las compañías. Por esta razón, las empresas utilizan el análisis del churn para identificar patrones y factores que influyen en la decisión de los clientes de abandonar el servicio.

La problemática del churn en el sector de las telecomunicaciones se debe a diversos motivos, que van desde la insatisfacción con los servicios hasta la competencia agresiva que ofrece mejores opciones. En este contexto, esta investigación se centra en desarrollar un modelo predictivo del churn para comprender mejor los factores que influyen en la decisión de los clientes de dar de baja el servicio. Esto permitirá a la empresa implementar estrategias efectivas de retención, reducir la pérdida de clientes y mantener su liderazgo en la industria de las telecomunicaciones en el Ecuador.

1.1. Antecedentes

Según una investigación del Instituto Ecuatoriano de Estadística y Censos (INEC), el acceso a internet en los hogares del Ecuador ha experimentado un crecimiento significativo en el año 2023 en comparación con el año anterior. A nivel nacional, el acceso aumentó del 60,4% al 62,2%, reflejando una tendencia ascendente en la adopción de servicios de internet entre los hogares ecuatorianos. Sin embargo, este aumento no se distribuye uniformemente en todos los sectores, ya que se observa una disminución en el acceso a internet en áreas urbanas, pasando del 70,1% en 2022 al 69,7% en 2023, mientras que en áreas rurales se registra un notable incremento del 38% al 44,4% durante el mismo período (INEC, 2023).

La Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador ha sido líder en el mercado nacional al ofrecer servicios de internet utilizando la tecnología FTTH (Fiber To The Home - Fibra Óptica hasta el Hogar). Esta tecnología de vanguardia ha revolucionado la industria de las telecomunicaciones al proporcionar una conexión directa de fibra óptica a los hogares, redefiniendo los estándares de calidad y velocidad en el

acceso a internet. Gracias a su enfoque innovador, la empresa se ha destacado por su compromiso con la excelencia y su capacidad para superar a la competencia en términos de ofrecer soluciones avanzadas en el ámbito de la conectividad residencial (CERES, 2023).

En un estudio, la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador, se posiciona como el internet más rápido del país, reconocido por grandes marcas mundiales como es OOKLA por cinco años consecutivos y a la vez por nPerf (empresa con sede en Lyon) quienes emiten resultados globales basados en un barómetro de las conexiones fijas a Internet en el Ecuador, mencionando que Netlife y Puntonet han ofrecido el mejor rendimiento general en redes fijas de Ecuador, durante los últimos dos semestres; mientras que Claro Ecuador ha conseguido la mejor latencia (24 ms), aunque Netlife se le acerca con (25 ms). Considerar también a la vez en solitario a Netlife con la mayor velocidad media de carga en los dos últimos semestres (NPERF, 2023).

Figura 1: Conexiones a Internet fijo en Ecuador.

Fuente: (NPERF, 2023)

Actualmente, la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador cuenta con más de 800 mil abonados a nivel nacional (Ekos, 2023), donde Guayaquil y Quito destacan como las ciudades con mayor cantidad de consumidores del servicio. Aunque en el año 2020 experimentó un crecimiento abrumador debido al impacto del COVID-19, este fue considerado como un año atípico. Para lo cual, desde el año 2023 hasta la fecha actual, el crecimiento ha disminuido y la empresa carece de una comprensión clara sobre qué clientes están en riesgo de abandonar el servicio de internet. Por lo tanto, es crucial desarrollar un algoritmo que identifique a estos clientes y nos permita tomar medidas proactivas para retenerlos. Esto implica la implementación de campañas o promociones específicas diseñadas para incentivar la permanencia de los clientes.

Al presente, la empresa cuenta con una gran cantidad de información tanto del Sistema de la Central Telefónica (llamadas entrantes de clientes) como del Sistema Transaccional de Operaciones (sistema de registro de facturación, tickets de soporte técnico, entre otros) permitiendo obtener data importante para el análisis y desarrollo del problema planteado. Serán consideradas variables para ayudar a entender el comportamiento de los clientes, con el objetivo principal de determinar la probabilidad de abandonar el servicio contratado con la empresa de telecomunicaciones.

Para lograrlo, analizaremos variables claves, que permitan obtener información valiosa sobre el comportamiento de los clientes y los factores que pueden influir en su decisión de cancelar el servicio. Los datos principales por utilizar son:

- **Datos del cliente:** Información demográfica como edad, género, ubicación geográfica, estado civil, nivel educativo, entre otros. Además, datos sobre la antigüedad del cliente,

el tipo de plan contratado, el historial de pagos, las interacciones anteriores con el servicio al cliente y cualquier cambio en los planes o servicios utilizados.

Consideraciones: Es importante mencionar que, bajo la Ley Orgánica de Protección de Datos Personales, no se utilizarán datos específicos como nombres del cliente, direcciones, correos electrónicos, teléfonos, entre otros. De esta manera, tanto la empresa como sus productos y servicios contratados por el cliente no se verán afectados.

- **Datos de comportamiento del cliente:** El uso del servicio de internet, como el ancho de banda utilizado, la frecuencia de conexión, los dispositivos utilizados para acceder al servicio, el registro de quejas por llamadas o visitas presenciales a los centros de atención al cliente donde se generan tareas o casos por indisponibilidad del servicio, las promociones otorgadas y las campañas en las que participó el cliente serán recopilados y analizados.
- **Datos de satisfacción del cliente:** Las encuestas de satisfacción del cliente, los comentarios y opiniones recopiladas a través de diferentes canales de comunicación, como correos electrónicos, llamadas telefónicas, chats en línea o redes sociales, serán utilizadas para proporcionar información sobre la percepción del cliente respecto a la calidad del servicio, la atención al cliente y otros aspectos relevantes.
- **Datos de facturación y pago:** La información sobre el historial de facturación, los métodos de pago utilizados, los problemas de facturación y cualquier retraso en los pagos serán analizados para obtener una variable clave como es el comportamiento de pago necesario para un mejor análisis de la data.

De acuerdo con lo mencionado, se recopilará información de una muestra de clientes antiguos que cancelaron el servicio para identificar patrones relevantes. Esta información será extraída de la base de datos de los últimos 6 meses a un año, dependiendo de las características de los procesos planteados para la extracción de información. Para este propósito, se emplearán algoritmos de machine learning para calcular esta tasa, y los resultados se presentarán de manera dinámica en una herramienta de visualización como Qlik Sense.

Una metodología estructurada será seguida, dividiendo el proyecto en diversas fases, desde la recopilación de datos hasta la presentación de resultados en la herramienta de visualización. Para profundizar en nuestro entendimiento y enfoque del proyecto, una variedad de conceptos relevantes serán consultados, lo que ayudará a clarificar nuestros objetivos y enfoque de trabajo.

Por tanto, la retención de clientes es considerada parte fundamental para las empresas. Ante la fuga de clientes, se busca identificar a aquellos con mayor probabilidad de renunciar a un producto o servicio. Por lo tanto, se intenta evitar esta fuga centrando los recursos y procedimientos de manera eficiente según las políticas comerciales. La fuga de clientes es vista como un conjunto de problemas para el negocio, ya que el tamaño de la cartera de clientes está directamente relacionado con la rentabilidad del negocio (Jélvez Caamaño, 2014).

1.2. Planteamiento del problema

De acuerdo con los datos obtenidos (ver figura 2 y figura 3), se constata que la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador experimenta un crecimiento constante durante la última década. Sin embargo, recientemente, se ha observado un notable incremento en el número de abonados que cancelan el servicio. Por tanto, es crucial que este desafío sea enfrentado para retener a los clientes en un mercado altamente competitivo y en constante evolución.

Figura 2: Tendencia Histórica de Clientes.

Realizado por: CHUQUER, William, 2024

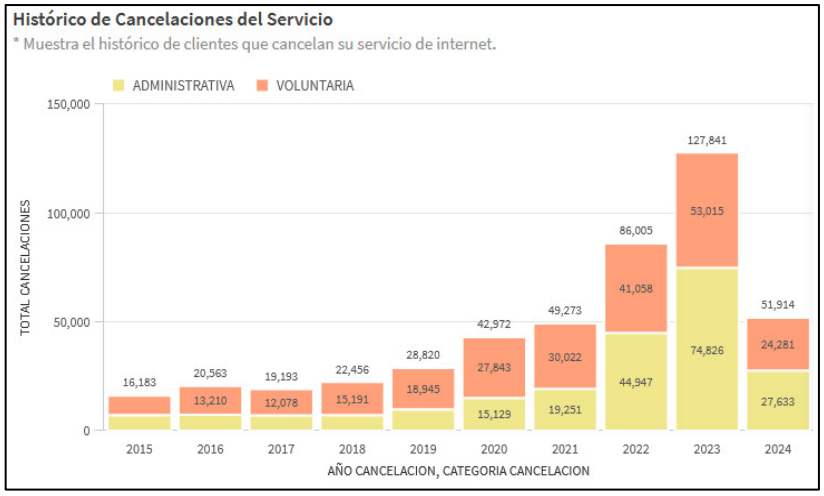


Figura 3: Histórico Cancelaciones del Servicio.

Realizado por: CHUQUER, William, 2024

Con base a los datos visualizados, se realizó un análisis que permitió obtener el churn rate de los clientes que cancelan el servicio durante la última década (información presentada hasta el 15 de mayo del 2024, ver figura 4). Por tanto, para abordar este problema, es importante que los patrones y tendencias presentes en los datos históricos del churn de clientes sean comprendidos y que esta información sea utilizada para desarrollar estrategias efectivas de retención de clientes.

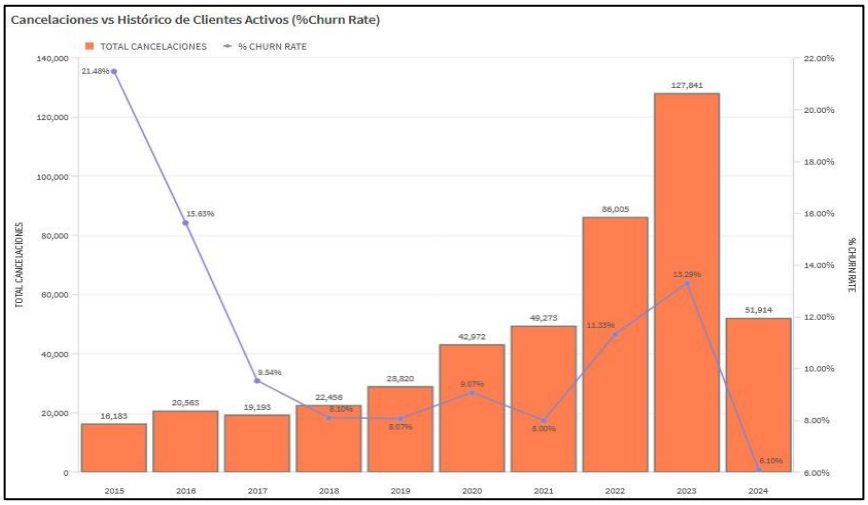


Figura 4: Cancelaciones versus Churn Rate (%).

Realizado por: CHUQUER, William, 2024

Es necesario que los datos demográficos, el comportamiento del usuario y los niveles de satisfacción del cliente sean priorizados e integrados de manera efectiva en un modelo predictivo de churn. Esta integración no solo mejorará la precisión del modelo predictivo, sino que también permitirá que acciones útiles sean aplicadas para mejorar la retención de clientes y para fortalecer la competitividad en el mercado ecuatoriano.

Por último, es fundamental que las diferentes técnicas de aprendizaje automático sean exploradas y evaluadas para desarrollar un modelo predictivo de churn. Identificar las técnicas más efectivas en términos de precisión y utilidad para la retención de clientes ayudará a que la empresa implemente estrategias más eficientes y efectivas para reducir el churn y mejorar la satisfacción del cliente.

Con lo antes expuesto, nos planteamos las siguientes preguntas:

- *¿Cuáles son los patrones y tendencias identificados en los datos históricos del churn de clientes en la empresa proveedora del servicio de internet para hogares en el Ecuador?*
- *¿Cómo integrar eficazmente los patrones y tendencias en un modelo predictivo para retener clientes en un mercado en constante evolución bajo estrategias efectivas?*

Encontrar respuestas a estas preguntas es esencial porque proporcionarán una comprensión detallada de los patrones y tendencias del churn de clientes, permitiendo a la empresa identificar las causas de la pérdida de clientes y desarrollar estrategias efectivas de retención. Integrar estos patrones en un modelo predictivo no solo prevendrá la pérdida de clientes, sino que también mejorará la satisfacción del cliente y la competitividad en un mercado cambiante. En última instancia, responder a estas preguntas garantizará la sostenibilidad y el éxito a largo plazo de la empresa en el mercado ecuatoriano.

1.3. Justificación

En la actualidad, tener acceso al internet se ha convertido en una herramienta fundamental para el desarrollo social en la actualidad. Su impacto en la vida cotidiana acerca a las personas al mundo tecnológico, permitiéndoles interactuar con plataformas y servicios en línea. Como resultado, el acceso generalizado a Internet se ha vuelto un tema prioritario para los gobiernos, quienes implementan políticas cada vez más avanzadas para garantizar su disponibilidad (Agencia de Regulación y Control de las Telecomunicaciones, 2020).

Este avance tecnológico genera cambios significativos en las organizaciones, impulsando su integración en la era digital a través de las diversas oportunidades que ofrece Internet y las tecnologías de la información y comunicación (TIC) en el ámbito comercial. Hoy en día, el crecimiento del uso de Internet está permitiendo reducir costos en la entrega de productos y servicios, superando las barreras geográficas y facilitando el intercambio entre proveedores y consumidores. Por lo tanto, el número de personas que utilizan servicios de Internet sigue en aumento (Gangeshwers, 2013).

El crecimiento del uso de Internet en Ecuador ha transformado los hábitos de consumo, lo que destaca la importancia estratégica de la retención de clientes para las empresas proveedoras de servicios de internet. Por lo tanto, la investigación sobre el churn de clientes es relevante y oportuna para la empresa, ya que impacta directamente en la rentabilidad y competitividad empresarial. Para abordarlo, es fundamental comprender los patrones y tendencias del churn y desarrollar un modelo predictivo que ayude a detectar comportamientos de los clientes y así tomar acciones preventivas en consecuencia de

ayudar a reducir la pérdida de clientes, permitiendo proporcionar una ventaja competitiva en un mercado dinámico como el ecuatoriano.

1.4. Objetivos

1.4.1. Objetivo general

Desarrollar un algoritmo dentro del Análisis Predictivo del Churn de Clientes en una empresa proveedora del servicio de internet para hogares en el Ecuador.

1.4.2. Objetivos específicos

- Analizar los datos históricos del Churn de Clientes que permitan identificar patrones y tendencias.
- Integrar datos demográficos, comportamiento del usuario y niveles de satisfacción del cliente en el análisis predictivo.
- Explorar diferentes técnicas de aprendizaje automático para desarrollar un modelo predictivo de Churn de Clientes.
- Diseñar un dashboard como herramienta de análisis y control de información que permita realizar un análisis visual y detallado de los datos obtenidos.
- Proporcionar recomendaciones para estrategias de retención de clientes basadas en los resultados del análisis predictivo.

CAPÍTULO II: FUNDAMENTO TEÓRICO

2. Marco teórico

2.1. Churn de clientes

El Churn Rate, también conocido como tasa de cancelación de clientes, es una métrica que proporciona información sobre la cantidad de clientes que han dejado de hacer negocios con una empresa o compañía en un período de tiempo determinado (Canal, 2022).

Esta medida es fundamental para comprender el abandono de clientes y se calcula en función de un conjunto específico de características que se eligen y definen previamente. Esta selección puede variar según el análisis que se pretenda realizar, abarcando a todos los clientes, ya sean activos, en riesgo de cancelación o que ya han cancelado sus servicios.

Numerosas empresas emplean el Churn Rate como una herramienta crucial para evaluar si están experimentando pérdidas en su base de clientes y para identificar áreas críticas que requieren una acción inmediata para evitar una disminución en sus ingresos o en la satisfacción del cliente.

2.1.1. Cálculo de la tasa de cancelación

El Churn Rate se calcula dividiendo el número de clientes que han cancelado sus servicios durante un período específico entre el total de clientes que la empresa tenía al inicio de ese mismo período. Este cálculo proporciona un análisis detallado del Churn de

clientes y permite evaluar la tasa de abandono en relación con el tamaño total de la base de clientes (Canal, 2022).

La fórmula básica para calcular el Churn Rate es:

$$Churn\ Rate = \frac{Número\ de\ clientes\ que\ cancelaron}{Total\ de\ clientes\ al\ inicio\ del\ período} \times 100\%$$

Esta fórmula proporciona el porcentaje de clientes que cancelaron sus servicios durante un período de tiempo específico en relación con el número total de clientes al inicio de ese mismo período. Multiplicar el resultado por 100 convierte el valor en un porcentaje.

2.1.2. Estado del servicio

Para el cálculo del Churn Rate en la empresa proveedora del servicio de internet para hogares en el Ecuador, se consideran tres tipos de estado clave en el servicio:

Estado activo: Este estado hace referencia a todos los clientes que se encuentran con sus facturas al día, su contrato vigente y el servicio de internet totalmente operativo.

Estado cancelado: Este estado hace referencia a todos los clientes que han cancelado su servicio por diversas razones, ya sean voluntarias o administrativas.

Estado in-corte: Este estado hace referencia a todos los clientes con facturas impagas por más de 30 días. El Sistema Transaccional de Operaciones otorga un período adicional de 15 días para que el cliente realice el pago, si no se realiza el pago dentro de este plazo, el sistema corta automáticamente el servicio y el cliente es transferido al proceso de cancelación administrativa.

2.1.3. Tipo de cancelaciones

En la empresa se manejan dos tipos de cancelaciones del servicio:

Cancelación administrativa: Se refiere a la cancelación automática del servicio aplicado por el Sistema Transaccional de Operaciones cuando un cliente ha permanecido en estado in-corte durante 60 días.

Cancelación voluntaria: Ocurre cuando el cliente decide, por su propia voluntad, comunicar o realizar el proceso de cancelación del servicio de internet.

2.2. Aprendizaje automático

El aprendizaje automático, conocido en inglés como Machine Learning (ML), tiene sus raíces en los años 50, cuando Alan Turing creó el famoso "Test de Turing" y Arthur Samuel desarrolló el primer algoritmo capaz de aprender, utilizado en un programa de damas. En los años 60 y 70, se acuñó el término "Inteligencia Artificial" y se crearon algoritmos de reconocimiento de patrones, pero el progreso fue lento. Sin embargo, en los años 80, surgieron sistemas expertos basados en reglas que renovaron el interés en el Machine Learning, aunque esto fue seguido por un periodo de estancamiento en la década de 1980 (Nalda, 2024).

La verdadera revolución del Machine Learning comenzó a principios de los 2000, con desarrollos como el sistema de ficheros distribuidos de Google y el paradigma de procesamiento distribuido "Map & Reduce". El lanzamiento de Hadoop, una plataforma Big Data Open Source, marcó un hito importante. A medida que la potencia de cálculo y la disponibilidad de datos aumentaron exponencialmente, el Machine Learning se desarrolló espectacularmente, cambiando su enfoque de "knowledge-driven" a "data-driven".

Actualmente, estamos en la tercera explosión del Machine Learning, donde su aplicación en diversos sectores empresariales, como el asegurador, el marketing y los call centers, está transformando significativamente las estrategias (Nalda, 2024).

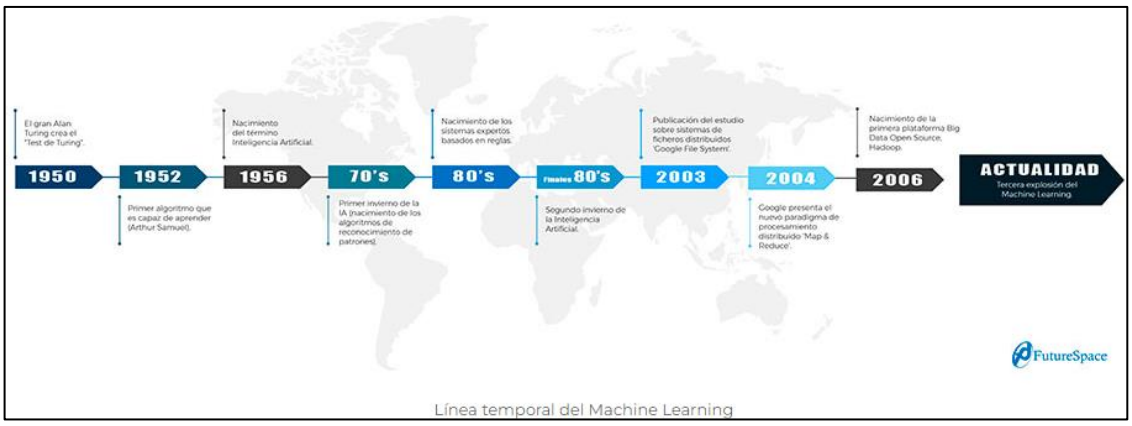


Figura 5: Línea Temporal del Machine Learning.

Fuente: (Nalda, 2024).

2.3. Tipos de aprendizaje automático

El aprendizaje automático está transformando el mundo a través de sus cuatro tipos principales. Estos tipos, que incluyen el aprendizaje supervisado, el no supervisado, el semisupervisado y el por refuerzo, abordan una amplia gama de problemas y aplicaciones, desde la regresión hasta la clasificación y el agrupamiento de datos. Cada una de estas categorías representa enfoques específicos que abordan diferentes aspectos y procesos dentro del ámbito del aprendizaje automático (Sánchez, 2020).

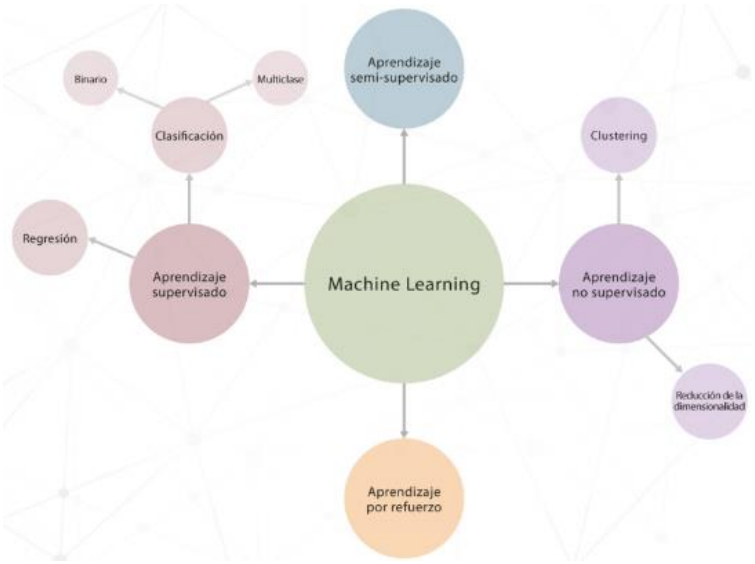


Figura 6: Tipos de Aprendizaje Automático.

Fuente: (Sánchez, 2020).

2.3.1. Aprendizaje supervisado

Este tipo de aprendizajes requiere conjuntos de datos etiquetados, lo que significa que se le indica al modelo qué debe aprender. Tomando como ejemplo, una heladería donde registran datos como clima, temperatura, día de semana, etc, junto con el número de helados vendidos por día en varios años. Para este caso, sería útil entrenar un modelo basado en los datos climáticos y otras características de un día en específico y pueda predecir la cantidad de helados que se venderán (Sánchez, 2020).

Dentro del aprendizaje supervisado, hay dos tipos de modelos dependiendo del tipo de etiqueta:

Modelos de clasificación: En este modelo se genera una etiqueta discreta a la salida, es decir, una etiqueta dentro de un conjunto finito de posibilidades. Son binarios, si la predicción resulta de clasificar entre dos clases o etiquetas (en salud, estar enfermo o no enfermo, un correo electrónico validar si es "spam" o no "spam"); o multiclase, en casos

que existen más de dos clases (por ejemplo, clasificar imágenes de mamíferos, análisis de emociones, etc.).

Modelos de regresión: Estos modelos producen un valor numérico real como salida, como el ejemplo mencionado anteriormente sobre la predicción de la cantidad de helados vendidos.

2.3.2. Aprendizaje no supervisado

El aprendizaje no supervisado se basa en datos no etiquetados, es decir, no tenemos una etiqueta específica que predecir. Este tipo de algoritmos se utilizan principalmente en la extracción de nuevos conocimientos o según su similitud agruparlos.

Una de las aplicaciones más comunes del aprendizaje no supervisado es el agrupamiento o clustering, donde el algoritmo define una métrica de similitud o distancia para comparar los datos entre sí y agruparlos en conjuntos afines. Esto puede ser útil para identificar grupos de clientes con características similares para campañas de marketing dirigidas, o para descubrir patrones ocultos en grandes conjuntos de datos, como en el análisis de información genómica (Sánchez, 2020).

Además, en ocasiones nos encontramos con conjuntos de datos que contienen una gran cantidad de características, lo que puede dificultar el procesamiento y análisis de los mismos. Para abordar este problema, se utilizan algoritmos de reducción de dimensionalidad, que aplican técnicas matemáticas y estadísticas para transformar el conjunto de datos original en uno nuevo con menos dimensiones, sacrificando un poco de información, pero facilitando el análisis y la visualización. Ejemplos de estos algoritmos

incluyen PCA (Análisis de Componentes Principales), t-SNE (t-Distributed Stochastic Neighbor Embedding) o ICA (Análisis de Componentes Independientes).

2.3.3. Aprendizaje semisupervisado

A menudo, resulta difícil obtener un conjunto de datos completamente etiquetado. Tomemos como ejemplo una empresa de productos lácteos que desea analizar la percepción de su marca a partir de los comentarios de los usuarios en redes sociales. Después de recopilar dieciséis mil mensajes que mencionan a la empresa, nos enfrentamos al desafío de que estos datos carecen de etiquetas, es decir, no sabemos si los comentarios son positivos, negativos o neutros.

Es en este punto donde entra en juego el aprendizaje semi-supervisado, que combina características del aprendizaje supervisado y no supervisado. En este enfoque, comenzamos etiquetando manualmente algunos de los comentarios. Con esta pequeña muestra de datos etiquetados, entrenamos uno o varios algoritmos de aprendizaje supervisado. Luego, utilizamos estos modelos para etiquetar automáticamente el resto de los comentarios.

Una vez que hemos etiquetado una cantidad significativa de datos utilizando estos modelos, entrenamos otro algoritmo de aprendizaje supervisado utilizando tanto los datos etiquetados manualmente como los generados por los modelos previos. De esta manera, logramos aprovechar la información proporcionada por los datos etiquetados manualmente y la capacidad de los algoritmos de aprendizaje supervisado para generalizar y etiquetar nuevos datos de manera eficiente (Sánchez, 2020).

2.3.4. Aprendizaje por refuerzo

El aprendizaje por refuerzo es una técnica del aprendizaje automático que se centra en recompensar los comportamientos deseables mientras penaliza los no deseados. Este método permite a un agente interpretar su entorno, tomar decisiones y aprender de las consecuencias de sus acciones a través de la prueba y el error. Su objetivo principal es maximizar una recompensa general a largo plazo para alcanzar una solución óptima (Sánchez, 2020).

Los juegos son uno de los entornos más comunes para aplicar el aprendizaje por refuerzo, como se ha demostrado en casos como AlphaGo o Pacman. En estos juegos, el agente recibe información sobre las reglas y aprende a jugar mediante la exploración y la experiencia. Aunque al principio puede actuar de manera aleatoria, con el tiempo desarrolla estrategias más sofisticadas.

Los juegos son uno de los entornos más comunes para aplicar el aprendizaje por refuerzo, como se ha demostrado en casos como AlphaGo o Pacman. En estos juegos, el usuario toma la información sobre las reglas y aprende a jugar mediante la exploración y la experiencia. Aunque al principio puede actuar de manera aleatoria, con el tiempo desarrolla estrategias más sofisticadas.

Además de los juegos, el aprendizaje por refuerzo ha sido aplicado en varias áreas como en la robótica, la optimización de recursos y en sistemas de control. En cada caso, el objetivo es utilizar la retroalimentación del entorno para mejorar el rendimiento del agente.

En síntesis, el aprendizaje automático es una de las herramientas más fuertes ya que convierte data en información útil y posterior se llega a una toma de decisiones. La elección del enfoque de aprendizaje adecuado depende del objetivo específico y las

características del conjunto de datos disponible. Es fundamental definir claramente el objetivo del aprendizaje para seleccionar el método más apropiado y obtener una solución que satisfaga las necesidades del problema en cuestión.

2.4. Herramientas

2.4.1. Alteryx

Es una herramienta multidimensional que permite preparar conjuntos de datos para obtener modelos predictivos, abarcando desde la limpieza de datos hasta su análisis. Facilita la integración y mapeo de datos de diferentes fuentes y ofrece diversas opciones operativas. Permite realizar modificaciones estándar mediante fórmulas a nivel de fila y operar tanto vertical como horizontalmente para análisis predictivos y espaciales. Además, posibilita la creación de aplicaciones analíticas avanzadas, elevando nuestra capacidad de preparación y análisis de datos (Lab, 2021).

¿Para qué sirve alteryx?

Alteryx ayuda en procesos de digitalización, simplificando la aplicación de analítica avanzada, sistemas predictivos, análisis de negocio y geográfico, o la administración y preparación de datos, entre otras potencialidades. Esta herramienta automatiza los procesos analíticos y, gracias al aprendizaje automático, genera modelos predictivos de manera rápida y sencilla, simplificando la toma de decisiones y mejorando los resultados comerciales (solutions, 2024).

2.4.2. Qlik sense

Qlik Sense es una plataforma de visualización y descubrimiento de datos que permite integrar diversas fuentes de datos y crear cuadros de mando interactivos, gráficos personalizados e informes detallados. Con Qlik Sense, los usuarios pueden visualizar datos de manera intuitiva y obtener información valiosa para una toma de decisiones informada (NUNSYS, 2024).

¿Para qué sirve qlik sense?

Qlik Sense ayuda a las organizaciones a comprender sus datos, tomar decisiones informadas y fomentar una cultura empresarial basada en datos. Sus principales funciones incluyen:

- **Análisis de datos:** Permite a los usuarios explorar y entender grandes volúmenes de datos rápidamente.
- **Toma de decisiones:** Facilita decisiones informadas mediante el acceso a información relevante y actualizada en tiempo real.
- **Visualizaciones personalizadas:** Permite crear visualizaciones, cuadros de mando y aplicaciones adaptadas a las necesidades de la empresa.
- **Exploración en tiempo real:** Los usuarios pueden navegar por los datos y ver resultados actualizados en tiempo real.
- **Análisis sin restricciones:** Elimina las limitaciones preconcebidas sobre la relación de datos, permitiendo un análisis más libre y completo.
- **Interfaz intuitiva:** Su función de arrastrar y soltar permite crear modelos analíticos sin necesidad de programación.

2.4.3. Python

Python es un lenguaje de programación ampliamente utilizado en el desarrollo de aplicaciones de aprendizaje automático debido a sus numerosas bibliotecas especializadas, como TensorFlow, Keras, y scikit-learn, que facilitan la implementación de algoritmos complejos. Su sintaxis sencilla y clara hace que sea accesible tanto para principiantes como para expertos, promoviendo una rápida iteración y experimentación. Al ser un lenguaje interpretado, permite modificar y probar código rápidamente sin necesidad de compilar, lo que es particularmente valioso en entornos de investigación y desarrollo donde la flexibilidad y la eficiencia son cruciales (Python, 2023).

¿Pará que sirve python?

Python es un lenguaje de programación versátil que se utiliza en una amplia variedad de aplicaciones, incluyendo:

- **Desarrollo web:** Python se usa para crear sitios web y aplicaciones web dinámicas, gracias a frameworks como Django y Flask.
- **Desarrollo de software:** Es utilizado para crear una amplia gama de software, desde aplicaciones de escritorio hasta sistemas de gestión empresarial.
- **Análisis de datos:** Python es ampliamente utilizado en ciencia de datos y análisis de datos debido a bibliotecas como Pandas, NumPy y SciPy, que proporcionan herramientas poderosas para manipular y analizar datos.
- **Aprendizaje automático e inteligencia artificial:** Python destaca con bibliotecas como TensorFlow, Keras y PyTorch que permiten la implementación de algoritmos de aprendizaje automático y la creación de modelos de inteligencia artificial.

- **Automatización y scripting:** Python es ideal para escribir scripts y automatizar tareas repetitivas, como procesamiento de archivos, administración de sistemas y tareas de mantenimiento.

2.5. Metodología CRISP-DM en la creación de modelos de aprendizaje automático

CRISP-DM, que significa "Cross-Industry Standard Process for Data Mining", es una metodología estándar desarrollada para guiar el proceso de minería de datos y la creación de modelos de aprendizaje automático. Su origen se remonta a mediados de la década de 1990, cuando un consorcio europeo de empresas, liderado por NCR Corporation, Daimler-Benz, SPSS, y OHRA, colaboró para estandarizar el enfoque del proceso de minería de datos.

El desarrollo de CRISP-DM comenzó oficialmente en 1996 y, tras una serie de talleres y pruebas con empresas de diversos sectores, se publicó la primera versión en 1999. La metodología rápidamente ganó aceptación debido a su flexibilidad y aplicabilidad en múltiples industrias, proporcionando un marco estructurado que facilita la gestión de proyectos de minería de datos desde su inicio hasta la implementación final (Kimball, 2013).

Entre varias metodologías de minería de datos se puede observar una similitud, pero es CRISP–DM la que se distingue por abarcar la aplicación de los resultados en el ámbito empresarial más completo.

Este es un modelo utilizado para el procesamiento en minería de datos, el cual detalla el abordaje de un problema. Para implementar la tecnología en los negocios, se necesita una metodología estructurada. Estos métodos generalmente se basan en la

experiencia personal y en los procedimientos estándar más conocidos. Entre ellos, CRISP-DM ha recibido un gran apoyo por parte de empresas privadas y organismos gubernamentales (Beatriz.Gil, 2019).

CRISP-DM se basa en un manual dividido en seis fases, algunas de las cuales son iterativas, lo que permite regresar a fases anteriores para revisarlas. Por lo tanto, el orden de las fases no es estrictamente lineal. La figura 7 muestra las fases de CRISP-DM y las posibles secuencias entre ellas.

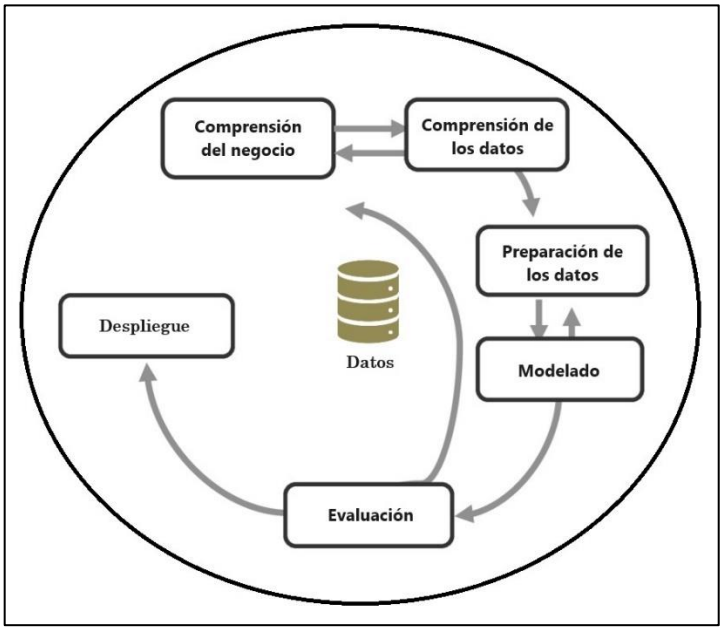


Figura 7: Metodología Crisp-Dm.
Fuente: (Claudia L. Hernández G, 2009).

En la figura 7, el círculo externo representa el ciclo de un proyecto de análisis de datos. De acuerdo con este método, el plan incluye los procesos de verificación y seguimiento. CRISP-DM enfoca el análisis de datos desde una perspectiva profesional que abarca un contexto más amplio, lo cual impacta en la creación del modelo.

Este enfoque tiene en cuenta la participación de clientes externos al equipo de desarrollo y entiende que el proyecto no termina al alcanzar un modelo óptimo; más bien,

es necesario llevar a cabo la implementación y el mantenimiento continuo del modelo (Claudia L. Hernández G, 2009).

Además, los resultados y conocimientos obtenidos son relevantes para otros proyectos, por lo que deben documentarse minuciosamente para que otros equipos de desarrollo puedan utilizarlos y beneficiarse de ellos (Claudia L. Hernández G, 2009).

2.5.1. Fases de la metodología CRISP-DM

I. Comprensión del negocio:

Propósito principal: Alinear los objetivos del proyecto de minería de datos con los objetivos comerciales. Por tanto, se trata de no iniciar proyectos de minería de datos que no tengan un impacto significativo dentro de la organización (Pete Chapman, 2007).

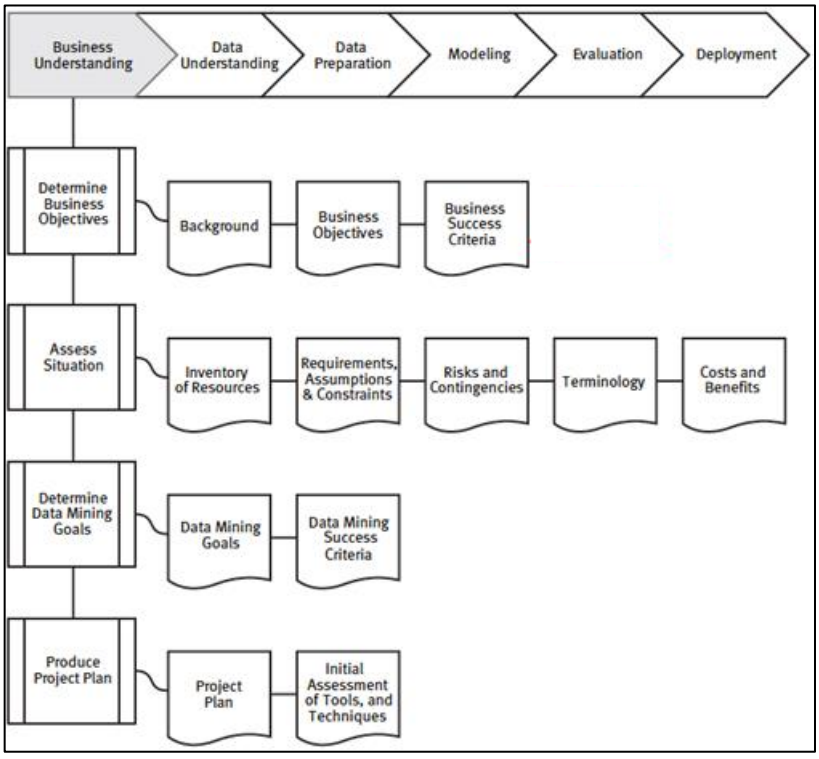


Figura 8: Fase de Comprensión del Negocio.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Perspectiva del negocio:** Comprender completamente el contexto y los objetivos de la organización.
- **Establecer objetivos del negocio:** Definir claramente los objetivos comerciales que se desean alcanzar.
- **Evaluación de la situación actual:** Analizar la situación actual de la compañía, incluyendo inventarios, fuentes de datos propias y terminologías adaptadas al negocio.
- **Creación del plan del proyecto:** Elaborar un plan detallado que incluya hitos, tareas y actividades necesarias para lograr los objetivos establecidos.

Proceso:

- **Reuniones colaborativas:** Iniciar reuniones con los encargados de las áreas implicadas en el proyecto para asegurar una comprensión común y establecer objetivos claros.
- **Análisis y documentación:** Documentar la situación actual de la empresa y los objetivos comerciales.
- **Definición de objetivos:** Clarificar cómo los objetivos comerciales se traducen en objetivos específicos para el análisis de datos.
- **Planificación del proyecto:** Desarrollar un plan de proyecto que incluya:
  - ✓ Hitos principales.
  - ✓ Tareas y actividades específicas.
  - ✓ Recursos necesarios.
  - ✓ Cronograma de implementación.

Impacto en el proyecto:

Si esta fase no se realiza correctamente, el resto del proyecto puede perder relevancia o no cumplir con las expectativas iniciales.

II. Comprensión de los datos:

Propósito principal: Identificar, recolectar y comprender los datos necesarios para el análisis, asegurando que sean relevantes y de alta calidad para alcanzar los objetivos del proyecto (Pete Chapman, 2007).

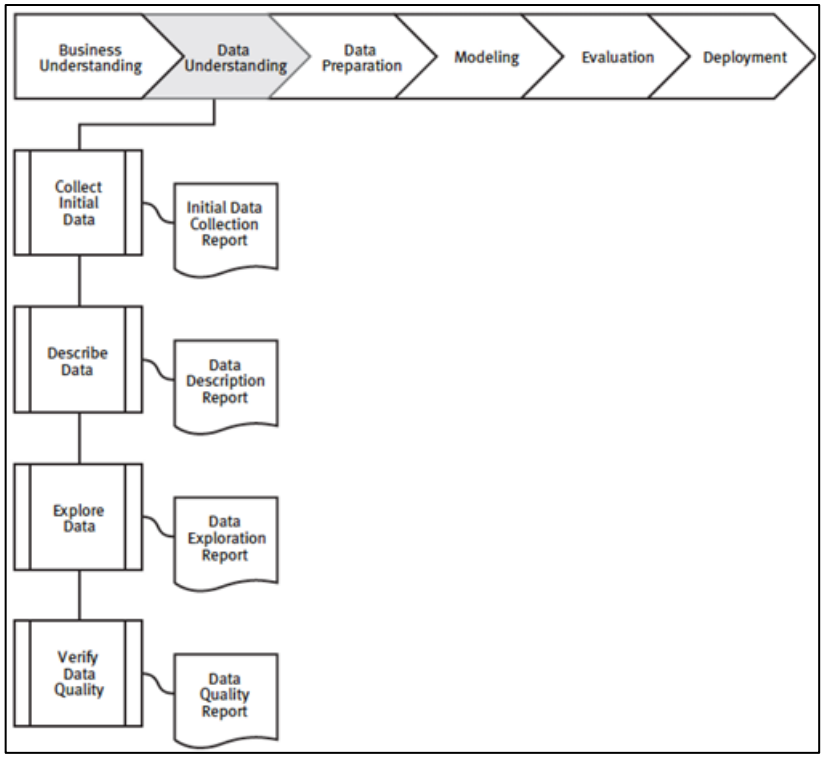


Figura 9: Fase de Comprensión de los Datos.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Identificación de fuentes de datos:** Determinar las fuentes de datos disponibles y relevantes para el proyecto.

- **Recolección de datos:** Recopilar datos desde diversas fuentes, asegurando la integridad y la exhaustividad de la información.
- **Descripción de los datos:** Proporcionar una descripción detallada del conjunto de datos, incluyendo su estructura, formato y contenido.
- **Exploración de datos:** Realizar análisis exploratorios para identificar patrones, tendencias y posibles anomalías en los datos.
- **Gestión de la calidad de los datos:** Evaluar y gestionar la calidad de los datos, detectando problemas y ofreciendo soluciones efectivas.

Importancia de la fase:

- **Comprensión del problema:** Es fundamental entender completamente el problema que se intenta resolver para recopilar datos exactos y relevantes.
- **Aplicación de conceptos comerciales:** Aplicar el concepto del negocio a la extracción de información y así asegurar el alineamiento de los objetivos comerciales con el análisis de datos.
- **Planificación estratégica:** Crear un plan para la guía del proceso de minería de datos.

Durante esta fase, es crucial asegurar que todos los datos relevantes sean considerados y que se comprendan los requisitos del negocio ya que esta estructura garantiza la comprensión de los datos estableciendo una base sólida para el éxito del proyecto de minería de datos.

III. Preparación de los datos:

Propósito principal: Seleccionar, limpiar y preparar conjuntos de datos adecuados para la fase de modelado. Este paso es crucial para asegurar que los datos sean precisos, completos y listos para el análisis (Pete Chapman, 2007).

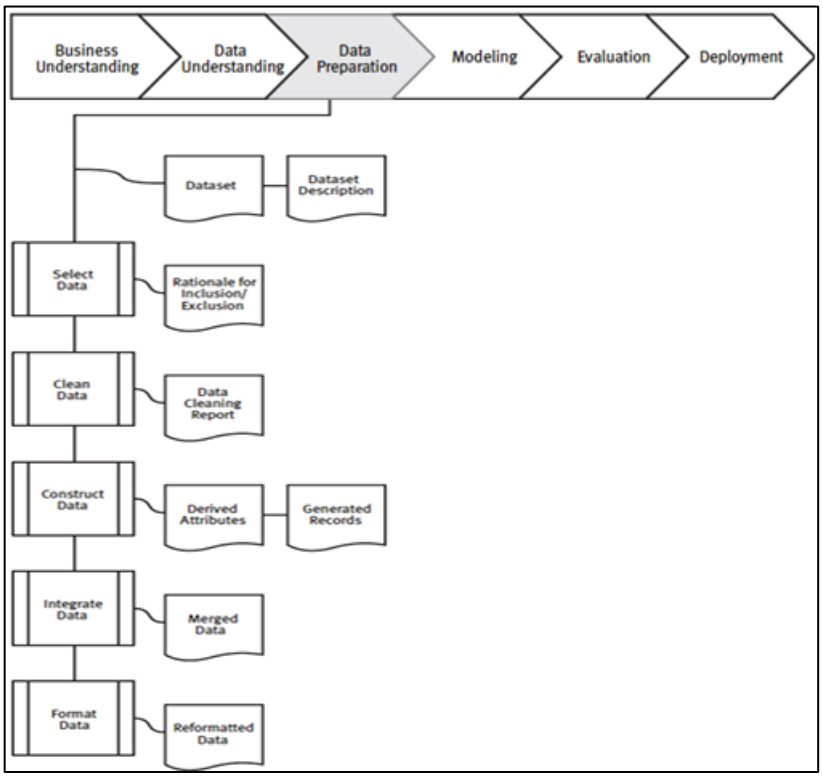


Figura 10: Fase de Preparación de los Datos.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Establecer el conjunto de datos:** Determinar y seleccionar los datos que se utilizarán en el proyecto.
- **Limpieza de datos:** Realizar actividades de limpieza de datos para eliminar errores, valores faltantes y duplicados.

- **Construcción del conjunto de datos:** Crear un conjunto de datos idóneo y bien estructurado que se utilizará en los modelos de minería de datos.
- **Integración de datos:** Combinar datos de varias fuentes si es necesario, asegurando coherencia y compatibilidad.
- **Formateo de datos:** Cambiar el formato de los datos si es necesario para que sean compatibles con las técnicas de modelado.

Importancia de la fase:

Esta etapa es crítica para el éxito del proyecto de minería de datos. Es importante tener precisión en esta fase teniendo preparados los datos correctamente y así evitar alargar los tiempos de ejecución del proyecto.

IV. Modelado:

Propósito principal: Generar modelos de conocimiento basados en los datos obtenidos en la etapa previa, adaptándolos a los objetivos del proyecto mediante técnicas específicas como la clasificación y la regresión (Pete Chapman, 2007).

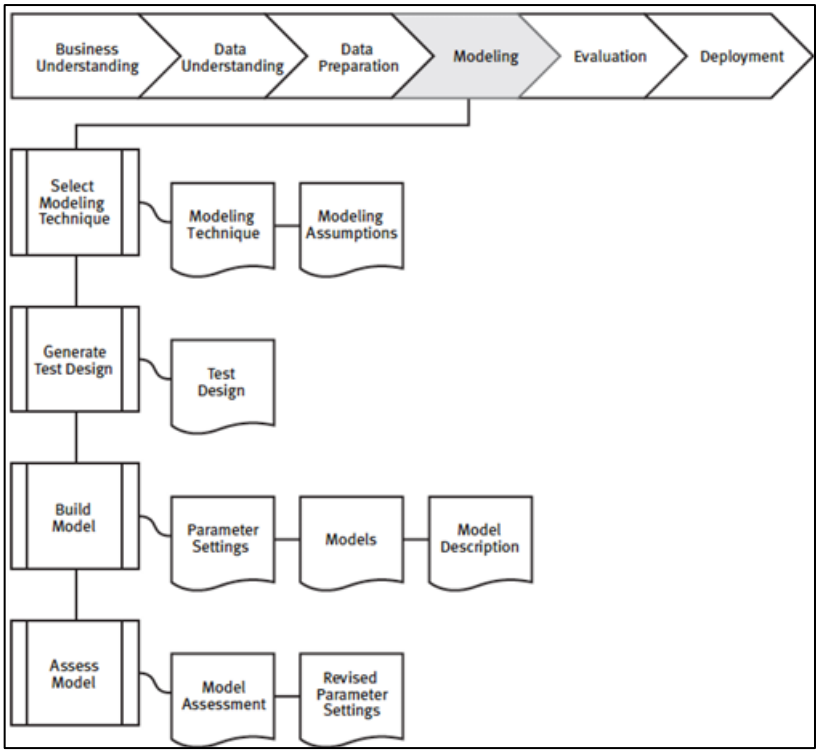


Figura 11: Fase de Modelado.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Selección de métodos de modelado:** Seleccionar los métodos de modelado que mejor se adapten a cumplir los objetivos del negocio con los datos disponibles.
- **Control de calidad del modelo:** Establecer una metodología para garantizar la calidad y precisión del modelo.
- **Creación del modelo:** Aplicar los métodos seleccionados al conjunto de datos para crear el modelo.
- **Evaluación y ajuste del modelo:** Evaluar la solidez del modelo y su impacto en los objetivos predeterminados, realizando ajustes según sea necesario.

Importancia de la fase:

- **Determinación del tipo de modelo:**
  - ✓ El tipo de modelo utilizado se selecciona en función de las necesidades del negocio y del tipo de variables que se analizan.
  - ✓ Es crucial especificar claramente los atributos y variables que se utilizarán en el modelo.
- **Iteración y retroalimentación:**
  - ✓ La repetición de esta etapa asegura que el modelo sea robusto y preciso, adaptándose a las condiciones y requerimientos del proyecto.

V. Evaluación:

Propósito principal: Evaluar el modelo frente a los criterios de éxito definidos y validar su efectividad y precisión para garantizar que cumpla con los objetivos comerciales (Pete Chapman, 2007).

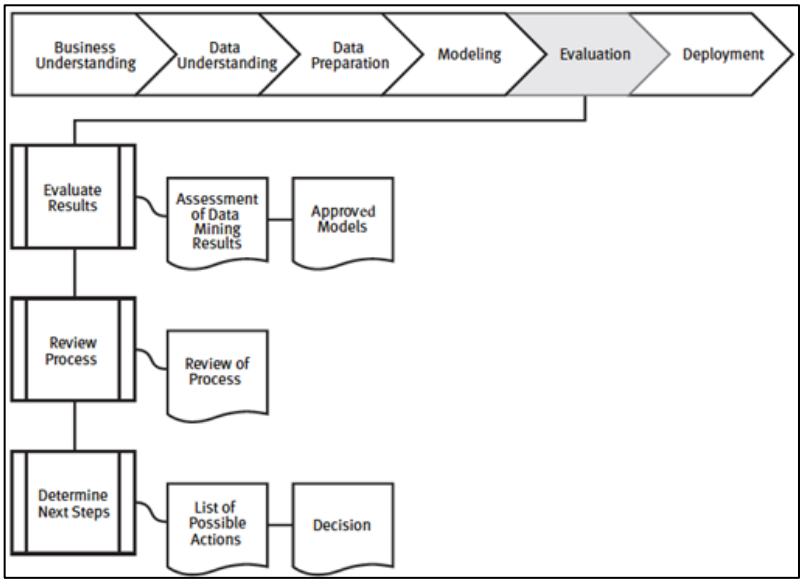


Figura 12: Fase de Evaluación.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Evaluación del modelo:** Analizar y evaluar el rendimiento de los modelos generados hasta el momento.
- **Revisión completa del proceso:** Revisar el proceso de minería de datos para identificar posibles errores o áreas de mejora.
- **Planificación de los próximos pasos:** Determinar los siguientes pasos a seguir, incluyendo la repetición de fases anteriores o la incorporación de nuevos temas de investigación si es necesario.

Importancia de la fase:

- **Validación de resultados:** Esta fase es crucial para validar los resultados obtenidos y asegurarse de que el modelo cumple con los objetivos comerciales.
- **Iteración para la mejora:** Proporciona una oportunidad para iterar y mejorar el modelo, aumentando la probabilidad de éxito del proyecto.
- **Documentación y aprendizaje:** Documentar los hallazgos y las lecciones aprendidas durante esta fase es esencial para el aprendizaje continuo y la mejora de futuros proyectos.

VI. Despliegue:

Propósito principal: Implementar, monitorear y mantener los modelos de datos en producción, asegurando que funcionen correctamente y aporten valor continuo al negocio (Pete Chapman, 2007).

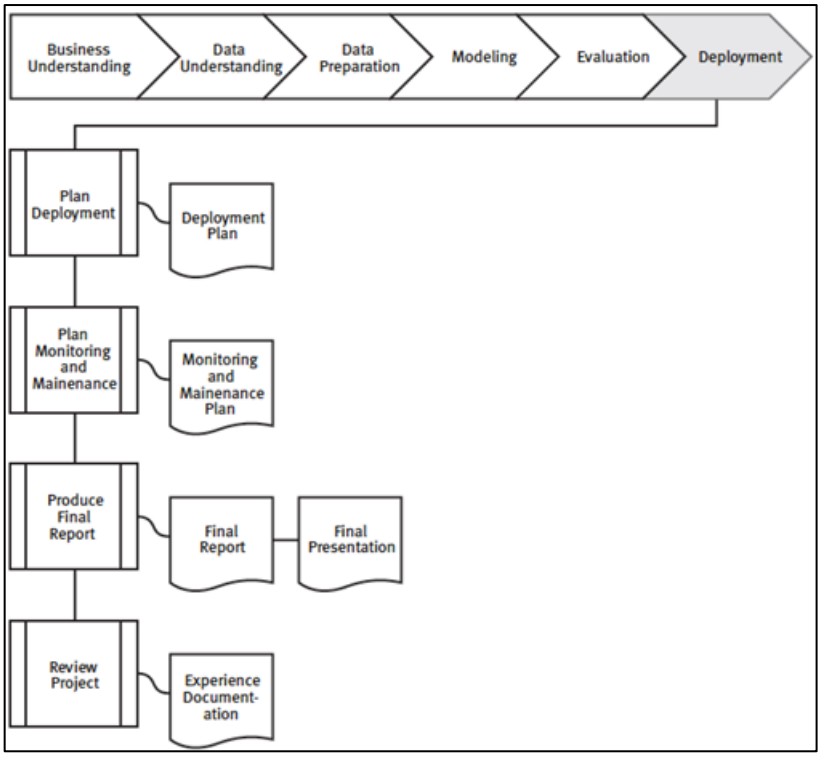


Figura 13: Fase de Despliegue.

Fuente: (Pete Chapman, 2007).

Objetivos a obtener:

- **Planificación del despliegue:** Diseñar un plan detallado para la implementación del modelo seleccionado en la organización.
- **Mantenimiento y seguimiento:** Establecer procesos para el mantenimiento operativo y el seguimiento continuo del modelo.
- **Revisión y documentación:** Revisar el proyecto en su totalidad, documentar las lecciones aprendidas y elaborar conclusiones y recomendaciones para futuros proyectos.

Importancia de la fase:

- **Implementación eficaz:** Una planificación adecuada del despliegue asegura que el modelo se integre eficazmente en la organización y comience a generar beneficios rápidamente.
- **Mantenimiento proactivo:** El mantenimiento y seguimiento continuo del modelo garantizan su funcionamiento correcto y minimizan el riesgo de fallos o errores.
- **Documentación y mejora:** Documentar las lecciones aprendidas y realizar un análisis exhaustivo del proyecto permite mejorar continuamente y aplicar estos conocimientos a futuros proyectos.

2.6. Modelos de predicción

Los modelos de predicción son herramientas analíticas que utilizan datos históricos y técnicas matemáticas o estadísticas para prever futuros eventos, comportamientos o valores. Estos modelos son fundamentales en diversos campos, como negocios, finanzas, medicina, ciencias sociales y más, ya que ayudan a tomar decisiones informadas basadas en las probabilidades y tendencias futuras.

Estos modelos se basan en la idea de que los patrones y tendencias identificados en los datos pasados pueden ayudar a predecir el comportamiento futuro (Mera, 2023).

La importancia de los modelos predictivos es que desempeñan papel fundamental al identificar oportunidades de mejora y prever eventos futuros en diversas áreas empresariales, como logística, previsión económica y ventas. Por ejemplo, en el ámbito del marketing, estos modelos ayudan a anticipar el comportamiento de los consumidores y sus decisiones de compra en función de sus acciones pasadas.

Entre las técnicas más utilizadas en modelos predictivos incluyen:

- **Árboles de decisión:** Permiten representar y analizar decisiones y sus posibles consecuencias.
- **Regresión lineal y logística:** Son utilizadas para modelar la relación entre variables independientes y dependientes.
- **Redes neuronales:** Modelan relaciones complejas entre entradas y salidas a través de múltiples capas de nodos.
- **Análisis bayesiano:** Utiliza la teoría de probabilidad para hacer predicciones basadas en evidencia previa.
- **Series temporales y datamining:** Se centran en el análisis de datos secuenciales y la extracción de patrones.
- **Máquinas de vectores de soporte (SVM):** Clasifican datos mediante la búsqueda de un hiperplano de separación óptimo.
- **K Nearest Neighbors (KNN):** Clasifica un punto de datos según la mayoría de sus vecinos más cercanos.
- **Gradient Boosting:** Combina múltiples modelos de bajo rendimiento para mejorar la precisión predictiva.
- **Respuesta incremental:** Se basa en el aprendizaje secuencial de modelos basados en nuevos datos.
- **Razonamiento con base en la memoria:** Utiliza experiencias pasadas para tomar decisiones futuras.
- **Regresión de mínimos cuadrados parciales:** Analiza la relación entre múltiples variables predictoras y una variable de respuesta.

La elección de la técnica adecuada depende de los objetivos específicos de la investigación y de la naturaleza de los datos disponibles. Para nuestro trabajo, se han seleccionado Regresión Logística, Árboles de Decisión y Random Forest debido a su idoneidad para predecir variables de interés en nuestro contexto de estudio.

2.6.1. Regresión logística

La regresión logística es un conjunto de técnicas estadísticas utilizadas para analizar la relación entre una variable dependiente cualitativa y una o más variables independientes, también conocidas como covariables. Esta técnica es especialmente útil cuando la variable dependiente es dicotómica (binaria) o tiene más de dos categorías (Jaén, 2019).

En regresión logística, hay tres clases principales:

- **Binaria:** En el caso de la regresión logística binaria o binomial, es un tipo de regresión logística utilizada cuando la variable de respuesta es binaria, es decir, tiene dos categorías. Se analiza una variable dependiente que puede tomar solo dos valores posibles, como "sí" o "no", "éxito" o "fracaso", etc.
- **Polinomial:** Es una extensión de la regresión logística binomial que permite modelar relaciones no lineales entre las variables predictoras y la variable de respuesta. Esto se logra al incluir términos polinomiales de las variables predictoras en el modelo. La regresión logística polinomial puede ser útil cuando la relación entre las variables no puede ser capturada de manera adecuada por un modelo lineal.

- **Multinomial:** Se utiliza cuando la variable de respuesta tiene más de dos categorías. En lugar de predecir la probabilidad de dos resultados posibles como en la regresión logística binomial, la regresión logística multinomial modela la probabilidad de ocurrencia de varias categorías. Es útil cuando la variable de respuesta es categórica y tiene tres o más niveles, como la clasificación de productos en varias categorías o la predicción de la preferencia de los consumidores entre múltiples opciones.

Las covariables pueden ser tanto cualitativas como cuantitativas, pero en el caso de las covariables cualitativas, es importante que sean dicotómicas, es decir, que puedan tomar valores de 0 o 1 para indicar su ausencia o presencia en el modelo. Si una covariable cualitativa tiene más de dos categorías, se realiza una transformación mediante la creación de variables ficticias (variables dummy) para cada categoría, de modo que cada una pueda ser incluida de forma individual en el modelo (Jaén, 2019).

Los modelos de regresión logística tienen tres finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, los odds ratio para cada covariable).
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente.

Variables dummy: Son variables binarias creadas para representar variables categóricas en un modelo de regresión logística. Estas variables se utilizan para codificar la presencia o ausencia de una categoría específica de una variable cualitativa.

2.6.2. Árboles de decisión

Los árboles de decisión son una técnica de minería de datos (Data Mining, DM) prepara, sondea y explora los datos para sacar la información oculta en ellos. Se aborda la solución a problemas de predicción, clasificación y segmentación (Vanesa Berlanga Silvente, 2013).

Los árboles de decisión son modelos versátiles que se utilizan tanto para clasificación, conocidos como árboles de clasificación, como para regresión, denominados árboles de regresión. Son ampliamente apreciados por su simplicidad y facilidad de interpretación, lo que los hace especialmente útiles para comparar resultados con hipótesis o preguntas de negocio. Su estructura visual y lógica transparente los convierte en una herramienta poderosa para analizar y tomar decisiones sobre el caso de estudio.

La estructura de un árbol de decisión (ver figura 14), es que cada nodo representa una regla de decisión y se aplica al elemento clasificado. Si se cumple la regla, el elemento se dirige por una de las flechas hacia una categoría o subgrupo adecuado, o hacia un nuevo nodo que contiene otra regla de clasificación. Es así como, el primer cuadrado representa el nodo inicial y las flechas indican las posibles direcciones basadas en las reglas de decisión (Chávez, 2018).

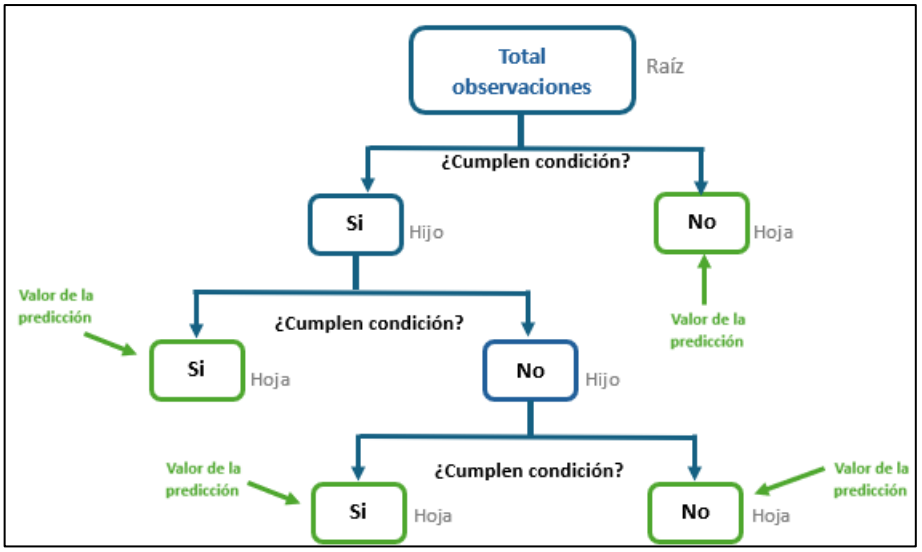


Figura 14: Flujograma de un árbol de decisión.

Fuente: (Chávez, 2018).

Los árboles de decisión constan de tres tipos de nodos:

- **Nodo raíz:** Es el primer nodo y es el que establece la división inicial que se basa en el campo principal de los datos.
- **Nodos internos o intermedios:** Después de la creación del nodo raíz, los nodos se dividen nuevamente en base a diferentes grupos de datos definidos por sus variables.
- **Nodos terminales (hojas):** Se encuentran en el final del diagrama y marcan el término de la clasificación, es decir, representan las categorías finales a las que se asignan los datos.

Es importante considerar la profundidad del árbol, que está determinada por el número máximo de nodos en cada rama. Una mayor profundidad puede aumentar la complejidad del modelo y su capacidad para capturar relaciones más sutiles en los datos, pero también puede aumentar el riesgo de sobreajuste.

Los árboles de clasificación son especialmente útiles porque:

- Son fáciles de interpretar y visualizar, lo que permite comprender todas las decisiones tomadas para llegar a los resultados.
- Funcionan correctamente con variables categóricas y en el manejo de valores atípicos presentan mayor robustez. Sin embargo, poseen como limitación el requerir de una base de datos bien entrenada y ejecutada.

2.6.3. Random forest

El algoritmo Random Forest es una herramienta común en el aprendizaje automático, desarrollada por Leo Breiman y Adele Cutler, que aprovecha la potencia de múltiples árboles de decisión para generar una predicción precisa. Este método es altamente flexible y fácil de usar, lo que ha contribuido a su popularidad en una amplia gama de problemas de clasificación y regresión (IBM, 2024).

Los métodos de aprendizaje por conjuntos, como Random Forest, funcionan mediante la combinación de las predicciones de múltiples clasificadores, como árboles de decisión, para identificar el resultado más común o promedio. Dos métodos de conjunto bien conocidos son el ensacado (bagging) y el impulso (boosting). El ensacado, introducido por Leo Breiman en 1996, implica la selección aleatoria de muestras de datos con reemplazo, lo que permite que los puntos de datos individuales se seleccionen más de una vez.

El algoritmo Random Forest extiende el método de ensacado incorporando aleatoriedad en la selección de características. Esta aleatoriedad garantiza que los árboles de decisión en el bosque estén poco correlacionados entre sí, lo que mejora la diversidad y

la precisión del modelo. A diferencia de los árboles de decisión tradicionales, que consideran todas las divisiones de características posibles, los Random Forest solo seleccionan un subconjunto aleatorio de características en cada nodo de división.

Al limitar la complejidad de cada árbol individual y diversificar la selección de características, Random Forest reduce el riesgo de sobreajuste y mejora la generalización del modelo. Esto se traduce en predicciones más precisas y robustas, incluso en conjuntos de datos ruidosos o con alta variabilidad (IBM, 2024).

Los algoritmos de Random Forest funcionan con tres hiperparámetros clave: el tamaño del nodo, la cantidad de árboles y la cantidad de características muestreadas, que se configuran antes del entrenamiento del modelo. Este clasificador puede utilizarse para resolver tanto problemas de regresión como de clasificación.

El algoritmo de Random Forest está constituido por un conjunto de árboles de decisión (ver figura 15), cada árbol está implementado a partir de una muestra extraída de un grupo de entrenamiento con reemplazo, también conocida como muestra de arranque. Además, se utiliza una muestra fuera de la bolsa (OOB) para la validación cruzada, lo que mejora la precisión del modelo. Para agregar más diversidad al conjunto de datos y reducir la correlación entre los árboles, se inyecta aleatoriedad mediante el agrupamiento de características.

En función del tipo de problema (regresión o clasificación), se determina la predicción final: para la regresión, se promedian las predicciones de los árboles individuales, mientras que, para la clasificación, se utiliza un voto mayoritario para determinar la clase predicha.

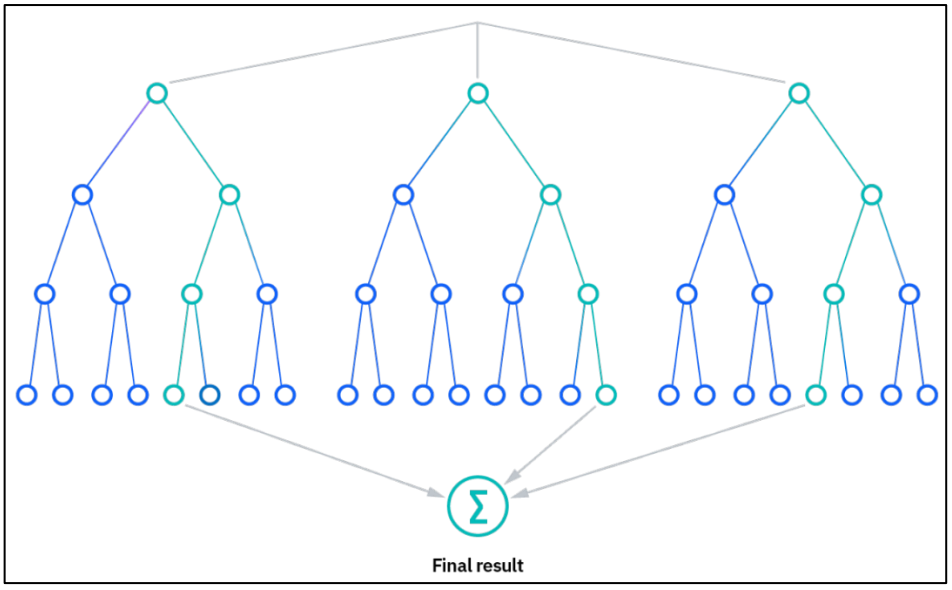


Figura 15: Flujo de Random Forest.

Fuente: (IBM, 2024).

2.7. Evaluación de desempeño del algoritmo

Es esencial definir las métricas a utilizar y distinguir cuál tiene un mejor desempeño para el objetivo establecido para el análisis predictivo del churn de clientes. Se implementará un enfoque riguroso que incluye la división de los datos y la selección de métricas de evaluación.

División de los datos: Para desarrollar y evaluar el modelo de predicción de churn, la base de datos será dividida utilizando el método Train-Test Split en un 70-30% para entrenamiento y prueba respectivamente. Este enfoque garantiza que el modelo sea entrenado con una parte de los datos (70%) y probado con información que nunca se ha ingresado en los modelos (30%), asegurando una previsibilidad real y evitando que los modelos simplemente memoricen la información de entrada.

La técnica de prueba dividida de entrenamiento (**Train-Test Split**), implica en dividir aleatoriamente el conjunto de datos en dos partes. La primera parte, que representa

entre el 70 % y el 80 % de los datos, se utiliza para entrenar el modelo de aprendizaje automático. La segunda parte, que comprende entre el 20 % y el 30 % restante de los datos, se utiliza para la verificación y evaluación del modelo.

Métricas de evaluación: Dado que el problema a abordar es de clasificación binaria (clientes activos vs. clientes en riesgo de churn), se utilizarán las siguientes métricas de evaluación:

- **Accuracy (Exactitud):** La proporción de predicciones correctas entre el total de predicciones realizadas. Esta métrica proporciona una visión general de la efectividad del modelo.

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Total\ de\ predicciones}$$

- **Precision (Precisión):** La proporción de verdaderos positivos entre el total de positivos predichos. Indica la exactitud de las predicciones positivas.

$$Precisión = \frac{Verdaderos\ Positivos\ (TP)}{Verdaderos\ Positivos\ (TP) + Falsos\ Positivos\ (FP)}$$

- **Recall (Sensibilidad):** La proporción de verdaderos positivos entre el total de positivos reales. Indica la capacidad del modelo para identificar correctamente los casos de churn.

$$Sensibilidad = \frac{Verdaderos\ Positivos\ (TP)}{Verdaderos\ Positivos\ (TP) + Falsos\ Negativos\ (FN)}$$

- **F1 score:** La media armónica de la precisión y la sensibilidad, proporcionando una única métrica que equilibra ambas.

$$F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad}$$

- **AUC-ROC (Área bajo la curva - Receiver Operating Characteristic):** Una medida que evalúa el desempeño del modelo en diferentes umbrales de clasificación. Un valor más alto indica un mejor desempeño del modelo.

$$AUC - ROC = \int_0^1 ROC(t)dt$$

Validación cruzada: Para nuestro estudio no se aplicará este método debido a que aplicaremos una técnica de evaluación y selección de variables predictoras llamado WOE (Weight of Evidence) + IV (Information Value) el cual será explicado en su momento cuando entremos a trabajar sobre la Preparación de los Datos específicamente con categorización de variables.

Comparación de modelos: Se probarán diferentes técnicas de aprendizaje automático, tales como:

- Regresión Logística
- Árboles de Decisión
- Bosques Aleatorios (Random Forest)

Cada modelo será evaluado utilizando las métricas mencionadas y los resultados serán comparados para identificar el modelo con el mejor desempeño en términos de precisión, sensibilidad, F1 score y AUC-ROC.

Interpretabilidad del modelo: Además del desempeño, será evaluada la interpretabilidad del modelo, es decir, la capacidad de entender cómo el modelo toma sus

decisiones. Modelos como los árboles de decisión y los bosques aleatorios serán interpretados mediante:

- **Importancia de características:** Evaluación de las características más relevantes para las predicciones.
- **Implementación del dashboard:** Un dashboard como herramienta de análisis y control de información será desarrollado utilizando herramientas como Qlik Sense para visualizar.

Recomendaciones basadas en los resultados: Finalmente, se proporcionarán recomendaciones para estrategias de retención de clientes obtenidos del modelo predictivo y las métricas de evaluación. Estas estrategias se centrarán en los factores más influyentes en la decisión de churn, ayudando a la empresa a implementar acciones proactivas para retener a sus clientes.

Consideraciones: De lo anteriormente expuesto, es importante definir y conocer brevemente conceptos sobre la Curva ROC y la Matriz de Confusión que son de suma importancia para nuestro trabajo.

2.7.1. Curva ROC (AUC-ROC)

La curva ROC (Receiver Operating Characteristic) y el área bajo la curva (AUC) son herramientas comunes en el análisis de modelos de clasificación en el aprendizaje automático. La curva ROC es un gráfico que representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación. TPR se refiere a la proporción de casos positivos que fueron correctamente identificados como positivos, mientras que FPR es la proporción de casos negativos

incorrectamente identificados como positivos. La curva ROC muestra cómo cambia el equilibrio entre TPR y FPR a medida que se ajusta el umbral de clasificación del modelo.

El AUC es el área bajo la curva ROC y proporciona una medida agregada del rendimiento del modelo en todos los umbrales de clasificación posibles. Cuanto mayor sea el AUC, mejor será el rendimiento del modelo para discriminar entre las clases positivas y negativas (ver figura 17). Un AUC de 1 indica un modelo perfecto que clasifica todas las instancias correctamente, mientras que un AUC de 0.5 sugiere un rendimiento similar al azar (Narkhede, 2018).

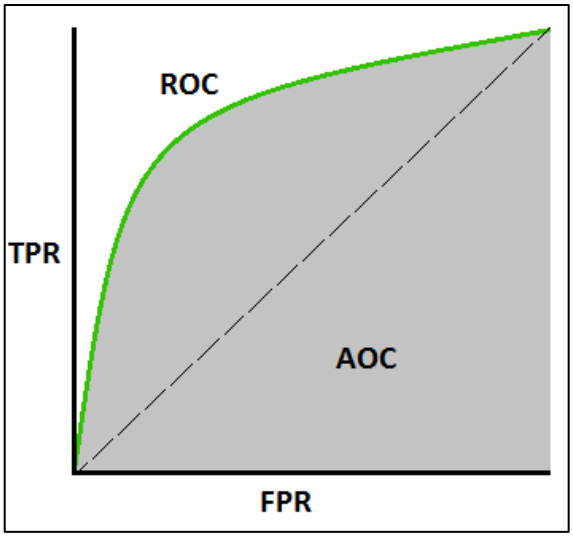


Figura 16: Curva ROC-AUC.
Fuente: (Narkhede, 2018).

La grafica establece la tasa de verdaderos positivos (TPR) en el eje Y (vertical) y la tasa de falsos positivos (FPR) en el eje X (horizontal). Cada punto en la curva representa un umbral diferente. La diagonal de referencia representa un clasificador aleatorio y se extiende desde el origen (0,0) hasta el punto (1,1). Este clasificador no tiene capacidad de discriminación y su curva ROC sería una línea diagonal.

AUC: Calcula el área bajo la curva ROC (AUC) para evaluar el rendimiento general del modelo. Cuanto mayor sea el AUC, mejor será el rendimiento del modelo en términos de discriminación entre clases positivas y negativas.

Es importante destacar que una curva ROC ideal se acercaría lo más posible al vértice superior izquierdo del gráfico, lo que indicaría una alta tasa de verdaderos positivos y una baja tasa de falsos positivos en todos los umbrales de clasificación.

2.7.2. Matriz de confusión

La matriz de confusión es una herramienta que se utiliza en el análisis de modelos de clasificación para evaluar el rendimiento del modelo en la predicción de las clases verdaderas de los datos. Esta matriz muestra un resumen de las predicciones realizadas por el modelo en comparación con las clases reales de los datos. Consiste en una tabla que resume las predicciones del modelo en comparación con los valores reales de las instancias.

Cada columna de la matriz representa las predicciones realizadas por el modelo para cada clase, a su vez cada fila muestra el número real de instancias pertenecientes a cada clase. De esta manera, la matriz nos permite identificar los aciertos y errores del modelo en la clasificación de las instancias.

La matriz de confusión también se conoce como matriz de error. Esta tabla proporciona un resumen conciso de las predicciones correctas e incorrectas del modelo, desglosadas por cada clase, lo que facilita la evaluación y medición del rendimiento del modelo de clasificación (Shin, 2020).

Una matriz de confusión típicamente tiene cuatro celdas, que representan cuatro posibles resultados de una predicción:

- **Verdaderos positivos (TP):** Instancias que fueron correctamente clasificadas como positivas por el modelo.
- **Verdaderos negativos (TN):** Instancias que fueron correctamente clasificadas como negativas por el modelo.
- **Falsos positivos (FP):** Instancias que fueron incorrectamente clasificadas como positivas por el modelo (también conocidos como errores de Tipo I).
- **Falsos negativos (FN):** Instancias que fueron incorrectamente clasificadas como negativas por el modelo (también conocidos como errores de Tipo II) (González, 2019).

La siguiente figura, representa la disposición típica de una matriz de confusión:

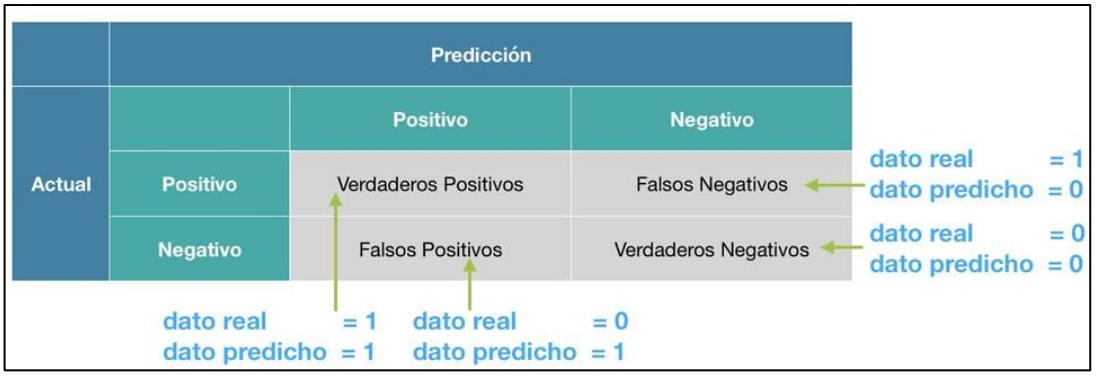


Figura 17: Matriz de Confusión.

Fuente: (González, 2019).

Con esta información, se pueden calcular diversas métricas de evaluación del rendimiento del modelo, como la precisión, la sensibilidad, la especificidad y el valor F1. Estas métricas proporcionan información detallada sobre la capacidad del modelo para realizar predicciones precisas en cada clase y pueden ser útiles para ajustar y mejorar el modelo.

CAPÍTULO III: DESARROLLO DEL MODELO

3. Desarrollo del modelo predictivo basado en la metodología CRISP-DM

Para el cumplimiento de los objetivos que se definieron en el presente trabajo de titulación se aplicará la metodología CRISP-DM debido a que es una adaptación utilizada en ciencia de datos conocido como "Descubrimiento del Conocimiento en Base de Datos" (KDD), esta consiste en ejecutar en orden una serie de pasos y así extraer información contundente de una base de datos

A continuación, se describirán los seis pasos que se aplicarán en el presente proyecto.

3.1. Comprensión del negocio

En la primera fase, se establece una conexión directa entre los aspectos operativos del negocio de la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador y los elementos técnicos necesarios para lograr los objetivos establecidos. Este enfoque se fundamenta en los criterios previamente identificados. Por lo tanto, el proceso de gestión y desarrollo del modelo se llevará a cabo con el fin de abordar eficazmente el desafío de reducir el churn de clientes, proporcionando una explicación clara de los factores relevantes y garantizando la máxima utilidad práctica del proyecto mediante la selección cuidadosa de variables pertinentes.

3.1.1. Perspectiva del negocio

La Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador tiene como objetivo ofrecer soluciones tecnológicas innovadoras de alto desempeño en su categoría, de tal manera que pueda satisfacer las necesidades de sus clientes en la conectividad mejorando así la calidad de vida de sus abonados. Para lograrlo, debe atraer a nuevos clientes interesados en el servicio, pero sobre todo, mantener a los clientes que ya forman parte como usuarios de la marca. Es por ello, que la empresa requiere de una herramienta que permita controlar a sus clientes y así evitar su deserción, por lo tanto, es importante utilizar un modelo de aprendizaje automático el cual se convierta en un instrumento de apoyo ya que mediante la predicción permitirá enmendar estas necesidades.

3.1.2. Objetivos del negocio

Desarrollar un algoritmo dentro del Análisis Predictivo del Churn de Clientes en una Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador.

3.1.3. Criterio de éxito

El modelo debe identificar la posible cancelación de servicios de un cliente analizando las interacciones con la empresa, permitiendo detectar puntos débiles y áreas de mejora en los distintos procesos, tiempos de atención y solución.

- **Aumentar la satisfacción del cliente:** El modelo debe ser entendible y fácil de usar, abarcando las necesidades del cliente desde las distintas áreas que son el frente de la empresa, para ayudar a prevenir la pérdida de clientes.

- **Aumentar la precisión en la medición de la deserción de clientes:** El modelo debe mejorar la detección de la deserción de clientes, permitiendo a la empresa tomar decisiones más concisas sobre promociones, ofertas y servicios.
- **Mejorar la comprensión de las necesidades del cliente:** El modelo debe permitir a la empresa entender el comportamiento de los clientes, aplicar fortalezas para mejorar la relación con ellos y alinear estos esfuerzos con los objetivos empresariales. Esto contribuirá a reducir la deserción mediante un modelo predictivo preciso, eficaz, entendible y fácil de interpretar para aquellos que lo analicen.

3.1.4. Evaluación de la situación

Actualmente, la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador registra sus ventas y cancelaciones a través del Sistema Transaccional de Operaciones, específicamente en el Módulo Financiero. Este módulo gestiona información de pagos, recaudaciones, cartera contable, facturas, entre otros datos que se almacenan en la plataforma Oracle Database, administrada por una empresa externa. Por lo tanto, para analizar las ventas y cancelaciones del servicio al final de cada cierre de mes, los datos relevantes son compartidos por QVD (QlikView Data - archivos que contienen una tabla de datos exportados desde Qlik Sense) para de esta manera generar reportes que permitan evaluar el comportamiento del cliente.

No existe un sistema, herramienta o algoritmo para determinar la probable deserción de clientes, puesto que este proceso se realiza manualmente, contactándose con

aquellos clientes que más reclamos tienen por el servicio otorgado o que han presentado quejas ante la Arcotel el cual es el organismo de control en Ecuador.

La creación de un algoritmo predictivo del Churn de Clientes es crucial para mejorar tanto el servicio al cliente como la eficiencia operativa. Es importante integrar información de todos los sistemas utilizados por la empresa la cual permitirá aplicar acciones tempranas de retención, de tal manera que los clientes deserten y busquen una mejor opción de servicio, especialmente dada la alta competencia en el mercado de las telecomunicaciones en el Ecuador.

Inventario de recursos

- **Recursos humanos**

  Equipo de proyecto: El equipo de proyecto está compuesto únicamente por mí, respaldado por el conocimiento y la experiencia adquiridos a través de la colaboración con expertos de los departamentos de Servicio al Cliente, Marketing, Facturación y Cobranzas, Tecnología y Customer Experience.

  Colaboración con expertos: Se trabajó estrechamente con expertos de cada departamento, asegurando una comprensión profunda de los desafíos relacionados con el negocio, el análisis de la información, las nuevas tendencias y las tecnologías aplicables.

- **Recursos tecnológicos**

Tabla 1. Equipamiento

Realizado por: CHUQUER, William, 2024.

DESCRIPCIÓN	CARACTERÍSTICAS	CANTIDAD
Laptop	✓ Core i5 10 Gen ✓ Windows 10 ✓ Procesador 64 bits ✓ 16 GB de Memoria RAM	4
Qlik Sense Server	✓ Xeon Gold 5320 ✓ Windows Server 2016 Standard ✓ Procesador 64 bits ✓ 128 GB de Memoria RAM	2
Alteryx Server	✓ Core i7 9 Gen ✓ Windows 10 ✓ Procesador 64 bits ✓ 64 GB de Memoria RAM	1

- **Recursos financieros**

Financiamiento: Este proyecto no cuenta con financiamiento de la empresa, ya que forma parte de los procesos empresariales orientados a la mejora continua y al desarrollo tanto personal como institucional. Los datos utilizados en este proyecto no representan ningún costo adicional para la empresa, ya que provienen de sus propios recursos.

- **Recursos temporales**

Duración del proyecto: El proyecto tendrá una duración definida de 4 meses para asegurar el cumplimiento de los objetivos del negocio. Durante este periodo, se ejecutarán todas las fases de la metodología CRISP-DM para garantizar el enfoque del proyecto.

- **Requisitos, supuestos y restricciones**

Tabla 2. Requisitos, supuestos y restricciones en el proyecto

REQUISITOS	SUPUESTOS	RESTRICCIONES
Modelo Preciso	El modelo desarrollado tendrá una alta precisión en la predicción de la deserción de clientes, basándose en las variables seleccionadas y las técnicas de validación aplicadas.	Los recursos limitados pueden reducir la capacidad de realizar un análisis exhaustivo de todas las variables relevantes, lo que podría comprometer la precisión del modelo en predecir la deserción de clientes debido a la falta de tiempo y recursos para optimizarlo adecuadamente.
Modelo Eficaz	Las técnicas de minería de datos utilizadas son eficaces para identificar patrones en los datos que ayuden a predecir la deserción de clientes.	Presencia de limitantes para acceder a datos históricos relevantes y de alta calidad podría reducir la capacidad del modelo para identificar patrones significativos de comportamiento de los clientes.
Modelo Utilizable	El modelo será diseñado y documentado con claridad, facilitando su comprensión y uso por parte de los involucrados en la toma de decisiones estratégicas y operativas.	La complejidad en la interpretación de los resultados del modelo podría limitar su adopción y utilidad práctica por parte de los involucrados no técnicos en la toma de decisiones estratégicas y operativas.
Modelo Disponible	Los datos necesarios para entrenar y validar el modelo están disponibles y son accesibles en todo momento.	La calidad de los datos disponibles puede variar y no siempre ser completa o precisa, lo que podría afectar la precisión del modelo.

- **Riesgos y contingencias**

Tabla 3. Riesgos y contingencias en el proyecto

RIESGOS	CONTINGENCIAS
Data de los clientes	Debido a restricciones legales, no es posible utilizar los datos personales de los clientes de la empresa. Se contará con información que no permita revelar la identidad de los abonados.
Modelo no preciso	Utilizar variables o datos precisos para ello es necesario utilizar técnicas de validación adicionales para mejorar la precisión del modelo.
Modelo no eficaz	Evitar el uso de variables o datos complejos para no afectar el rendimiento del modelo. Es importante realizar cambios en el diseño del modelo para mejorar su eficacia.
Calidad de los datos	Existe el riesgo de que la calidad de los datos recopilados no sean consistentes en la investigación. Es importante implementar un proceso riguroso de validación y limpieza para identificar y corregir errores, eliminar duplicados y completar datos faltantes.

3.1.5. Definición de los objetivos del modelado

- Analizar los datos históricos de los clientes para ser utilizados en el modelo predictivo de tal manera que se puedan identificar patrones y características comunes entre aquellos que han cancelado el servicio.
- Identificar las variables más significativas a ser utilizadas en el modelo.
- Implementar el modelo predictivo adecuado para la deserción de clientes en la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador.

3.1.6. Plan del proyecto

El plan del proyecto establece una duración de 4 meses, periodo necesario para obtener los resultados deseados dentro de la planificación y construcción de los modelos a competir para la deserción de clientes.

3.2. Comprensión de los datos

En la segunda fase, se lleva a cabo la recopilación inicial de los datos con el objetivo de establecer una primera asociación con el problema, así podemos familiarizarnos con los datos, analizar su estructura e identificar patrones iniciales que orienten las primeras hipótesis.

Similar al punto anterior, esta etapa se enfoca a la comprensión, directamente a la información almacenada en las bases de datos de la empresa. Es importante distinguir claramente los tipos de datos almacenados y los tipos de información que representa a cada variable, puesto que esta información se combina con la comprensión del negocio para obtener una visión integral que facilite el análisis y el desarrollo del problema planteado.

3.2.1. Recolección de datos iniciales

Historial de cancelaciones de clientes:

- **Descripción:** Informes mensuales que contiene información sobre los clientes que han cancelado sus servicios, incluyendo las fechas de cancelación y, en algunos casos, los motivos proporcionados por los clientes.
- **Relevancia:** Analizar estos datos es fundamental para identificar patrones de cancelación, como picos en determinadas épocas del año, y comprender las razones comunes detrás de las cancelaciones, lo que permitirá al modelo prever situaciones similares en el futuro.

Encuestas de satisfacción del cliente:

- **Descripción:** Son retroalimentaciones directas realizadas a los clientes sobre su experiencia con el servicio, donde describe los motivos de satisfacción o insatisfacción.
- **Relevancia:** Este tipo de datos permite medir el nivel de satisfacción de los clientes y relacionarlo con el riesgo de churn. Al integrar estos datos en el modelo, se puede identificar cuáles aspectos del servicio son más críticos para la retención de clientes.

Historial de facturaciones de clientes:

- **Descripción:** Informes mensuales que contiene información sobre el historial detallado de facturación de clientes, incluyendo recargos mensuales, descuentos aplicados, promociones, pagos realizados y cualquier otra información financiera relevante.
- **Relevancia:** Los problemas relacionados con la facturación, como errores en descuentos, promociones o recargos inesperados, pueden ser un factor significativo en la decisión de cancelar un servicio. El análisis de estos datos puede ayudar a identificar clientes en riesgo debido a insatisfacciones financieras.

Información demográfica de cliente:

- **Descripción:** Se obtiene datos como género, edad, ubicación geográfica, y otros datos que puedan ser relevantes para entender su comportamiento.
- **Relevancia:** Los datos demográficos son esenciales para segmentar a los clientes y comprender cómo las características personales influyen en la

decisión de cancelar. Este tipo de segmentación permite al modelo ajustar las predicciones según el perfil del cliente.

Historial de clientes retenidos:

- **Descripción:** Informes mensuales que contiene información sobre el historial de clientes retenidos justamente cuando solicitaron en algún momento cancelar su servicio, incluyendo las circunstancias de la que pensaban desertar del mismo.
- **Relevancia:** Analizar estos datos históricos es clave para identificar patrones y comportamientos similares en los clientes actuales que podrían estar considerando cancelar. Estas características históricas permiten que el modelo sea más preciso y proactivo en la identificación de clientes en riesgo.

3.2.2. Métodos de recolección de datos iniciales

- **Sistemas de la empresa:** El conjunto de datos que se utilizará para este análisis pertenece a la Empresa Proveedora del Servicio de Internet para Hogares en el Ecuador. Estos datos se extraen de los QVDs suministrados por la empresa externa y se integra en la herramienta Qlik Sense para la creación de reportes detallados.
- **Interacción con áreas internas:** Este proceso ha permitido que la colaboración de áreas claves dentro de la empresa, aporten con la recolección de información adicional sobre las interacciones con los clientes y sus expectativas y experiencias del servicio.

Con relación a la siguiente tabla que se muestra a continuación, se detallan los reportes a utilizar de los distintos sistemas utilizados por la empresa.

Tabla 4. Recopilación de información en los sistemas de la empresa

SISTEMAS DE LA EMPRESA		
Sistema Transaccional de Operaciones	**Sistema de Análisis de Red**	**Sistema de la Central Telefónica**
• Ventas, Histórico de Clientes (Comercial) • Cartera, Pagos, Facturación (Facturación y Cobranzas) • Soporte – Casos y Tareas (Tecnología, Servicio al Cliente, Call Center) • Atractivo del Servicio – Estudio de Marca, Satisfacción Clientes (Satisfacción y Encuestas).	• Potencia de Red • Índex de Cobertura	• Llamadas Entrantes (recurrencia de llamadas – rellamadas)

3.2.3. Descripción de los datos

De acuerdo con la tabla 4, se ha recolectado información de los tres sistemas de la empresa, generando un **total de 22 bases de datos** de los diferentes procesos necesarios para el estudio. Es importante consolidar estas 22 bases de datos en una única base de datos integrada, ya que este paso permitirá una exploración más detallada y la determinación de los tipos de datos presentes. Es así como una vez comprendida y definida la información, se garantizará su integridad y utilidad para el análisis y desarrollo del modelo predictivo.

En la sección de la ***Preparación de los datos***, se detallará la consolidación de estas bases utilizando un flujo de trabajo en Alteryx. Este flujo abordará tanto a los clientes cancelados como los activos e in-corte para las Categorías de Cancelaciones

Administrativas como Voluntarias. A continuación, se describirán las bases de datos, especificando el sistema de origen y el reporte asociado.

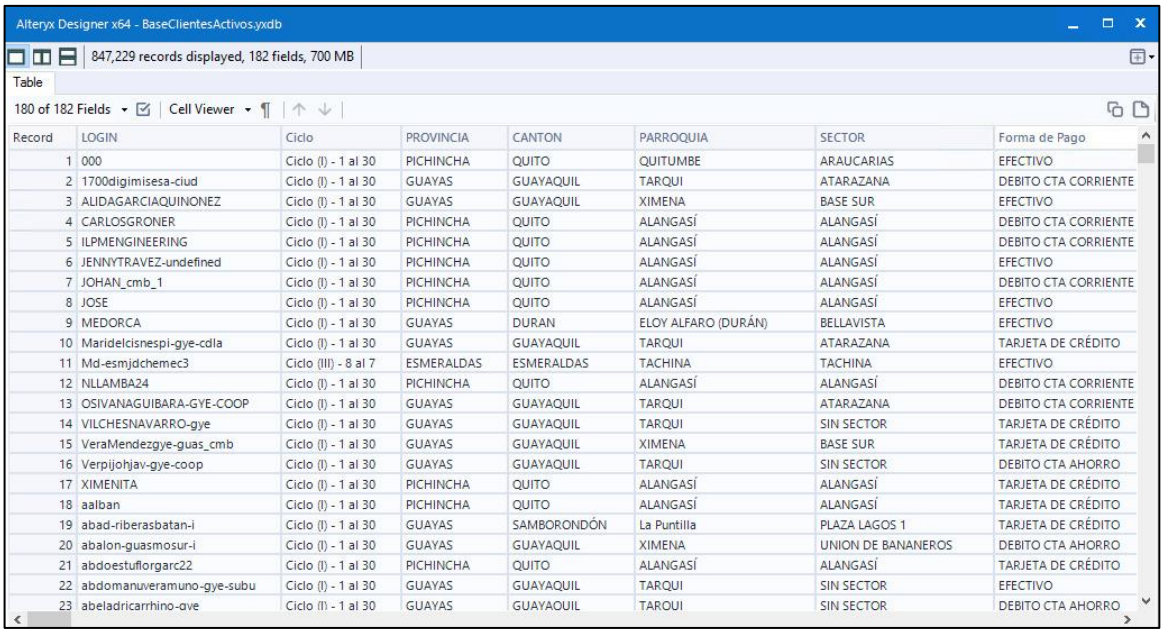
Tabla 5. Descripción del conjunto de bases de datos

Sistema Transaccional de Operaciones		
Bases de Datos	**Reporte Asociado**	**Descripción**
Bdd_clientes_activos_inCorte_2.2	Histórico de Clientes	Muestra información representativa de los clientes activos e in-corte y de los cancelados en las distintas bases de datos. Se considera información demográfica del cliente, temas de pagos, facturación, tickets de soporte, entre otros.
Bdd_clientes_cancelados_2.2	Histórico de Clientes	
Bdd_cartera	Cartera	
Bdd_tareas_Proceso_retencion	Atractivo del Servicio	
Bdd_tareas_info_proceso_cancel	Facturación	
Bdd_tareas_informacion_traslado	Soporte	
Bdd_tareas_ipcc	Soporte	
Bdd_tareas_retencion	Soporte	
Bdd_tareas_Proceso_fidelizacion	Atractivo del Servicio	
Bdd_casos_tecnicos_por_login	Soporte	
Bdd_casos_backbone_por_login	Soporte	
Bdd_rechazos_debitos	Pagos	
Bdd_incortes_totales	Pagos	
Bdd_downgrades_totales	Ventas	
Bdd_upgrades_totales	Ventas	
Sistema de Análisis de Red		
Bases de Datos	**Reporte Asociado**	**Descripción**
Bdd_index_cobertura_all	Índex de Cobertura	Muestra data importante de los clientes activos donde se obtiene información del comportamiento de la red y los equipos internos del cliente.
Bdd_index_cobertura_30	Índex de Cobertura	
Bdd_index_cobertura_90	Índex de Cobertura	
Bdd_index_cobertura_180	Índex de Cobertura	
Bdd_total_disp_por_tipo	Índex de Cobertura	
Bdd_Potencia_equipoONT_45d	Potencia de Red	
Sistema de la Central Telefónica		
Bases de Datos	**Reporte Asociado**	**Descripción**
Bdd_recurrencia_llamadas	Llamadas Entrantes	Muestra data importante de las llamadas realizadas por los clientes a la central telefónica.

1. Cantidad de datos

En la sección de la ***Preparación de los datos***, se describirá el proceso utilizado para obtener el conjunto de datos final. Pues, tras ejecutar el flujo en Alteryx tanto para clientes activos e in-corte, así como para clientes cancelados, se generaron dos conjuntos de datos **"BaseClientesActivos.yxdb"** el cual cuenta con 847,229 registros y 180 variables, y **"BaseClientesCancel.yxdb"** el cual cuenta con 119,107 registros y 167 variables que se utilizarán para crear el conjunto de datos final, el cual será la base para el desarrollo del algoritmo del Análisis Predictivo del Churn de Clientes.

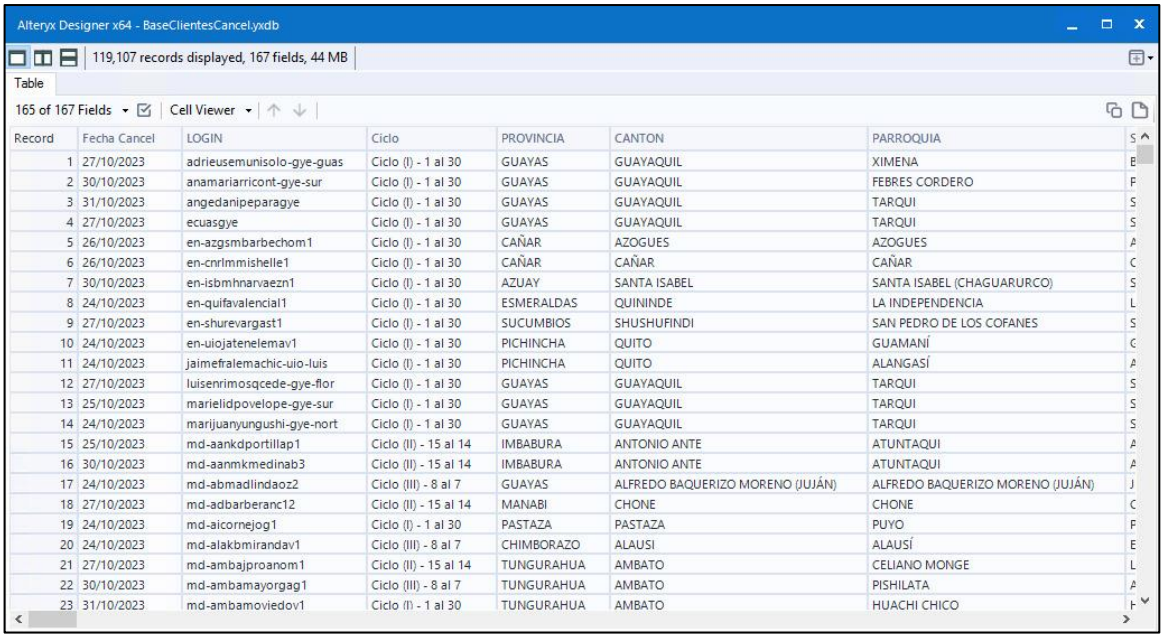
En la figura 18 y figura 19, observamos el conjunto de datos elaborado en Alteryx después de la consolidación del Flujo **"Base_Analisis_Churn"**, el cual será detallado en la sección mencionada inicialmente.



Record	LOGIN	Ciclo	PROVINCIA	CANTON	PARROQUIA	SECTOR	Forma de Pago
1	000	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	QUITUMBE	ARAUCARIAS	EFECTIVO
2	1700digimisesa-ciud	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	ATARAZANA	DEBITO CTA CORRIENTE
3	ALIDAGARCIAQUINONEZ	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	XIMENA	BASE SUR	EFECTIVO
4	CARLOSGRONER	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	DEBITO CTA CORRIENTE
5	ILPMENGINEERING	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	DEBITO CTA CORRIENTE
6	JENNYTRAVEZ-undefined	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	EFECTIVO
7	JOHAN_cmb_1	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	DEBITO CTA CORRIENTE
8	JOSE	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	EFECTIVO
9	MEDORCA	Ciclo (I) - 1 al 30	GUAYAS	DURAN	ELOY ALFARO (DURÁN)	BELLAVISTA	EFECTIVO
10	Maridelcisnespi-gye-cdla	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	ATARAZANA	TARJETA DE CRÉDITO
11	Md-esmjdchemec3	Ciclo (III) - 8 al 7	ESMERALDAS	ESMERALDAS	TACHINA	TACHINA	EFECTIVO
12	NLLAMBA24	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	DEBITO CTA CORRIENTE
13	OSIVANAGUIBARA-GYE-COOP	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	ATARAZANA	DEBITO CTA CORRIENTE
14	VILCHESNAVARRO-gye	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	SIN SECTOR	TARJETA DE CRÉDITO
15	VeraMendezgye-guas_cmb	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	XIMENA	BASE SUR	TARJETA DE CRÉDITO
16	Verpijohjav-gye-coop	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	SIN SECTOR	DEBITO CTA AHORRO
17	XIMENITA	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	TARJETA DE CRÉDITO
18	aalban	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	TARJETA DE CRÉDITO
19	abad-riberasbatan-i	Ciclo (I) - 1 al 30	GUAYAS	SAMBORONDÓN	La Puntilla	PLAZA LAGOS 1	TARJETA DE CRÉDITO
20	abalon-guasmosur-i	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	XIMENA	UNION DE BANANEROS	DEBITO CTA AHORRO
21	abdoestuflorgarc22	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	ALANGASÍ	TARJETA DE CRÉDITO
22	abdomanuveramuno-gye-subu	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	SIN SECTOR	EFECTIVO
23	abeladricarrhino-gye	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	SIN SECTOR	DEBITO CTA AHORRO

Figura 18: Total registros del conjunto de datos BaseClientesActivos.

Realizado por: CHUQUER, William, 2024



Record	Fecha Cancel	LOGIN	Ciclo	PROVINCIA	CANTON	PARROQUIA	S
1	27/10/2023	adrieusemunisolo-gye-guas	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	XIMENA	B
2	30/10/2023	anamariarricont-gye-sur	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	FEBRES CORDERO	P
3	31/10/2023	angedanipeparagye	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	S
4	27/10/2023	ecuasgye	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	S
5	26/10/2023	en-azgsmbarbechom1	Ciclo (I) - 1 al 30	CAÑAR	AZOGUES	AZOGUES	A
6	26/10/2023	en-cnrlmmishelle1	Ciclo (I) - 1 al 30	CAÑAR	CAÑAR	CAÑAR	C
7	30/10/2023	en-isbmhnarvaezn1	Ciclo (I) - 1 al 30	AZUAY	SANTA ISABEL	SANTA ISABEL (CHAGUARURCO)	S
8	24/10/2023	en-quifavalencial1	Ciclo (I) - 1 al 30	ESMERALDAS	QUININDE	LA INDEPENDENCIA	L
9	27/10/2023	en-shurevargast1	Ciclo (I) - 1 al 30	SUCUMBIOS	SHUSHUFINDI	SAN PEDRO DE LOS COFANES	S
10	24/10/2023	en-uiojatenelemav1	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	GUAMANÍ	G
11	24/10/2023	jaimefralemachic-uio-luis	Ciclo (I) - 1 al 30	PICHINCHA	QUITO	ALANGASÍ	A
12	27/10/2023	luisenrimosqcede-gye-flor	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	S
13	25/10/2023	marielidpovelope-gye-sur	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	S
14	24/10/2023	marijuanyungushi-gye-nort	Ciclo (I) - 1 al 30	GUAYAS	GUAYAQUIL	TARQUI	S
15	25/10/2023	md-aankdportillap1	Ciclo (II) - 15 al 14	IMBABURA	ANTONIO ANTE	ATUNTAQUI	A
16	30/10/2023	md-aanmkmedinab3	Ciclo (II) - 15 al 14	IMBABURA	ANTONIO ANTE	ATUNTAQUI	A
17	24/10/2023	md-abmadlindaoz2	Ciclo (III) - 8 al 7	GUAYAS	ALFREDO BAQUERIZO MORENO (JUJÁN)	ALFREDO BAQUERIZO MORENO (JUJÁN)	J
18	27/10/2023	md-adbarberanc12	Ciclo (II) - 15 al 14	MANABI	CHONE	CHONE	C
19	24/10/2023	md-aicornejog1	Ciclo (I) - 1 al 30	PASTAZA	PASTAZA	PUYO	P
20	24/10/2023	md-alakbmirandav1	Ciclo (III) - 8 al 7	CHIMBORAZO	ALAUSI	ALAUSÍ	E
21	27/10/2023	md-ambajproanom1	Ciclo (II) - 15 al 14	TUNGURAHUA	AMBATO	CELIANO MONGE	L
22	30/10/2023	md-ambamayorgag1	Ciclo (III) - 8 al 7	TUNGURAHUA	AMBATO	PISHILATA	A
23	31/10/2023	md-ambamoviedov1	Ciclo (I) - 1 al 30	TUNGURAHUA	AMBATO	HUACHI CHICO	H

Figura 19: Total registros del conjunto de datos BaseClientesCancel.

Realizado por: CHUQUER, William, 2024

2. Tipo de valores

En la siguiente sección de la ***Exploración de los datos***, se describirá el proceso para visualizar los tipos de datos iniciales del dataset final, con el objetivo de seleccionar y estructurar la información más relevante para el análisis requerido. Las variables incluyen diferentes tipos de datos como: flotante (float64), entero (int64) y categóricas (object).

En la figura 20, se muestra una representación de los diferentes tipos de datos del dataset elaborado para la exploración de los datos en Python (Google Colab).

```
# Con data.info() podemos ver las variables categóricas
# (Dtype=object)
datos.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 966336 entries, 0 to 966335
Columns: 180 entries, Empresa to DPTO_1CANCELACION
dtypes: float64(60), int64(1), object(119)
memory usage: 1.3+ GB
```

Figura 20: Tipos de datos iniciales del dataset.

Realizado por: CHUQUER, William, 2024

3.2.4. Exploración de los datos

Los datos utilizados para la exploración de los datos proviene de la unión de los conjuntos de datos “BaseClientesActivos.yxdb” y “BaseClientesCancel.yxdb”, resultando en el archivo denominado “Base_Analisis_Churn”. Dado el elevado número de variables presentes en el dataset, se presentará un resumen fundamental que facilitará la interpretación de la importancia de cada una en el contexto del estudio. Este análisis preliminar es crucial para identificar patrones y relaciones clave que contribuirán al desarrollo del modelo predictivo seleccionado. A continuación, se expone una breve descripción de cada uno los campos relevantes del dataset:

- **LOGIN:** Identificador único de los clientes activos, in-corte y cancelados.
- **Estado Servicio:** Situación actual en la que se encuentra un servicio, este puede ser activo o cancelado.
- **TIPO_ORDEN:** Indica si el servicio instalado es nuevo o trasladado.
- **CATEGORIA.CANCEL:** Muestra los tipos de categorías existentes al momento de cancelar un servicio, voluntarias o administrativas.
- **MOTIVO1_CANCEL:** Muestra el primer motivo que impulsa al cliente a la cancelación del servicio de internet.
- **PROVINCIA:** Región administrativa que indica la ubicación general del servicio de internet.
- **CANTON:** Divisiones dentro de la provincia que especifican la ubicación del servicio de internet.
- **PARROQUIA:** Subdivisión del cantón que detalla con mayor precisión el área donde se ofrece el servicio de internet.

- **SECTOR**: Área específica dentro de la parroquia que señala el punto exacto del servicio de internet.
- **Permanencia(días):** Muestra el número de días que un cliente ha permanecido con su servicio activo.
- **PLAN ACTUAL:** Representa al plan que tiene contratado el cliente.
- **Tipo Negocio:** Representa el servicio de internet que tiene instalado el cliente los cuales son HOME, PYME y PRO.
- **Forma de Pago:** Muestra el tipo de pago que un cliente utilizó en su contrato para pagar el servicio.
- **ComportamientoPago:** Muestra el tipo de comportamiento de pago que un cliente tiene durante el tiempo que ha permanecido con el servicio (Excelente Pagador, Buen Pagador, Regular, Mal y Pésimo Pagador).
- **ONT_CLIENTE_MODELO:** Describe el tipo de equipo según su modelo que tiene el cliente dependiendo del plan contratado.
- **#DocumentosSaldo:** Muestra un conteo saldos pendientes en sus cuotas mensuales durante el tiempo que permanece o ha permanecido con el servicio un cliente.
- **Total_llamadas:** Muestra el número total de llamadas que un cliente a efectuado durante el tiempo que ha permanecido con el servicio.
- **TUVO_PROMO_PLAN1:** Indica si el cliente durante su permanencia con el servicio tuvo promociones o no.
- **TieneExtEmpresa:** Indica si el cliente tiene como producto adicional un extender de la empresa o no.

- **TieneExtPropio:** Indica si el cliente tiene su propio extender o no donde se encuentra instalado el servicio.
- **Rechazos de Débitos:** Los rechazos de débitos indican si los pagos del servicio de internet se realizan de manera puntual. Un alto número de rechazos puede ser un factor que contribuye a la cancelación del plan por parte del cliente.
- **Nivel de Satisfacción del cliente:** Este dato se obtiene a través de encuestas mensuales realizadas a todos los clientes para evaluar su grado de satisfacción con el servicio.
- **Número de In-Cortes:** Muestra el número de veces que el cliente pasó de estado activo a estado in-corte debido a la falta de pago del servicio.
- **Fecha último In-Corte:** Muestra la fecha en la que el cliente pasó de estado activo a estado in-corte debido a la falta de pago del servicio.
- **#TareasProcesoRetencion:** Muestra el número de tareas efectuadas por proceso retención durante el tiempo de permanencia del cliente.
- **%IndexRSSI:** Muestra en nivel de potencia de la señal que un dispositivo inalámbrico recibe del router, este valor nos permite evaluar la calidad de una conexión inalámbrica y optimizar la red adecuadamente.
- **%IndexSNR:** Muestra el nivel de la calidad de señal emitido por el ruido de comunicación en el sistema que puede afectar a la señal Wifi.

Para asegurar que los datos utilizados en el desarrollo del modelo predictivo del churn de clientes sean de alta calidad y permitan una comprensión clara y precisa del

comportamiento de los usuarios, se lleva a cabo el siguiente proceso en la exploración de los datos:

a) Estructuración de los datos:

Se examina la organización y el formato de los conjuntos de datos para identificar y comprender las variables relevantes para el Análisis Predictivo del Churn de Clientes.

En la figura 20, observamos que el conjunto de datos final contiene una cantidad de 966,336 registros y 180 variables, distribuidas en tres tipos: flotantes (float64, con 60 features), enteros (int64 con 1 features) y categóricas (object con 119 features) totalmente estructuradas. Luego, para una mejor comprensión de los datos, en la figura 21 se presenta la forma del dataset donde presenta el 31.53% de NaN, esto quiere decir que existen datos ausentes o no están disponibles para determinadas variables, mientras que en la figura 22, proporciona un resumen que detalla el tipo de dato y la naturaleza de cada variable, permitiendo identificar la presencia de valores nulos o únicos en el estudio del dataset.

```
# Ver la forma del dataset

print( f"Features (columnas): {datos.shape[1]}")

print( f"Registros (filas): {datos.shape[0]}")
print( f"Celdas: {datos.shape[0] * datos.shape[1]}")

print(f"Porcentaje de NaN: {datos.isna().mean().mean() * 100  :.2f}%")

Features (columnas): 180
Registros (filas): 966336
Celdas: 173940480
Porcentaje de NaN: 31.53%
```

Figura 21: Forma del dataset.

Realizado por: CHUQUER, William, 2024

```python
# Crear un DataFrame de resumen
resumen = pd.DataFrame({
    'Nombre Columna': datos.columns,
    'Tipo de Dato': datos.dtypes,
    'Valores Nulos': datos.isnull().sum(),
    'Valores Únicos': datos.nunique()
})

# Mostrar el resumen completo
print(resumen)
```

```
                     Nombre Columna Tipo de Dato  Valores Nulos  \
Empresa                     Empresa       object              0
LOGIN                         LOGIN       object              0
Ciclo                         Ciclo       object              0
PROVINCIA                 PROVINCIA       object              0
CANTON                       CANTON       object              0
...                             ...          ...            ...
Fecha Cancel           Fecha Cancel       object         847229
CATEGORIA.CANCEL    CATEGORIA.CANCEL       object         847229
AREA_1CANCELACION  AREA_1CANCELACION       object         847229
USR_1CANCELACION    USR_1CANCELACION       object         847229
DPTO_1CANCELACION  DPTO_1CANCELACION       object         847229

                   Valores Únicos
Empresa                         2
LOGIN                      966066
Ciclo                           3
PROVINCIA                      23
CANTON                        179
...                           ...
Fecha Cancel                  305
CATEGORIA.CANCEL                4
AREA_1CANCELACION               6
USR_1CANCELACION              143
DPTO_1CANCELACION               9

[180 rows x 4 columns]
```

Figura 22: Dataframe resumen.

Realizado por: CHUQUER, William, 2024

b) Aplicación de estadística básica:

Realizar un análisis estadístico preliminar para obtener una visión general de las distribuciones, tendencias y posibles correlaciones que puedan influir en la cancelación del servicio.

En la figura 23, se presenta el resultado de ***"describe( )"***, que proporciona una estadística descriptiva de las variables numéricas. Este análisis descriptivo sirve como guía para evaluar la información relevante e identificar patrones que son cruciales para el análisis de los factores que afecten la cancelación del servicio.

```
# Estadísticas descriptivas para las columnas numéricas
datos.describe()
```

	PRECIO_VENTA	Permanencia(dias)	#DocumentosSaldo	SaldoTotal($)	MesesLlamadas	Total_llamadas	DiasAntesCancel	#Registros	#DiasRegistros	TotalConn	...
count	966336.000000	966336.000000	502983.000000	502983.000000	265415.000000	265415.000000	265415.00000	828384.000000	828384.000000	8.283840e+05	...
mean	31.015798	994.654143	1.213270	35.226099	1.465188	7.154219	75.86101	1442.092037	67.316922	5.591690e+03	...
std	9.930006	870.060463	0.547179	19.292764	0.846804	96.996529	53.02639	518.014112	23.574015	9.803569e+03	...
min	0.000000	0.000000	0.000000	-847.820000	1.000000	1.000000	1.00000	0.000000	1.000000	0.000000e+00	...
25%	27.500000	320.000000	1.000000	25.760000	1.000000	1.000000	28.00000	1208.000000	69.000000	2.474000e+03	...
50%	29.990000	769.000000	1.000000	33.590000	1.000000	2.000000	69.00000	1664.000000	74.000000	4.662000e+03	...
75%	34.990000	1382.000000	1.000000	40.880000	2.000000	4.000000	118.00000	1729.000000	74.000000	7.530000e+03	...
max	400.000000	4689.000000	12.000000	1014.880000	15.000000	6118.000000	184.00000	5155.000000	240.000000	7.902681e+06	...

8 rows × 61 columns

Figura 23: Análisis Descriptivo Variables Numéricas.

Realizado por: CHUQUER, William, 2024

Por ejemplo, si queremos comprender mejor la variable numérica **PRECIO_VENTA**, observamos que la media del precio mensual del plan de internet es aproximadamente $31 dólares. El valor mínimo registrado es de $0 dólares, que corresponde a nuevos contratos que aún no han generan factura. Por otro lado, el plan de internet más caro registrado tiene un precio de $400 dólares.

En la figura 24, se presenta un análisis descriptivo de las variables categóricas del dataset. Este resumen proporciona un conteo de valores para cada variable, identifica si los valores son únicos y determina si algún valor se encuentra entre los más frecuentes. Además, muestra la frecuencia del valor más común, lo cual es crucial para comprender patrones de comportamiento y características relevantes que podrían influir en la predicción del churn de clientes.

```
# Seleccionar las columnas categóricas
categorical_columns = datos.select_dtypes(include=['object']).columns

# Resumen detallado para cada variable categórica
for col in categorical_columns:
    print(f'Columna: {col}')
    print(f'Conteo de Valores: {datos[col].count()}')
    print(f'Valores Únicos: {datos[col].nunique()}')
    print(f'Valor Más Frecuente (Top): {datos[col].mode()[0]}')
    print(f'Frecuencia del Valor Más Frecuente: {datos[col].value_counts().max()}')
    print('-'*40)
```

```
----------------------------------------
Columna: Ciclo
Conteo de Valores: 966336
Valores Únicos: 3
Valor Más Frecuente (Top): Ciclo (I) - 1 al 30
Frecuencia del Valor Más Frecuente: 392457
----------------------------------------
Columna: PROVINCIA
Conteo de Valores: 966336
Valores Únicos: 23
Valor Más Frecuente (Top): GUAYAS
Frecuencia del Valor Más Frecuente: 366053
----------------------------------------
Columna: CANTON
Conteo de Valores: 966336
Valores Únicos: 179
Valor Más Frecuente (Top): QUITO
Frecuencia del Valor Más Frecuente: 304566
----------------------------------------
Columna: PARROQUIA
Conteo de Valores: 966336
Valores Únicos: 641
Valor Más Frecuente (Top): TARQUI
```

Figura 24: Análisis Descriptivo Variables Categóricas.

Realizado por: CHUQUER, William, 2024

Por ejemplo, en la columna **Ciclo** se observan tres valores únicos, que corresponden a los tres ciclos de facturación utilizados por la empresa. El ciclo más frecuente es el **Ciclo (I) – 1 al 30**, que abarca el periodo del primero al 30 del mes para facturar el mes de servicio de internet. De los 966,336 registros totales, el **Ciclo (I) – 1 al 30** es el más común con 392,457 clientes que tienen este ciclo de facturación.

c) Detección de inconsistencias y datos faltantes:

Identificar y gestionar valores atípicos, nulos o vacíos que podrían comprender la calidad del análisis y precisión del modelo predictivo.

Como se mostró en la figura 21, los valores **"NaN"** representan el 31,53% del total de la información en el dataset. En la figura 25, se detalla la distribución de estos valores faltantes, identificando a 155 variables (de un total de 180) con ocurrencias de NaN.

```
# Identificación de valores faltantes
valores_faltantes = datos.isnull().sum()
print(valores_faltantes[valores_faltantes > 0])

# Opcional: Visualización de valores faltantes
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
sns.heatmap(datos.isnull(), cbar=False, cmap='viridis')
plt.show()
```

```
TIPO_CUENTA_CONTRATO    272627
BANCO                   272627
PRIMER PLAN                  1
ZONA_OLT                  4221
NOMBRE_OLT                1479
                          ...
Fecha Cancel            847229
CATEGORIA.CANCEL        847229
AREA_1CANCELACION       847229
USR_1CANCELACION        847229
DPTO_1CANCELACION       847229
Length: 155, dtype: int64
```

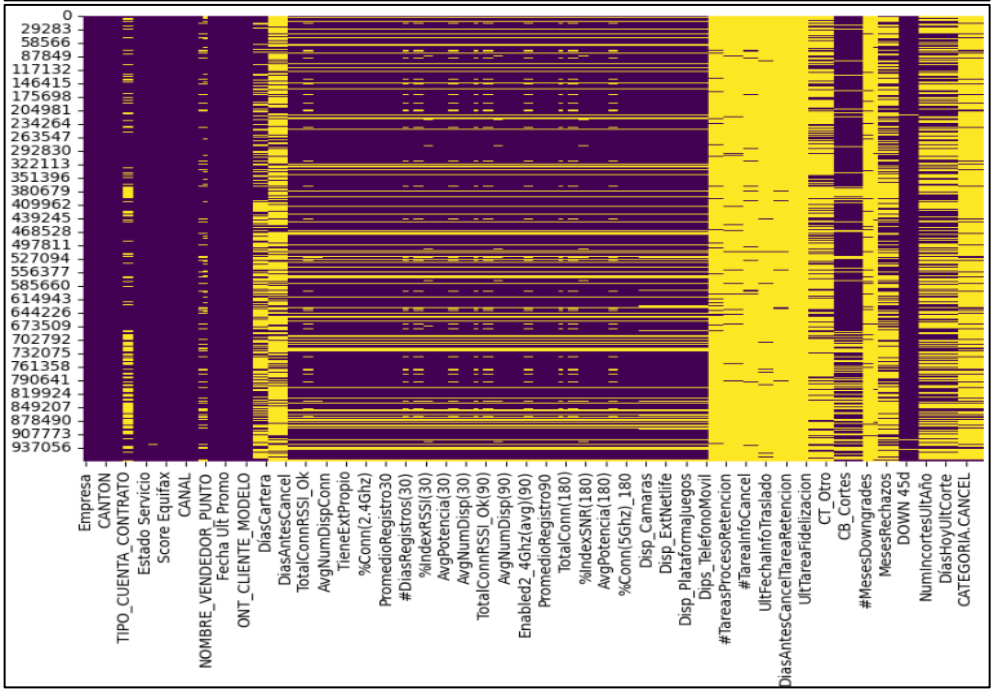


Figura 25: Valores faltantes, nulos o vacíos

Realizado por: CHUQUER, William, 2024

d) Visualización de los datos:

Emplear gráficas y herramientas de visualización para representar de manera efectiva la información y los patrones identificados, facilitando la interpretación y toma de decisiones específicamente para la limpieza de datos.

En la figura 26, se muestra un resumen visual de los histogramas de todas las variables numéricas, lo que permite una comprensión más clara y rápida de la distribución de los datos. Además, en la figura 27, se presenta un resumen general de los ***overviews de las features***, incluyendo la presencia de valores nulos, no nulos, estadísticas descriptivas y gráficas de distribución. Estos *overviews* proporcionan una visión clara y concisa del contenido y la estructura del conjunto de datos, lo que ayuda a identificar las características más relevantes y así asegurar la calidad del análisis posterior.

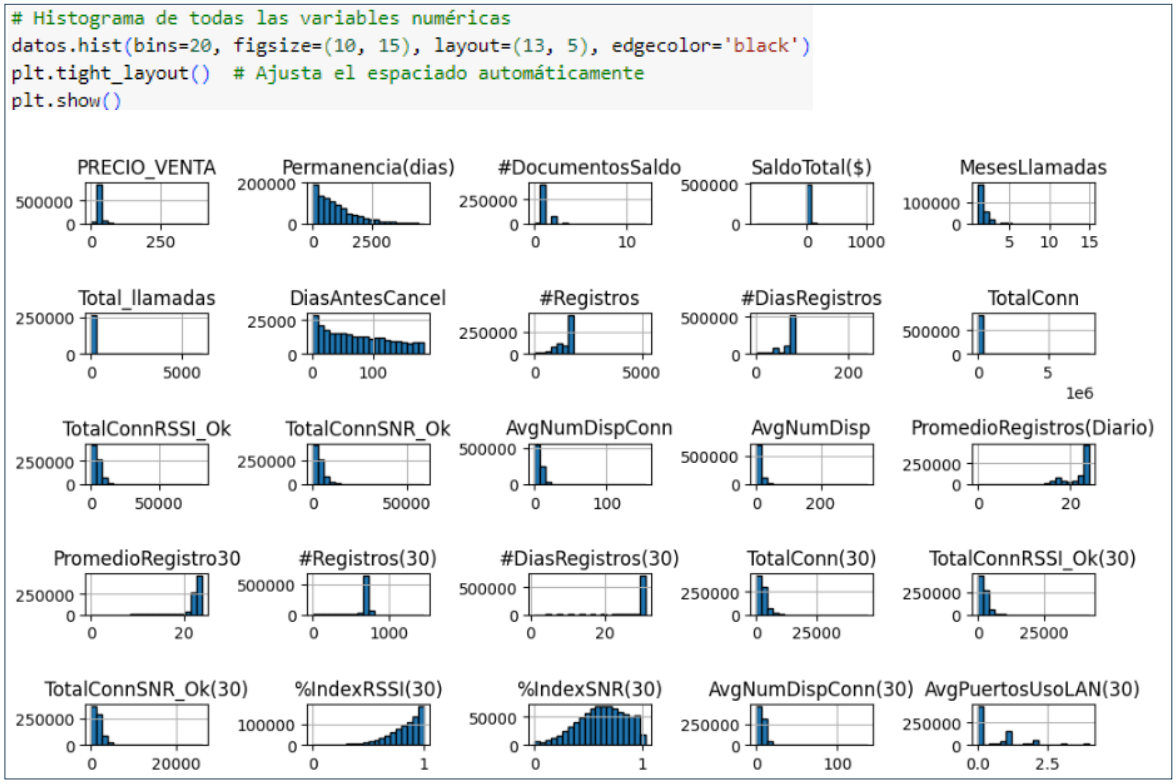


Figura 26: Histogramas Variables Numéricas.

Realizado por: CHUQUER, William, 2024

```
# Overview de las features
for i in datos.columns:
    contador_nan = datos[i].isna().sum()
    contador_dash = (datos[i] == "-").sum()
    contador_cero = (datos[i] == "0").sum()  + (datos[i] == 0).sum()
    registros = datos.shape[0]

    print(

        # General
        f"Columna:      \033[1m{i.upper()}\033[0m",
        f"Tipo de Dato: {datos[i].dtype}",

        "",

        # Valores Nulos
        f"NaN:          {contador_nan}      {contador_nan/ registros * 100 :.2f}%",
        f"Dash:         {contador_dash}      {contador_dash / registros * 100 :.2f}%",
        f"Total Nulos:  {contador_nan + contador_dash}      {(contador_nan + contador_dash)/ registros * 100 :.2f}%",
        f"Ceros:        {contador_cero}      {contador_cero / registros * 100 :.2f}%",

        "",

        # Valores No Nulos
        f"Valores Únicos:   {datos[i].nunique()}",
        f"Distribución: \n\n{datos[i].value_counts()}",

        # Espacio
        "\n\n\n",

        sep = "\n"
        )
# Estadística para valores numéricos
if datos[i].dtype != "object":
    print(

        f"Máx.:     {datos[i].max()}",
        f"Media:    {datos[i].mean()}",
        f"Mediana:  {datos[i].median()}",
        f"Mín.:     {datos[i].min()}",
        "",
        sep = "\n"
    )

# Graficar distribución
if datos[i].nunique() > 100 and datos[i].dtype == "object":

    print("Gráfico complejo de procesar", "\n\n\n", )

else:
    # Manejar valores no numéricos antes de graficar
    if datos[i].dtype == 'object':
        numeric_data = pd.to_numeric(datos[i], errors='coerce').dropna()  # Convertir a numérico, ignorando errores
        if numeric_data.size > 0:  # Comprueba si quedan valores numéricos
            plt.figure(figsize=(9,2))
            sns.histplot(data=numeric_data, bins=10)
            plt.xticks(rotation=90)
            plt.show()
        else:
            print(f"Columna '{i}' no contiene valores numéricos para graficar.\n\n\n")
    else:
        plt.figure(figsize=(9,2))
        sns.histplot(data = datos[i], bins = 10)
        plt.xticks(rotation=90)
        plt.show()
```

```
PRECIO_VENTA
29.99     112364
27.50     112315
23.00      77787
33.00      76397
29.50      69823
           ...
14.00          1
20.90          1
31.90          1
87.00          1
275.00         1
Name: count, Length: 117, dtype: int64



Máx.:     400.0
Media:    31.015797879826476
Mediana:  29.99
Mín.:     0.0
```

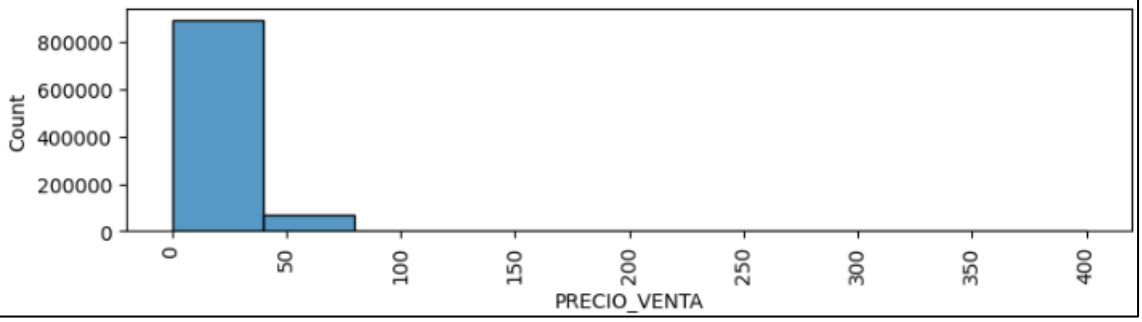


Figura 27: Resumen overviews de las features.

Realizado por: CHUQUER, William, 2024

3.2.5. Verificación de la calidad de los datos

La verificación de la calidad de los datos es fundamental para garantizar que sean confiables y adecuados para el modelado predictivo del churn de clientes. Asegurar la calidad de los datos es esencial para lograr un modelo preciso y obtener resultados que reflejen con exactitud el comportamiento de los usuarios, lo que permitirá tomar decisiones informadas y efectivas.

a) Revisar tamaño del dataset:

Se valoró esta información en la sección de la Exploración de los datos, punto 1: Estructuración de los datos donde el dataset contiene 966,336 registros y 180 variables compuestas por variables numéricas y categóricas.

b) Datos nulos:

En la figura 25 se identificó una cantidad significativa de datos nulos o NaN, lo que destaca la necesidad de tratarlos adecuadamente para mejorar la calidad del conjunto de datos. Esta revisión es crucial para garantizar la integridad del análisis y evitar ajustes futuros innecesarios. Es esencial que tanto las variables numéricas como las categóricas estén completas para cada cliente. Por lo tanto, en la sección de ***Preparación de los datos*** se llevará a cabo la limpieza de datos aplicado mediante un flujo en Alteryx, con el objetivo de asegurar una mayor calidad y fiabilidad de los datos para el modelo predictivo.

c) Calidad de los datos:

En términos generales, la calidad de los datos no es óptima, como se evidenció durante la exploración inicial, donde se detectó una considerable cantidad de variables con valores nulos. Aunque no se encontraron errores significativos en el dataset debido al volumen de información procesada, es crucial mejorar la calidad de los datos para asegurar un análisis más preciso. Por esta razón, en la sección de ***Preparación de los datos***, procederemos a categorizar las variables, lo que permitirá mejorar la precisión del análisis y contribuir al cumplimiento de los objetivos planteados en este estudio.

3.3. Preparación de los datos

Una vez completado el estudio en el apartado de la Comprensión de los datos, es necesario establecer una relación clara entre los datos explorados y la construcción del modelo predictivo para el Análisis del Churn de Clientes. En otras palabras, es esencial que la base de datos se ajuste a las necesidades de los modelos que se desarrollarán en el futuro, dado que, dependiendo del enfoque teórico o de modelación que se utilice, habrá diferentes requisitos que deben cumplirse para desarrollar el modelo de manera efectiva.

Este apartado se centra en el tratamiento de valores faltantes, el ajuste del formato de las variables, la elección de las variables pertinentes para el modelo, y la definición de los datos de referencia basados en el período de tiempo examinado. Además, se deben considerar otras modificaciones necesarias en la base de datos original para garantizar que sea adecuada para el desarrollo del proyecto de titulación.

Ahora bien, cabe recordar que como se mencionó en la sección de la Descripción de los datos, se utilizaron un total de 22 bases de datos correspondientes a diferentes procesos de la empresa. En esta sección, no solo detallaremos la consolidación de las bases de datos en un único dataset, sino que también presentaremos todos los procesos realizados, los cuales se ilustran en la figura 28 para la Preparación de los datos utilizando la herramienta Alteryx.

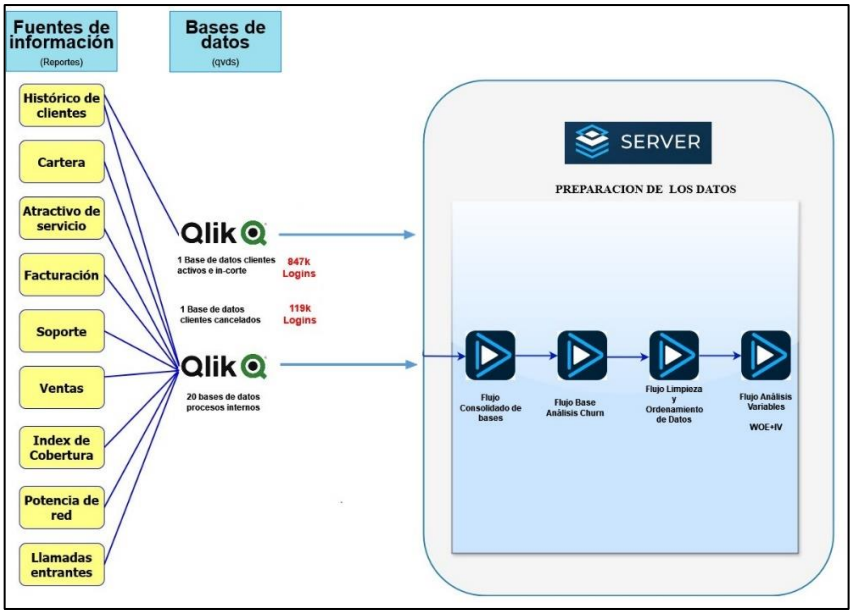


Figura 28: Diagrama de procesos - Preparación de los Datos.

Realizado por: CHUQUER, William, 2024

Como se muestra en el diagrama de procesos, se realizó la extracción de las fuentes de información a partir de los distintos reportes generados en la herramienta Qlik Sense. Estas bases se integraron en un primer flujo dentro de Alteryx (total de cuatro flujos elaborados en Alteryx para esta fase), donde el resultado de este proceso fue la obtención de dos bases de datos finales, “BaseClientesActivos.yxdb” y “BaseClientesCancel.yxdb”, que son las que hemos mencionado durante la Comprensión de los datos.

En la figura 29, se representa el primer flujo elaborado en Alteryx para visualizar la consolidación de las bases, es importante mencionar que cada bloque con procesos internos corresponde a un contenedor el cual se utiliza para organizar y gestionar secciones específicas del flujo, esto con el fin de facilitar la organización, el control, y la comprensión del flujo trabajado, especialmente como este tipo de flujos complejos.

En la figura 30, se presenta una vista más detallada del flujo, destacando el contenedor de llamadas y algunos de los procesos clave. Dentro de estos contenedores, se

emplean diversas herramientas, representadas por íconos, cada una con una función específica en el flujo. Por ejemplo, se utilizan herramientas como: *Select, Join, filtros, ordenamientos, creación de fórmulas*, así como herramientas de machine learning, que serán fundamentales en el entrenamiento de los modelos seleccionados.

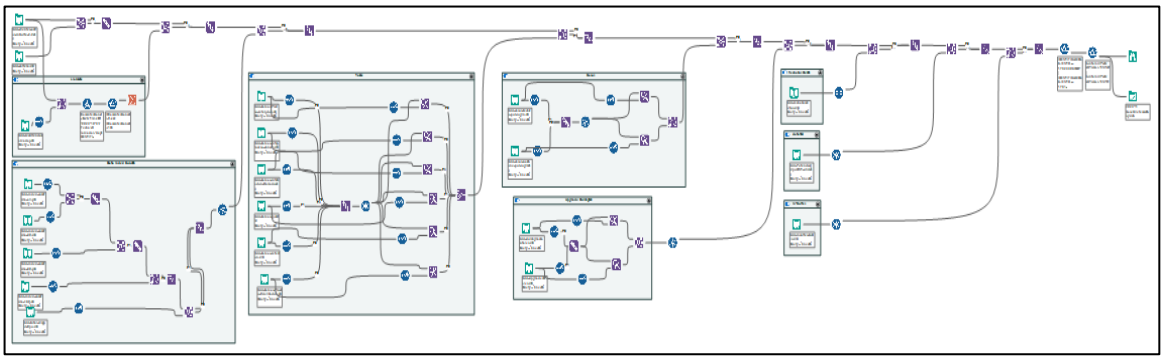


Figura 29: Flujo consolidado de bases.

Realizado por: CHUQUER, William, 2024

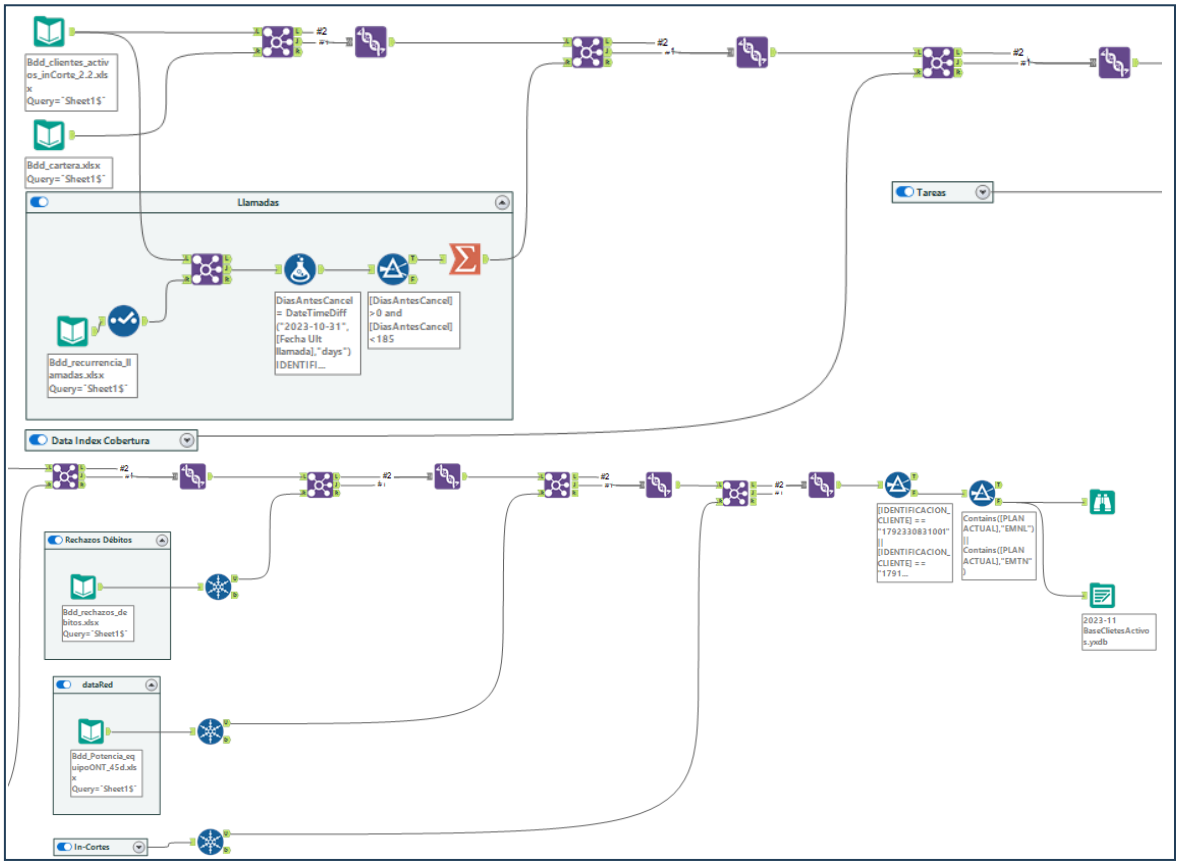


Figura 30: Flujo consolidado de bases ampliado.

Realizado por: CHUQUER, William, 2024

El siguiente paso consistió en consolidar ambas bases resultantes en una sola y para ello se elaboró el segundo flujo de trabajo que se utilizó en la sección de la Exploración de los datos y que representamos en la figura 31 como un flujo sencillo en comparación con los procesos más complejos que se aplican en el estudio.

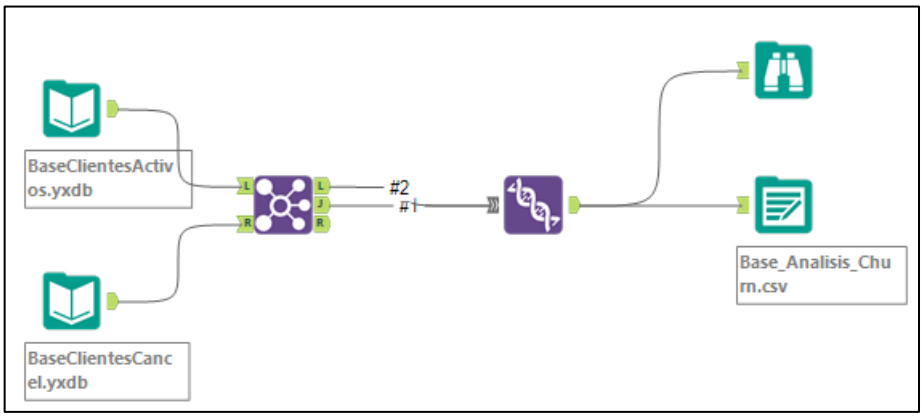


Figura 31: Flujo Base Análisis Churn.
Realizado por: CHUQUER, William, 2024

3.3.1. Selección de los datos

Tal como se mencionó en la sección de la Verificación de la calidad de los datos, literal c, debemos categorizar las variables para mejorar la precisión del análisis y eliminar variables que no son relevantes para este estudio. Por tanto, es necesario excluir los campos que no serán considerados de las 180 variables iniciales quedando con un total de 87 variables para el análisis. Vamos a nombrar algunas variables que fueron eliminadas: EMPRESA, CICLO, TIPO_CUENTA_CONTRATO, BANCO, PRIMER PLAN, ZONA_OLT, NOMBRE_OLT, FECHA_INSTALACION_REAL, PRECIO_VENTA, FECHA_ANTERIOR_TRASLADO, USR_VENDEDOR, NOMBRE_VENDEDOR, CANAL, TUVO_PROMO_PLAN1, TUVO_NDI, FECHA_ULT_NDI, ULT_TIPO_PAGO, #Registros, #DiasRegistros, TotalConnRSSI, TotalConnSNR.

3.3.2. Limpieza de datos

Recordando el dataset “Base_Analisis_Churn”, generada en el segundo flujo, contiene un total de 966,336 registros con 180 variables y que, durante la Exploración de los datos, se identificó que el 31,53% de los valores eran NaN. Ahora bien, para abordar este problema, se implementó un tercer flujo de limpieza y ordenamiento de datos. Para ello, utilizamos dos funciones, la primera función *Find Replace* y diversas fórmulas que permiten buscar y reemplazar valores en el dataset, mientras que para la segunda función *Data Cleansing* nos permite limpiar y preparar datos de manera eficiente como eliminar espacios en blanco, convertir mayúsculas a minúsculas, eliminar caracteres no deseados, eliminar o reemplazar filas, y otras tareas de depuración.

Este proceso se aplicó a las 180 variables (reducidas a 87 variables) donde se estableció una regla: si una variable presentaba más del 70% de sus valores como NaN, se procedía a eliminar y para el caso de tener registros con NaN en las variables presentes, se procedía a poner cero en las columnas de tipo entero, evitando así campos vacíos que podrían distorsionar el análisis.

Este enfoque de descartar variables o registros con una gran cantidad de valores NaN se denomina filtrado de datos o depuración de datos.

En la figura 32, se presenta el tercer flujo elaborado en Alteryx, donde se realiza la Limpieza y ordenamiento de datos necesarios para esta fase del proyecto, con el objetivo de preparar el conjunto de datos para la implementación del modelo predictivo. Además, en la figura 33, se proporciona una vista más detallada del flujo aplicado.

Consideraciones: Es importante destacar que en este flujo, también se llevó a cabo las categorizaciones necesarias para mejorar la calidad de los datos. Estas categorizaciones

serán explicadas en detalle en la siguiente sección dedicada a la Construcción de nuevos datos.

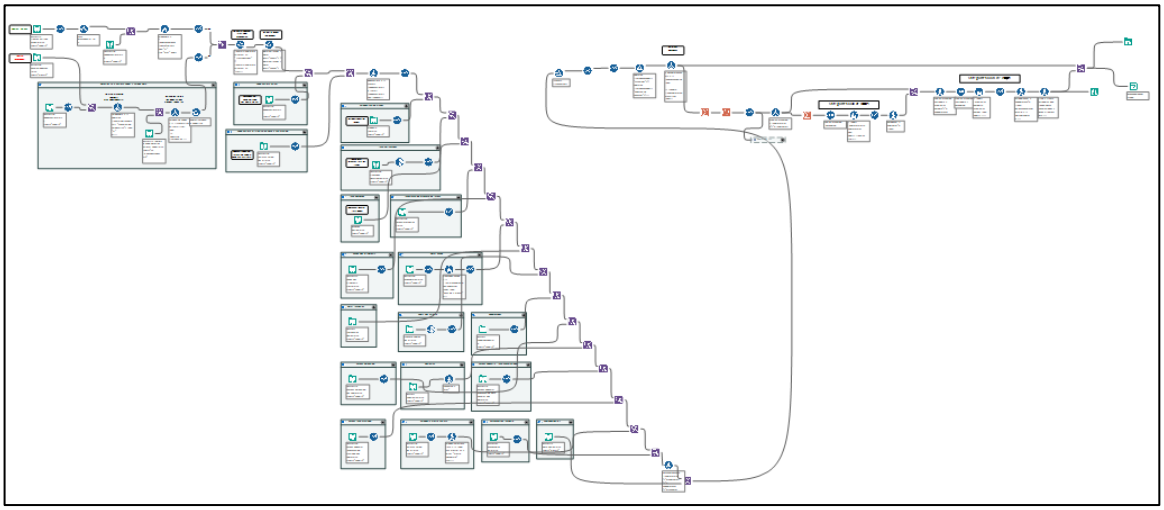


Figura 32: Flujo limpieza y ordenamiento de datos.

Realizado por: CHUQUER, William, 2024

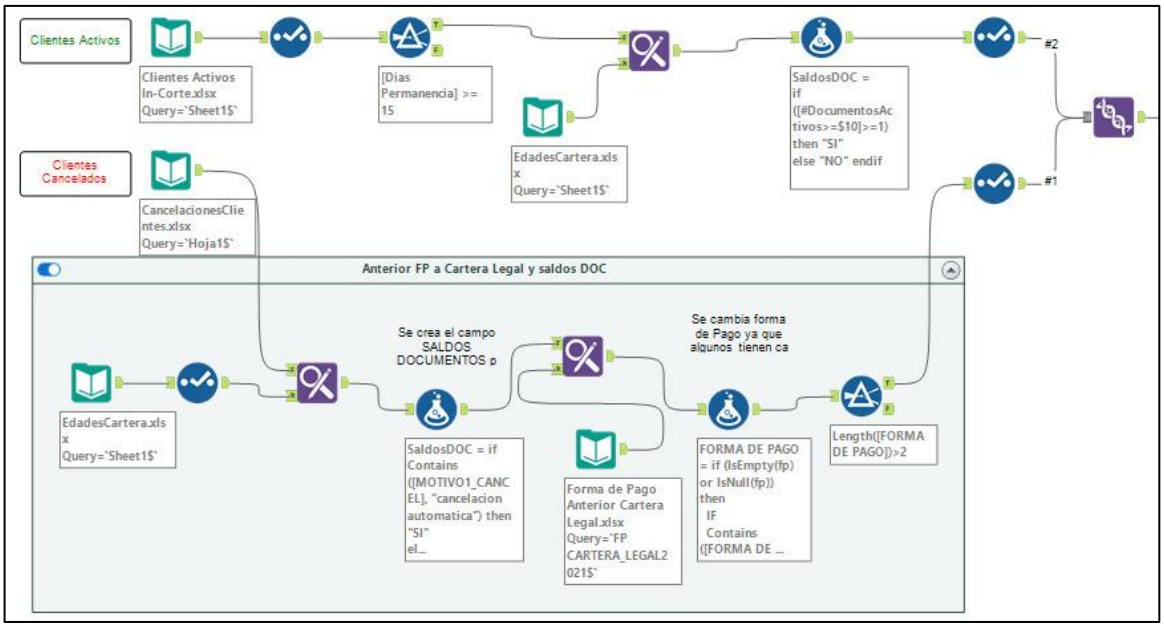


Figura 33: Categorización de Catera con forma de pago y saldos.

Realizado por: CHUQUER, William, 2024

3.3.3. Construcción de nuevos datos

En esta sección, se definirán todas las categorizaciones realizadas durante el estudio y, adicional a ello, se presentará el cuarto flujo de Análisis de Variables mediante WOE + IV. Por tanto, antes de profundizar en el concepto y el objetivo de trabajar con WOE + IV, es importante mencionar todas las categorizaciones:

- Para los clientes activos, se seleccionó una base de datos con un tiempo de permanencia mayor a un año para entrenar al algoritmo con datos más representativos y variados.
- La forma de pago de los clientes es un factor crucial; sin embargo, se encontraron datos dispersos con significados similares. Estos datos se categorizaron para crear un conjunto homogéneo de categorías, facilitando el reconocimiento por parte del algoritmo. Por ejemplo, las formas de pago como transferencia, débito y canje se unificaron bajo la categoría "efectivo", según lo acordado con el responsable del proceso.
- Se consideró importante registrar cuántos meses un cliente ha llamado reiteradamente, por esta razón, se incorporó al dataset el número de llamadas realizadas durante su permanencia.
- En cuanto a la información de satisfacción obtenida a través de encuestas, los motivos de satisfacción se clasificaron de la siguiente manera: una satisfacción alta se valoró como 3, neutral como 2, e insatisfacción como 1.
- Para la información sobre la intención de contratar el servicio nuevamente, se clasificaron las respuestas en una escala de 1 a 5, donde 1 es "nada dispuesto" y 5 es "muy dispuesto."

- Al evaluar los datos, se identificó dispersión en varias variables de la base consolidada, como permanencia, total de casos, tipos de problemas (cortes, atenuaciones, wifi, etc.), provincias, y tipos de orden. Para mejorar la consistencia, estas variables se categorizaron adecuadamente. Por ejemplo, en el caso de la permanencia, los registros con menos de 365 días se categorizaron como "Menor a 1 año," los que tenían entre 365 y 1085 días como "Entre 1 y 3 años," y los que superaban los 1085 días como "Mayor a 3 años."
- Se categorizó también la existencia de casos abiertos con las siguientes reglas: si había un caso registrado, se asignaba un valor de 1; de lo contrario, 0. Las categorías específicas, como casos de atenuaciones, cortes, problemas wifi, entre otros, se manejaron de manera similar, asignando 1 si existían y 0 si no.
- Para la categoría de provincias, se priorizó según la cantidad de clientes: Pichincha y Guayas se mantuvieron como categorías individuales, mientras que las provincias restantes se agruparon en R1 (región sierra) y R2 (región costa) según su relevancia.
- El tipo de orden, debido a su variedad y baja frecuencia, se simplificó bajo la categoría "reubicación por traslado."
- También se consideró si un cliente contaba con más de un servicio de internet. En ese caso, se asignó un valor de 1; si no, 0.
- El motivo de cancelación se categorizó en campos como problemas económicos, cambio de domicilio, mejor oferta de la competencia, entre otros. Para los clientes activos, se utilizó la categoría "Activos."

- En cuanto a rechazos de débitos, se establecieron rangos: sin rechazos, menos de 6 rechazos, entre 6 y 12 rechazos, y más de 12 rechazos.
- Finalmente, se categorizaron variables relacionadas con tareas y cortes, asignando un valor de 1 si el cliente había experimentado estos eventos, y 0 si no. Lo mismo se aplicó para la frecuencia de llamadas: si el cliente llamó más de una vez por quejas, se asignó un valor de 1; de lo contrario, 0.

WOE (Weight of Evidence): Son técnicas estadísticas utilizadas en el análisis de datos con el objetivo de usarlos para ordenar las categorías de una variable de manera que refleje su relación con la variable dependiente, mejorar la interpretabilidad y rendimiento del modelo (Chakraborty, 2021).

$$WoE = \left[ln \frac{Relative\ frequecy\ of\ Goods}{Relative\ frequecy\ of\ Bads}\right] * 100$$

IV (Information Value): Son técnicas estadísticas utilizadas en el análisis de datos con el objetivo de evaluar la capacidad predictiva de una variable independiente respecto a una variable dependiente binaria (Chakraborty, 2021).

$$IV = \sum (DistributionGood_i - DistributionBad_i) * WoE_i$$

En el proyecto, WOE + IV se utilizarán para seleccionar y transformar variables que tienen una relación significativa con el churn de clientes.

En la figura 34, se presenta el cuarto flujo elaborado en Alteryx, donde se realiza el Análisis de Variables en conjunto con el WOE + IV necesarios para esta fase del proyecto, con el objetivo de mejorar la eficiencia y precisión del modelo predictivo. Además, en la figura 35, se proporciona una vista más detallada del flujo aplicado.

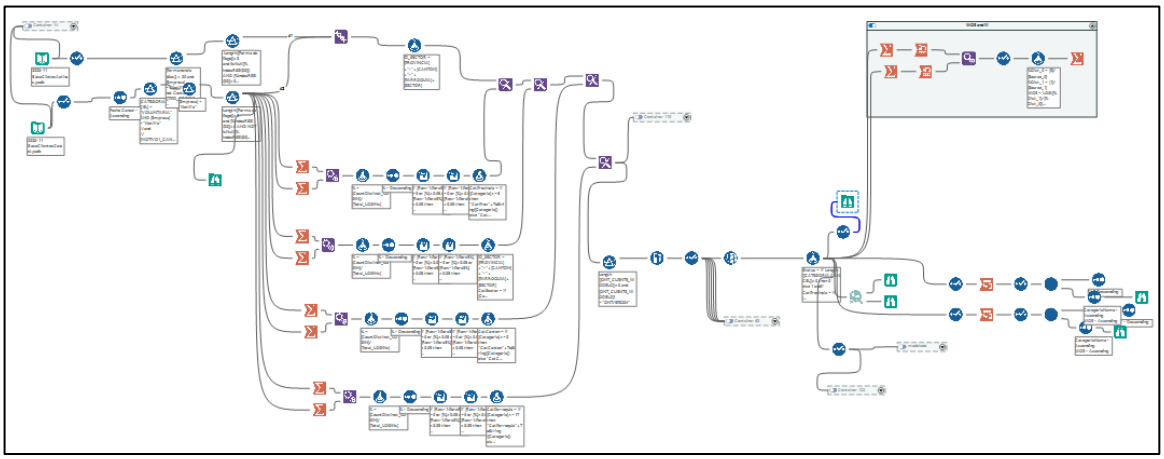


Figura 34: Flujo Análisis Variables WOE + IV.

Realizado por: CHUQUER, William, 2024

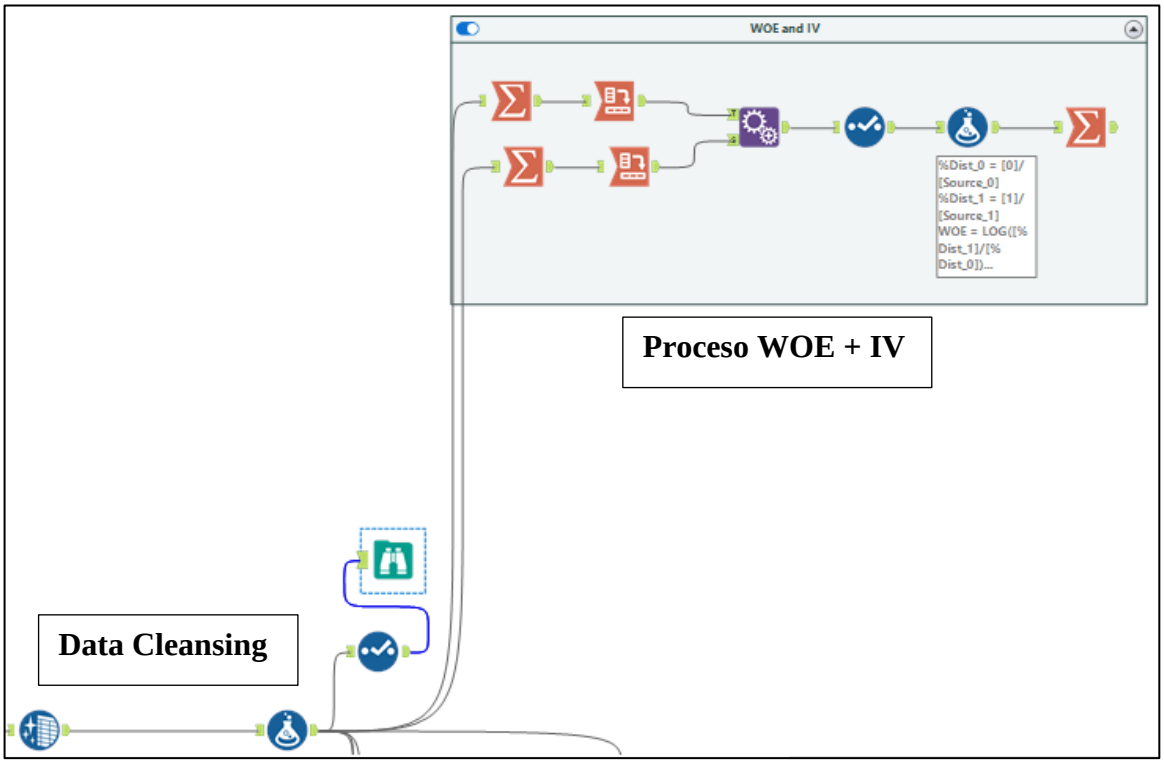


Figura 35: Data Cleansing y Proceso WOE + IV en el flujo de análisis de variables.

Realizado por: CHUQUER, William, 2024

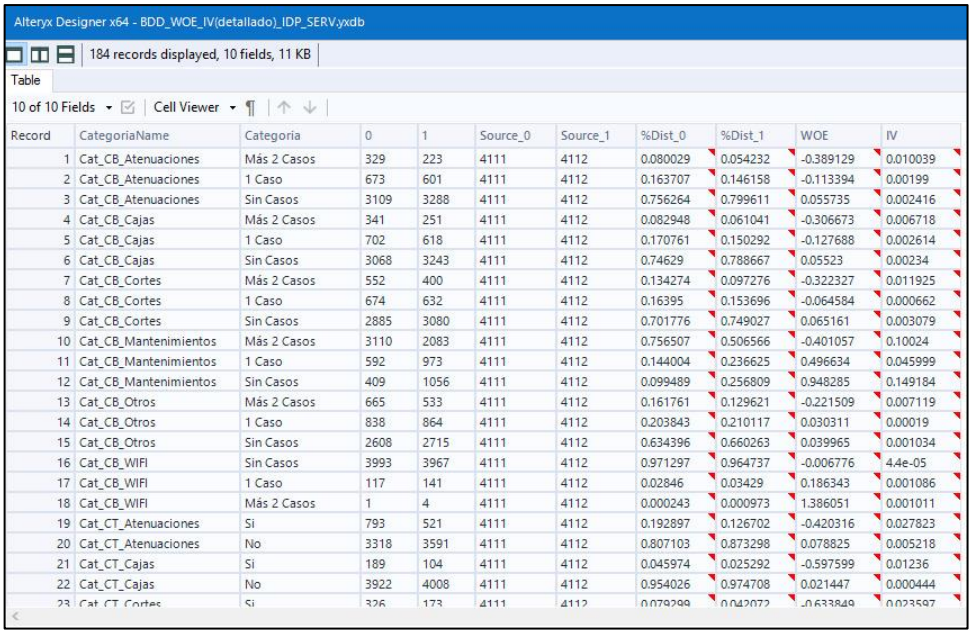
3.3.4. Proceso de selección de variables y preparación para el modelado

Debido a la construcción del cuarto flujo, podemos identificar y seleccionar las variables más relevantes para el modelo predictivo. Este proceso nos permite enfocarnos en aquellas variables del dataset que poseen un mayor poder predictivo según su valor de Information Value (IV). Las variables seleccionadas serán claves para mejorar la precisión y efectividad del modelo.

Tabla 6. Reglas relacionadas con Information Value (IV)

IV (Information Value)	Predictive Power
< 0.02	Predictor Demasiado Débil
0.02 a 0.1	Predictor Débil
0.1 a 0.3	Predictor Medio
0.3 a 0.5	Predictor Fuerte
> 0.5	Predictor Extremadamente Fuerte

A continuación, para entender mejor como se aplicaron las reglas basadas en el Information Value (IV), presentaremos un ejemplo que considera únicamente las categorías creadas en el punto anterior. Este enfoque nos permite evaluar el poder predictivo de cada categoría y determinar si es viable incluirlas en la construcción del modelo predictivo. Por tanto, en la figura 36, se muestra la base de datos con las categorías creadas, incluyendo la información del WOE e IV. Adicional, en la figura 37, se presenta un ranking de las categorías con mayor poder predictivo que fueron analizadas de las bases de datos de categorías presentes en la figura 36.



Alteryx Designer x64 - BDD_WOE_IV(detallado)_IDP_SERV.yxdb

184 records displayed, 10 fields, 11 KB

Table

10 of 10 Fields | Cell Viewer

Record	CategoriaName	Categoria	0	1	Source_0	Source_1	%Dist_0	%Dist_1	WOE	IV
1	Cat_CB_Atenuaciones	Más 2 Casos	329	223	4111	4112	0.080029	0.054232	-0.389129	0.010039
2	Cat_CB_Atenuaciones	1 Caso	673	601	4111	4112	0.163707	0.146158	-0.113394	0.00199
3	Cat_CB_Atenuaciones	Sin Casos	3109	3288	4111	4112	0.756264	0.799611	0.055735	0.002416
4	Cat_CB_Cajas	Más 2 Casos	341	251	4111	4112	0.082948	0.061041	-0.306673	0.006718
5	Cat_CB_Cajas	1 Caso	702	618	4111	4112	0.170761	0.150292	-0.127688	0.002614
6	Cat_CB_Cajas	Sin Casos	3068	3243	4111	4112	0.74629	0.788667	0.05523	0.00234
7	Cat_CB_Cortes	Más 2 Casos	552	400	4111	4112	0.134274	0.097276	-0.322327	0.011925
8	Cat_CB_Cortes	1 Caso	674	632	4111	4112	0.16395	0.153696	-0.064584	0.000662
9	Cat_CB_Cortes	Sin Casos	2885	3080	4111	4112	0.701776	0.749027	0.065161	0.003079
10	Cat_CB_Mantenimientos	Más 2 Casos	3110	2083	4111	4112	0.756507	0.506566	-0.401057	0.10024
11	Cat_CB_Mantenimientos	1 Caso	592	973	4111	4112	0.144004	0.236625	0.496634	0.045999
12	Cat_CB_Mantenimientos	Sin Casos	409	1056	4111	4112	0.099489	0.256809	0.948285	0.149184
13	Cat_CB_Otros	Más 2 Casos	665	533	4111	4112	0.161761	0.129621	-0.221509	0.007119
14	Cat_CB_Otros	1 Caso	838	864	4111	4112	0.203843	0.210117	0.030311	0.00019
15	Cat_CB_Otros	Sin Casos	2608	2715	4111	4112	0.634396	0.660263	0.039965	0.001034
16	Cat_CB_WIFI	Sin Casos	3993	3967	4111	4112	0.971297	0.964737	-0.006776	4.4e-05
17	Cat_CB_WIFI	1 Caso	117	141	4111	4112	0.02846	0.03429	0.186343	0.001086
18	Cat_CB_WIFI	Más 2 Casos	1	4	4111	4112	0.000243	0.000973	1.386051	0.001011
19	Cat_CT_Atenuaciones	Si	793	521	4111	4112	0.192897	0.126702	-0.420316	0.027823
20	Cat_CT_Atenuaciones	No	3318	3591	4111	4112	0.807103	0.873298	0.078825	0.005218
21	Cat_CT_Cajas	Si	189	104	4111	4112	0.045974	0.025292	-0.597599	0.01236
22	Cat_CT_Cajas	No	3922	4008	4111	4112	0.954026	0.974708	0.021447	0.000444
23	Cat_CT_Cortes	Si	326	173	4111	4112	0.079299	0.042072	-0.633849	0.023597

Figura 36: Base de datos - Categorías aplicando WOE + IV.

Realizado por: CHUQUER, William, 2024

CategoriaName	IV	Predictive Power
CatTareaInfoCancel	2,5421	Predictor extremadamente Fuerte
CatEquipoApagado	1,0447	Predictor extremadamente Fuerte
ONT_CLIENTE_MODELO	0,4368	Fuerte Predictor
CatTareaRetencion	0,3208	Fuerte Predictor
Cat_CB_Mantenimientos	0,2385	Medium predictor
CatPotencia	0,2262	Medium predictor
CatMesesLlamadas	0,2051	Medium predictor
CatLlamadasUltAño	0,2050	Medium predictor
CatPermanencia	0,1529	Medium predictor
CatTareaInfoTraslado	0,0831	Débil Predictor
CatDispConn(30)	0,0759	Débil Predictor
CatCanton	0,0709	Débil Predictor
CatProvincia	0,0559	Débil Predictor
CatNumDisp(30)	0,0556	Débil Predictor
Score Equifax	0,0553	Débil Predictor
FormaPago	0,0484	Débil Predictor
CatSector	0,0477	Débil Predictor
CatGamaPrecios	0,0397	Débil Predictor
Categoria2.4Ghz(30)	0,0296	Débil Predictor
Categoria5Ghz(30)	0,0296	Débil Predictor
TieneFact(Inst o Trasl)	0,0240	Débil Predictor
CatTareaProcesoRetencion	0,0226	Débil Predictor
ComportamientoPago	0,0144	Demasiado Débil Predictor
Cat_CT_Wifi	0,0128	Demasiado Débil Predictor
Cat_UsoPuertosLan(30)	0,0097	Demasiado Débil Predictor
CatTareaIPCC	0,0094	Demasiado Débil Predictor
Cat_CT_Otros	0,0087	Demasiado Débil Predictor
TieneExtPropio	0,0063	Demasiado Débil Predictor
CategoriaSNR(30)	0,0050	Demasiado Débil Predictor
TieneExtNetlife	0,0043	Demasiado Débil Predictor
CatDowngrade	0,0035	Demasiado Débil Predictor
Cat_CB_Cajas	0,0033	Demasiado Débil Predictor
Cat_CB_Cortes	0,0033	Demasiado Débil Predictor
Cat_CT_Cortes	0,0026	Demasiado Débil Predictor
Cat_In-Cortes	0,0025	Demasiado Débil Predictor
Cat_CB_WIFI	0,0018	Demasiado Débil Predictor
CategoriaRSSI(30)	0,0014	Demasiado Débil Predictor
CatUpgrade	0,0014	Demasiado Débil Predictor
Cat_CB_Atenuaciones	0,0013	Demasiado Débil Predictor
Cat_CB_Otros	0,0006	Demasiado Débil Predictor
Cat_CT_Cajas	0,0005	Demasiado Débil Predictor
TUVO_PROMO_PLAN1	0,0002	Demasiado Débil Predictor
Cat_CT_Atenuaciones	0,0002	Demasiado Débil Predictor

Figura 37: Ranking poder predictor de las variables categorizadas.

Realizado por: CHUQUER, William, 2024

3.4. Modelado

Tras completar la fase de comprensión y preparación de los datos, es crucial seleccionar los modelos sobre los cuales se aplicará el conjunto de entrenamiento. Estos modelos serán evaluados para determinar su capacidad predictiva y precisión. En este proyecto, se emplearán tres modelos predictivos supervisados, aplicando técnicas de modelado donde los parámetros de cada modelo serán ajustados para optimizar su rendimiento.

3.4.1. Proceso de selección de técnica de modelado

Antes de abordar en la construcción de los modelos predictivos supervisados, es fundamental dividir el conjunto de datos en dos grupos con características equivalentes para permitir una evaluación adecuada del desempeño de los modelos. Se empleará el 70% de los datos para entrenamiento del modelo, y el 30% sobrante para pruebas. Por tanto, es importante asegurar que la división de los datos mantenga una distribución equilibrada en todo el conjunto de datos, evitando se presente cualquier sesgo que pueda afectar a los resultados.

Como se indicó en la sección 2.6, se utilizarán tres modelos propuestos para el Análisis Predictivo del Churn de Clientes, estos modelos a competir son: Regresión Logística, Árboles de Decisión y Random Forest. Estos modelos utilizarán el mismo conjunto de variables descritos en las secciones 3.3.1 y 3.3.4, garantizando una evaluación justa y comparativa de su desempeño.

3.4.2. Construcción del modelo

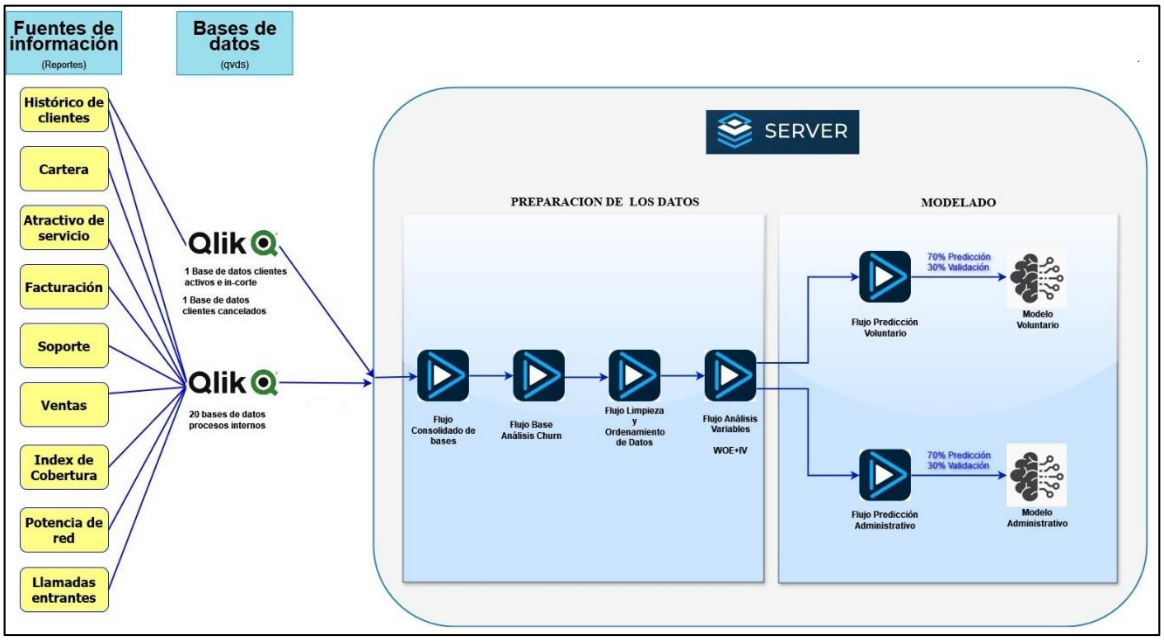


Figura 38: Diagrama de procesos - Modelado.

Realizado por: CHUQUER, William, 2024

Como se muestra en el diagrama de procesos de la figura 38, la fase de Preparación de los datos abarcó todo el proceso hasta el cuarto flujo utilizando la herramienta Alteryx, ahora bien, a partir del Flujo de Análisis de Variables (WOE + IV), se inició la construcción de dos flujos para evaluar el desempeño del modelo de aprendizaje en el Análisis Predictivo del Churn de Clientes: uno para cancelaciones administrativas y otro para cancelaciones voluntarias.

Ambos flujos se presentan en la figura 39, cada uno de ellos dentro de su respectivo contenedor, donde se visualizan los procesos de entrenamiento de los tres modelos seleccionados, aplicando la parametrización de variables con la técnica Train-Test Split, utilizando el 70% de los datos para entrenamiento del modelo, y el 30% sobrante para pruebas.

El conjunto de entrenamiento es fundamental para enseñar al modelo a identificar patrones, mientras que el conjunto de pruebas evalúa su desempeño en datos no vistos previamente. Esto permite obtener una evaluación preliminar de la funcionalidad del modelo y su capacidad para generalizar a un volumen de datos mayor.

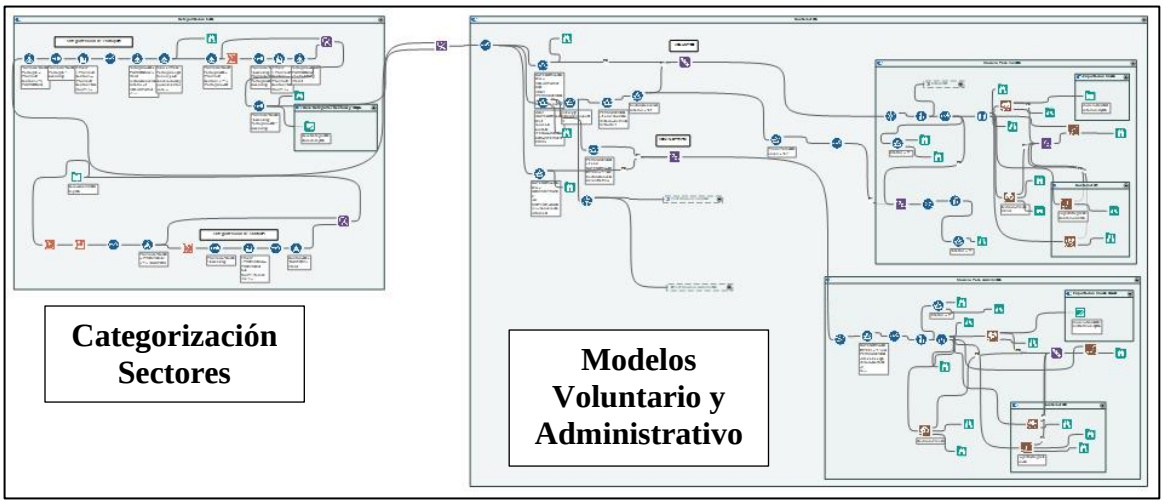


Figura 39: Flujo Fase de Modelado en Alteryx.
Realizado por: CHUQUER, William, 2024

En la figura 39, se detalla un proceso diseñado para categorizar los sectores del país en función de las provincias y cantones. Esta categorización se realiza porque el comportamiento del servicio y de los clientes varían significativamente entre diferentes regiones. Al hacerlo, se pueden identificar tendencias o patrones específicos, como por ejemplo, las razones comunes detrás de las cancelaciones en diversas áreas o que en un análisis de churn para una empresa de servicios de internet, categorizar por provincias y cantones podría revelar que algunas zonas presentan tasas de churn más altas debido a factores como la calidad del servicio o la competencia más intensa en esas regiones. Por tanto, esta información permite a la empresa tomar decisiones informadas, tal como mejorar la infraestructura o ajustar sus estrategias de retención en esos lugares.

Una vez completada la categorización por sectores, se procedió a desarrollar procesos adicionales para categorizar tanto los motivos de cancelación como la permanencia de los clientes. Estos pasos se llevaron a cabo con el objetivo de mejorar el análisis previo a la unificación de las bases de datos para el entrenamiento de los modelos predictivos, los cuales competirán para ofrecer el resultado esperado. Las categorizaciones realizadas se presentan de manera más visual en la figura 40.

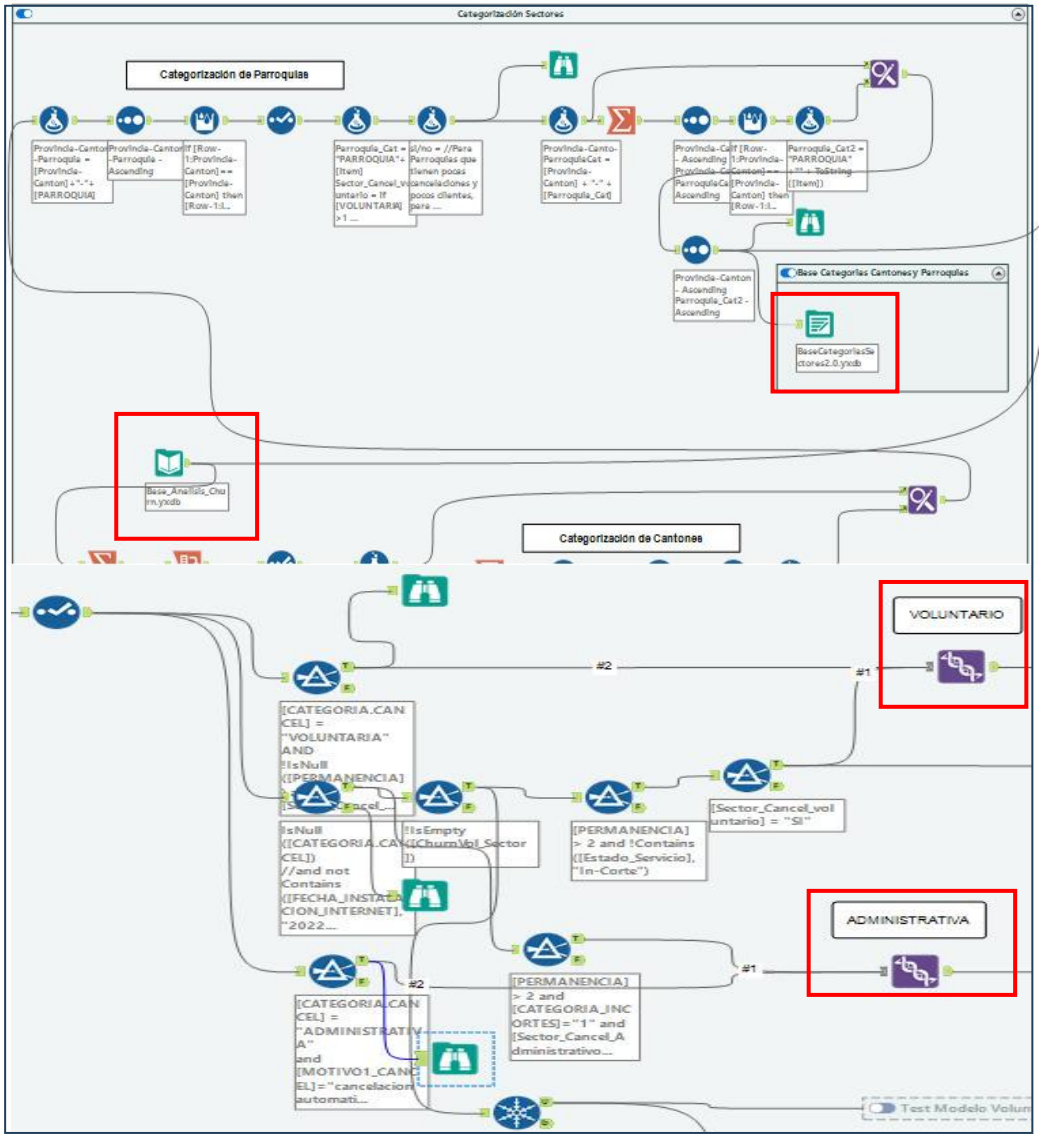


Figura 40: Flujo Modelado - Categorizaciones.

Realizado por: CHUQUER, William, 2024

En la figura 40, se puede visualizar de mejor manera el flujo de la fase del modelado en relación con las categorizaciones realizadas. Los cuadros en color rojo destacan a la base de datos principal utilizada a lo largo del proyecto denominada “Base_Analisis_Churn”, así como la nueva base que se está generando con los procesos de categorización de sectores previamente descritos. Además, se muestran los pasos de las demás categorizaciones, culminando en la unión de los resultados que utilizaremos para el diseño de los modelos predictivos tanto para cancelaciones voluntarias como las administrativas.

Retomando el flujo de la fase de modelado presentado en la figura 39, a continuación se visualizan los procesos en detalle tanto en la figura 41, como en la figura 42 sobre los modelos elaborados para Cancelaciones Voluntarias y Cancelaciones Administrativas respectivamente. Allí se pueden identificar los íconos correspondientes a los tres modelos entrenados, al final, se presentan los resultados individuales de cada modelo así como la comparación entre ellos, donde Random Forest ha demostrado ser el modelo con el mejor resultado de predicción en este estudio.

Consideraciones: Los resultados detallados y la competencia entre los tres modelos, serán descritos de mejor manera en la fase de evaluación.

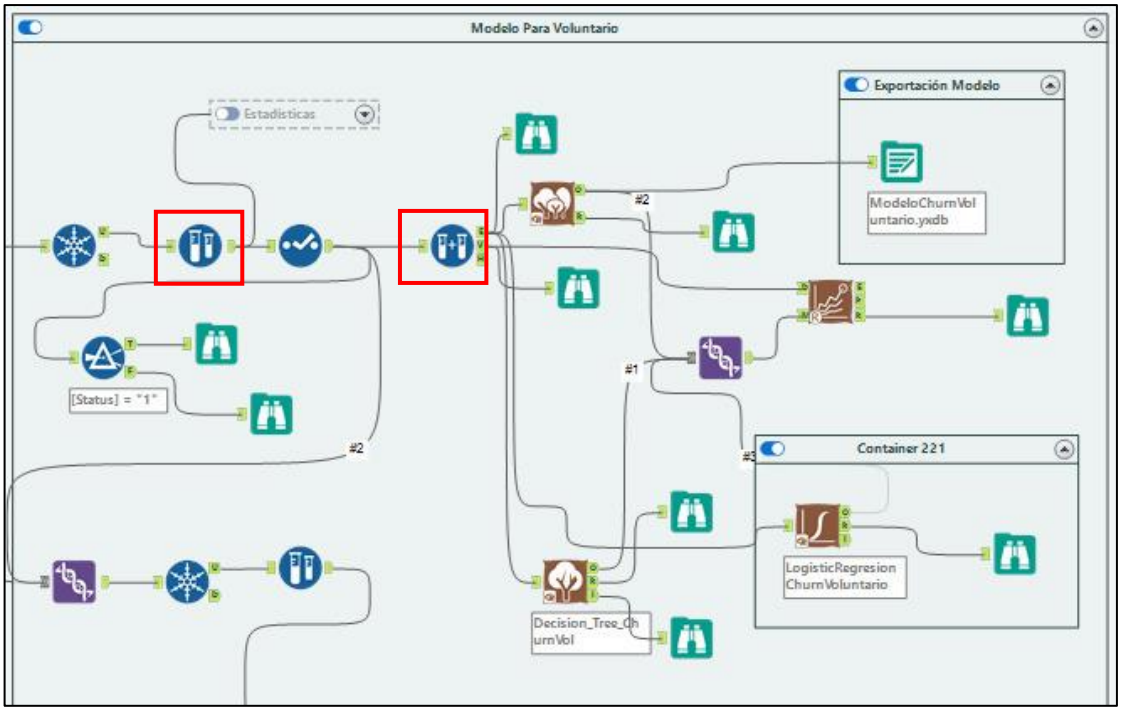


Figura 41: Flujo Modelo Predicción - Voluntario.

Realizado por: CHUQUER, William, 2024

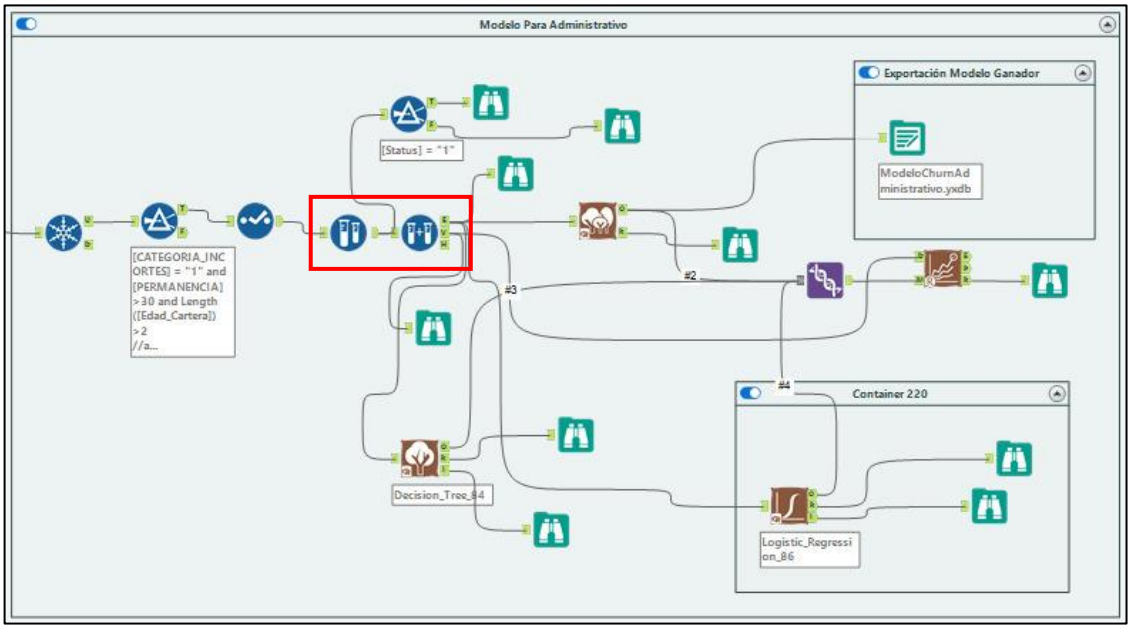


Figura 42: Flujo Modelo Predicción - Administrativo.

Realizado por: CHUQUER, William, 2024

En ambas figuras podemos observar con íconos de color café a los tres modelos, tanto Regresión Logística, Árboles de decisión, Random Forest y al final el ícono llamado *Model Comparison* el cual nos muestra los resultados obtenidos del entrenamiento efectuado, pues evalúa y selecciona el modelo predictivo más efectivo entre los candidatos. Ahora bien, en los cuadros de color rojo hemos encerrado a la herramienta *Oversample Field y Create Samples* los cuales describimos a continuación:

Oversample Field: Se utiliza para ajustar la proporción de clases en un conjunto de datos desequilibrado, con el objetivo de mejorar el rendimiento de un modelo predictivo. En nuestro caso, el campo de interés se establece en 50%, esto significa que después de aplicar Oversample Field, el conjunto de datos resultante será equilibrado con un 50% de los registros perteneciendo a la clase de interés y el otro 50% a las demás clases. En otras palabras, la herramienta ajustará los datos de manera que la clase deseada represente la mitad del total de registros logrando así una distribución equilibrada (ver figura 43).

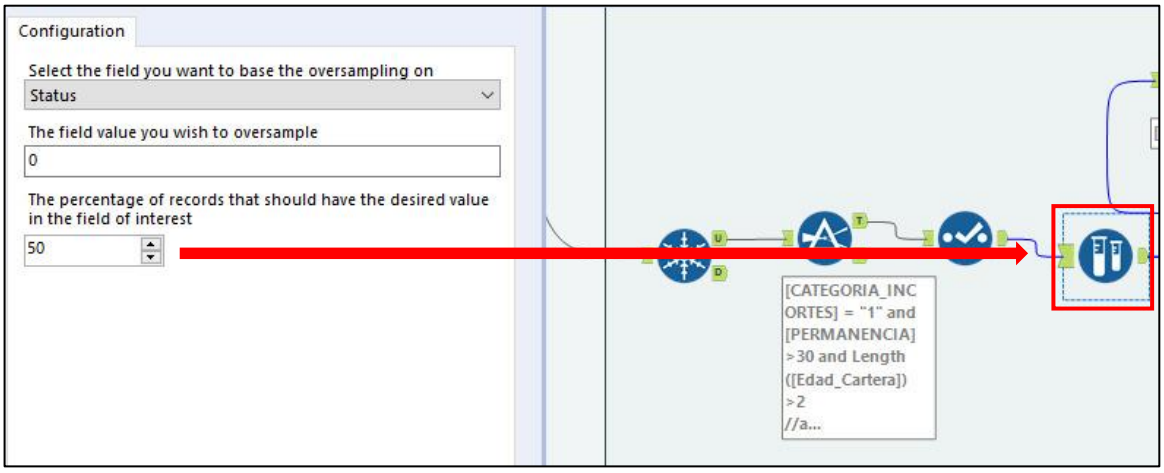


Figura 43: Oversample Field al 50% de equilibrio.

Realizado por: CHUQUER, William, 2024

Create Samples: Permite dividir los datos en diferentes subconjuntos para distintos propósitos como entrenamiento, validación, entre otros. En nuestro caso lo utilizamos para aplicar la técnica del Train-Test Split con una estimación del 70% y una validación del 30%. Por lo tanto, al usar un 70%-30% en *Create Samples* en Alteryx, es lo mismo que aplicar un proceso de Train-Test Split (ver figura 44).

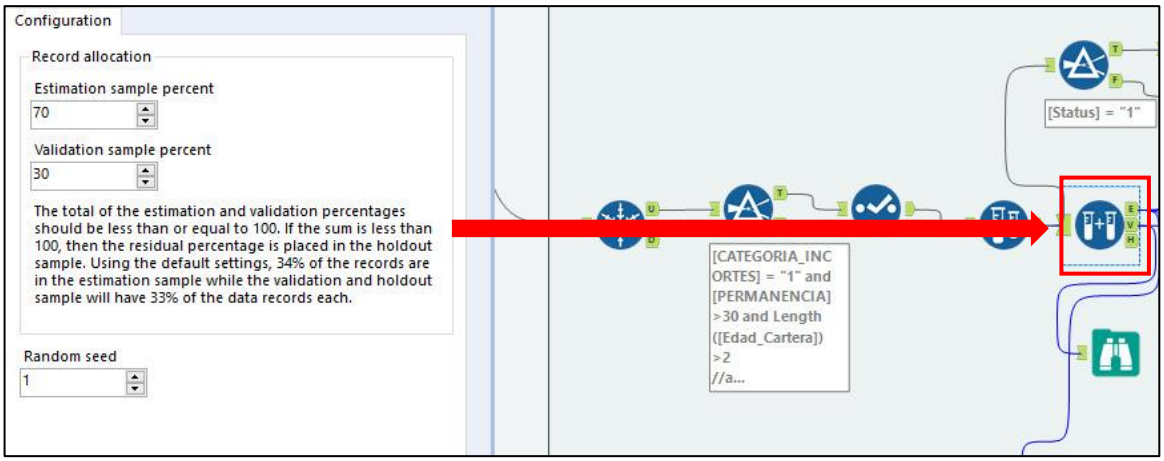


Figura 44: Create Samples - Representación del Train-Test Split en Alteryx.
Realizado por: CHUQUER, William, 2024

3.5. Evaluación

La importancia de esta fase en el proyecto, pues se mide la calidad de las predicciones generadas por los modelos predictivos desarrollados. Tras completar el entrenamiento de los modelos, se realiza una validación utilizando una parte de los registros etiquetados como churn. Esta validación permite evaluar cómo los modelos predicen la cancelación de clientes en datos que no fueron utilizados durante el entrenamiento.

Es fundamental analizar detalladamente los resultados obtenidos para cada modelo y comparar estos resultados con los objetivos del proyecto y del negocio. Esta etapa implica la evaluación exhaustiva de cada modelo mediante la matriz de confusión, la cual

proporciona información sobre la precisión, sensibilidad y especificidad de las predicciones. Este análisis permitirá determinar qué modelo ofrece el mejor rendimiento en la predicción de la cancelación de clientes y alinearse con los objetivos del negocio.

3.5.1. Métricas de desempeño

Para la evaluación de modelos, utilizaremos las siguientes métricas:

- Matriz de confusión
- La técnica de Train-Test Split
- Curva ROC

a) Regresión logística

Cancelaciones administrativas:

Para la regresión logística presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 45, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 16626 y 16182, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of Logistic_Regression_114

	Actual_0	Actual_1
Predicted_0	16626	505
Predicted_1	133	16182

Figura 45: Matriz de Confusión - Regresión Logística: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 16626

Verdaderos Negativos (TN): 16182

Falsos Positivos (FN): 505

Falsos Negativos (FP): 133

- En la posición (0, 0), el valor 16626 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).
- En la posición (0, 1), el valor 505 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).
- En la posición (1, 0), el valor 133 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 16182 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 16626 / (16626+133) = 99,21%

  Recall: 16626 / (16626+505) = 97,05%

En la tabla 7, presentamos las medidas de ajuste y error que el modelo de regresión logística nos entregó.

Tabla 7. Medidas de ajuste y error: R. Logística - Cancelaciones Administrativas

Medidas de ajuste y error – Cancelaciones administrativas					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Regresión Logística	0.9809	0.9807	0.9949	0.9921	0.9697

Como podemos apreciar en la tabla 7, la curva ROC logra un nivel adecuado con un AUC de 0.9949. Esto lo representamos en la figura 46 donde la curva se aproxima a la esquina superior izquierda, lo que indica un excelente rendimiento del modelo en términos de su capacidad para distinguir entre las clases.

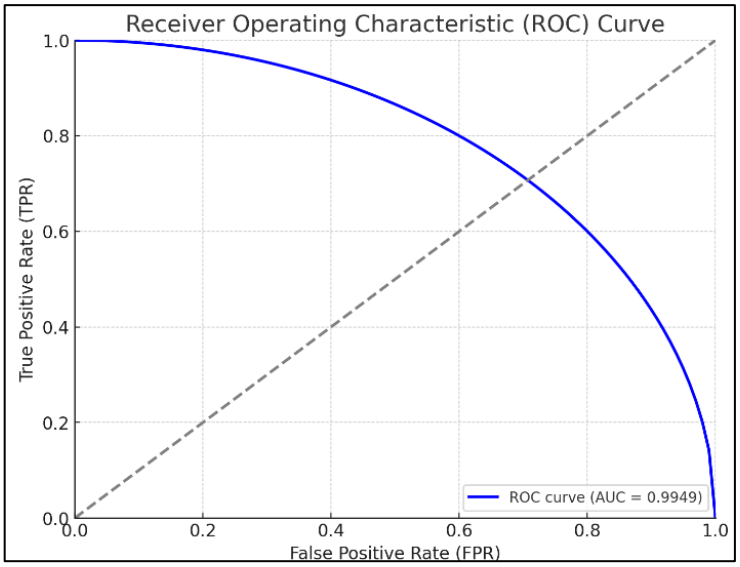


Figura 46: Curva ROC - R. Logística: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Cancelaciones voluntarias:

Para la regresión logística presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 47, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 7411 y 8295, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of Logistic_Regression_114

	Actual_0	Actual_1
Predicted_0	7411	874
Predicted_1	1689	8295

Figura 47: Matriz de Confusión - Regresión Logística: Cancelaciones Voluntarias.
Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 7411

Verdaderos Negativos (TN): 8295

Falsos Positivos (FN): 874

Falsos Negativos (FP): 1689

- En la posición (0, 0), el valor 7411 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).
- En la posición (0, 1), el valor 874 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).

- En la posición (1, 0), el valor 1689 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 8295 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 7411 / (7411+1689) = 81,44%

  Recall: 7411 / (7411+874) = 89,45%

En la tabla 8, presentamos las medidas de ajuste y error que el modelo de regresión logística nos entregó.

Tabla 8. Medidas de ajuste y error: R. Logística - Cancelaciones Voluntarias

Medidas de ajuste y error – Cancelaciones voluntarias					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Regresión Logística	0.8597	0.8662	0.9337	0.8144	0.9047

Como podemos apreciar en la tabla 8, la curva ROC logra un nivel adecuado con un AUC de 0.9337. Esto lo representamos en la figura 48 donde la curva se aproxima a la esquina superior izquierda, lo que indica un buen rendimiento del modelo en términos de su capacidad para distinguir entre las clases.

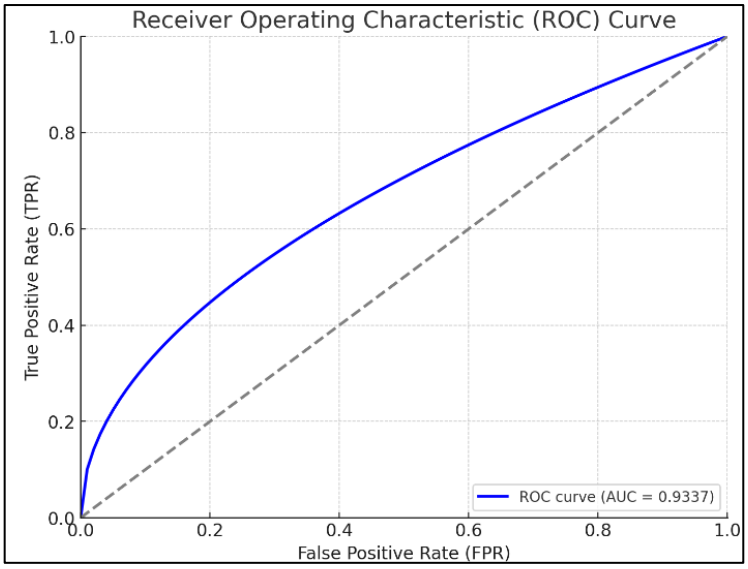


Figura 48: Curva ROC - R. Logística: Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

b) Árboles de decisión

Cancelaciones administrativas:

Para los árboles de decisión presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 49, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 16637 y 16195, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of Decision_Tree_55

	Actual_0	Actual_1
Predicted_0	16637	492
Predicted_1	122	16195

Figura 49: Matriz de Confusión - Árboles de decisión: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 16637

Verdaderos Negativos (TN): 16195

Falsos Positivos (FN): 492

Falsos Negativos (FP): 122

- En la posición (0, 0), el valor 16637 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).
- En la posición (0, 1), el valor 492 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).
- En la posición (1, 0), el valor 122 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 16195 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 16637 / (16637+122) = 99,27%

  Recall: 16637 / (16637+492) = 97,13%

En la tabla 9, presentamos las medidas de ajuste y error que el modelo de árboles de decisión nos entregó.

Tabla 9. Medidas de ajuste y error: A. de decisión - Cancelaciones Administrativas

Medidas de ajuste y error – Cancelaciones Administrativas					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Árboles de decisión	0.9816	0.9814	0.9919	0.9927	0.9705

Como podemos apreciar en la tabla 9, la curva ROC logra un nivel adecuado con un AUC de 0.9919. Esto lo representamos en la figura 50 donde la curva se aproxima a la esquina superior izquierda, lo que indica un buen rendimiento del modelo pero la Regresión Logística es ligeramente superior en términos de su capacidad para distinguir entre las clases.

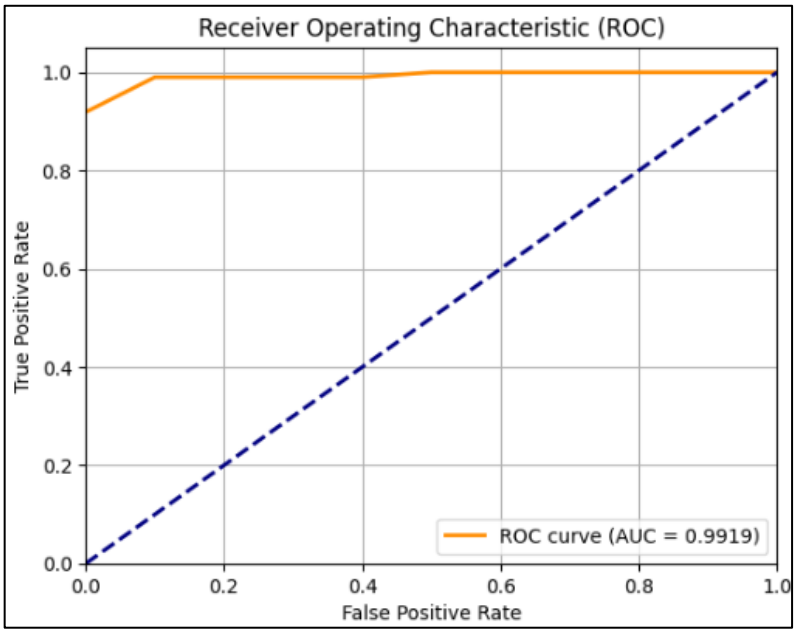


Figura 50: Curva ROC - A. de decisión: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Cancelaciones voluntarias:

Para los árboles de decisión presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 51, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 7390 y 8245, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of Decision_Tree_55

	Actual_0	Actual_1
Predicted_0	7390	924
Predicted_1	1710	8245

Figura 51: Matriz de Confusión - Árboles de decisión: Cancelaciones Voluntarias.
Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 7390

Verdaderos Negativos (TN): 8245

Falsos Positivos (FN): 924

Falsos Negativos (FP): 1710

- En la posición (0, 0), el valor 7390 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).

- En la posición (0, 1), el valor 924 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).
- En la posición (1, 0), el valor 1710 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 8245 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 7390 / (7390+1710) = 81,21%

  Recall: 7390 / (7390+924) = 88,88%

En la tabla 10, presentamos las medidas de ajuste y error que el modelo de árboles de decisión nos entregó.

Tabla 10. Medidas de ajuste y error: A. de decisión - Cancelaciones Voluntarias

Medidas de ajuste y error – Cancelaciones voluntarias					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Árboles de decisión	0.8558	0.8623	0.9175	0.8121	0.8992

Como podemos apreciar en la tabla 10, la curva ROC logra un nivel adecuado con un AUC de 0.9175. Esto lo representamos en la figura 52 donde la curva se aproxima a la esquina superior izquierda, lo que indica un buen rendimiento del modelo aunque un poco menos precisa que la Regresión Logística.

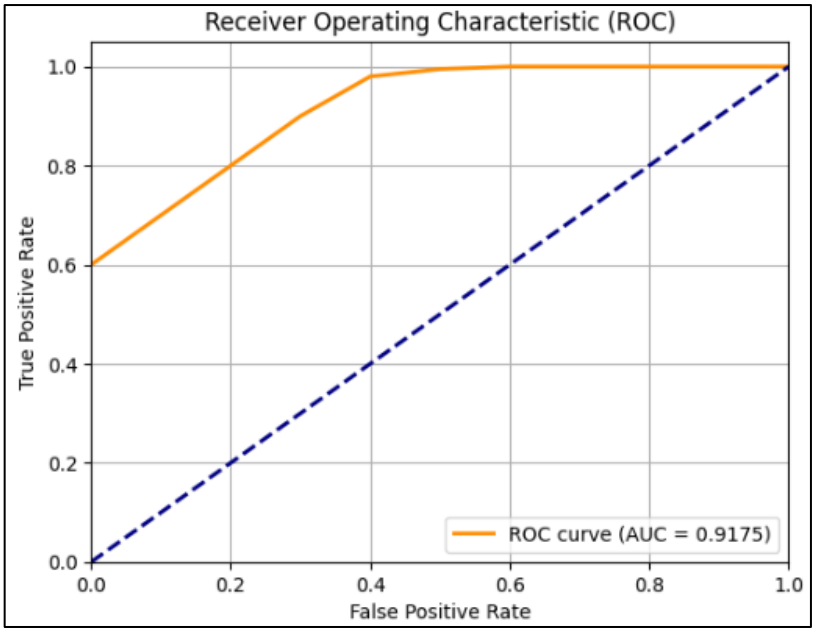


Figura 52: Curva ROC - A. de decisión: Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

c) Random Forest

Cancelaciones administrativas:

Para Random Forest presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 53, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 16672 y 16141, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of ForestModel

	Actual_0	Actual_1
Predicted_0	16672	546
Predicted_1	87	16141

Figura 53: Matriz de Confusión - Random Forest: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 16672

Verdaderos Negativos (TN): 16141

Falsos Positivos (FN): 546

Falsos Negativos (FP): 87

- En la posición (0, 0), el valor 16672 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).
- En la posición (0, 1), el valor 546 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).
- En la posición (1, 0), el valor 87 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 16141 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 16672 / (16672+87) = 99,48%

  Recall: 16672 / (16672+546) = 96,83%

En la tabla 11, presentamos las medidas de ajuste y error que el modelo de random forest nos entregó.

Tabla 11. Medidas de ajuste y error: R. Forest - Cancelaciones Administrativas

Medidas de ajuste y error – Cancelaciones administrativas					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Random Forest	0.9811	0.9808	0.9964	0.9948	0.9673

Como podemos apreciar en la tabla 11, la curva ROC logra un nivel adecuado con un AUC de 0.9964. Esto lo representamos en la figura 54 donde la curva se aproxima a la esquina superior izquierda, lo que indica un excelente rendimiento del modelo siendo este superior a los anteriores modelos en términos de su capacidad para distinguir entre las clases.

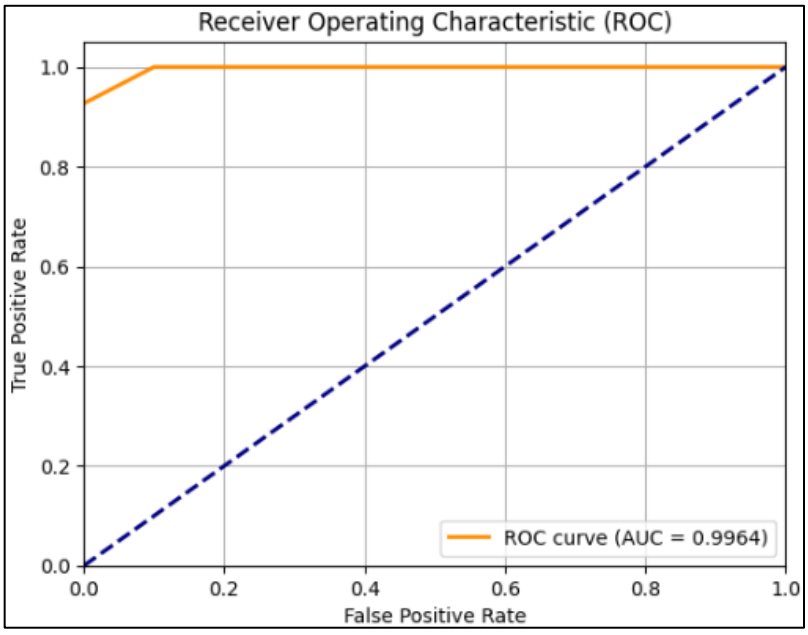


Figura 54: Curva ROC - R. Forest: Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Cancelaciones voluntarias:

Para Random Forest presentamos los resultados obtenidos en la matriz de confusión, las medidas de ajuste y error, y la curva ROC.

En la figura 55, se muestra el resultado de la matriz de confusión, en esta matriz los valores en la diagonal principal 7378 y 8446, representan las instancias correctamente clasificadas por el modelo, ya que los valores en la diagonal principal corresponden a las clasificaciones correctas. En un modelo ideal, estos valores suelen ser los más altos en la matriz de confusión, lo cual cumple para este caso. Además, dado que las variables en el análisis son dicotómicas (Yes/No), los valores asignados en la matriz son Predicted_0 y Predicted_1, donde Predicted_0 representa “No” y Predicted_1 representa “Si”.

Confusion matrix of ForestModel

	Actual_0	Actual_1
Predicted_0	7378	723
Predicted_1	1722	8446

Figura 55: Matriz de Confusión - Random Forest: Cancelaciones Voluntarias.
Realizado por: CHUQUER, William, 2024

Verdaderos Positivos (TP): 7378

Verdaderos Negativos (TN): 8446

Falsos Positivos (FN): 723

Falsos Negativos (FP): 1722

- En la posición (0, 0), el valor 7378 representa el número de instancias correctamente clasificadas como positivas (verdaderos positivos).

- En la posición (0, 1), el valor 723 representa el número de instancias incorrectamente clasificadas como negativas (falsos negativos).
- En la posición (1, 0), el valor 1722 representa el número de instancias incorrectamente clasificadas como positivas (falsos positivos).
- En la posición (1, 1), el valor 8446 representa el número de instancias correctamente clasificadas como negativas (verdaderos negativos).
- Podemos calcular también:

  Precisión: 7378 / (7378+1722) = 81,08%

  Recall: 7378 / (7378+723) = 91,08%

En la tabla 12, presentamos las medidas de ajuste y error que el modelo de random forest nos entregó.

Tabla 12. Medidas de ajuste y error: R. Forest - Cancelaciones Voluntarias

Medidas de ajuste y error – Cancelaciones voluntarias					
Model	**Accuracy**	**F1**	**AUC**	**Accuracy_0**	**Accuracy_1**
Random Forest	0.8662	0.8736	0.9384	0.8108	0.9211

Como podemos apreciar en la tabla 12, la curva ROC logra un nivel adecuado con un AUC de 0.9384. Esto lo representamos en la figura 56 donde la curva se aproxima a la esquina superior izquierda, lo que indica un buen rendimiento del modelo siendo este superior a los anteriores modelos en términos de su capacidad para distinguir entre las clases.

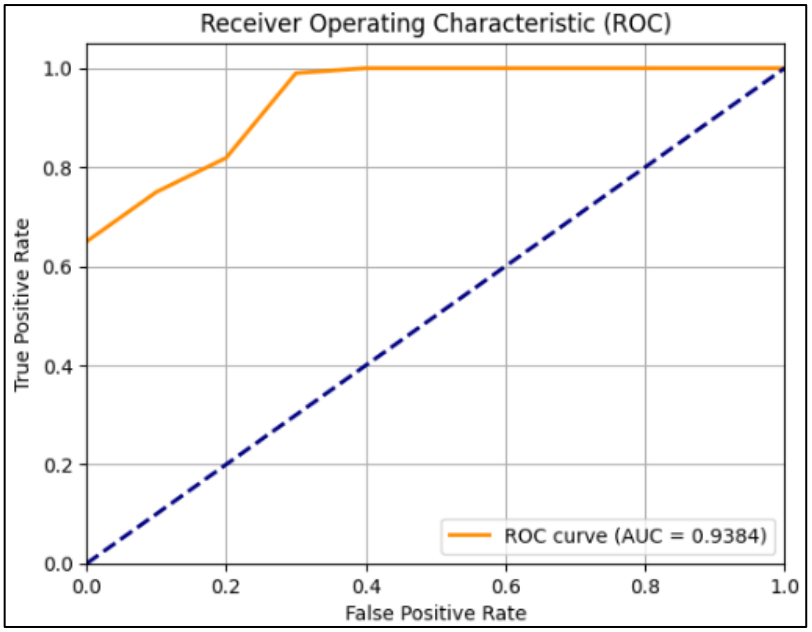


Figura 56: Curva ROC - R. Forest: Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

d) Cuadro comparativo de modelos: Cancelación administrativa

Tabla 13. Comparación modelos entrenados - Cancelación Administrativa

Modelo	Precisión (%)	Curva ROC (%)	Recall (%)
Regresión Logística	99.21%	99.49%	97.05%
Árboles de Decisión	99.27%	99.19%	97.13%
Random Forest	99.48%	99.64%	96.83%

e) Cuadro comparativo de modelos: Cancelación voluntaria

Tabla 14. Comparación modelos entrenados - Cancelación Voluntaria

Modelo	Precisión (%)	Curva ROC (%)	Recall (%)
Regresión Logística	81.44%	93.37%	89.45%
Árboles de Decisión	81.21%	91.75%	88.88%
Random Forest	81.08%	93.84%	91.08%

Según el análisis generado de los modelos adaptados en los cuadros comparativos, se determina que Random Forest es el mejor en términos de rendimiento durante el entrenamiento efectuado con el conjunto de datos utilizado en este estudio. Este modelo es el más adecuado para alcanzar el objetivo de predecir la deserción de clientes, lo que nos permitirá implementar estrategias efectivas para reducir la tasa de abandono.

Consideraciones: En el capítulo 4, donde se presentan los resultados obtenidos, se detallará el modelo seleccionado, explicando sus resultados y justificando porque es el más adecuado para nuestro proyecto.

3.6. Despliegue

En esta fase final del proyecto, se desarrollarán dos flujos los cuales nos permita obtener la información que asocie a la reportería mediante la construcción de dashboards en Qlik Sense. Estos dashboards presentarán los resultados de manera clara y comprensible. Además, se compartirá con las áreas estratégicas informes mensuales para que puedan utilizarlo en la gestión de clientes y en la toma de decisiones efectivas. Esto permitirá una evaluación continua del proceso y del impacto que cause este estudio para reducir la deserción de los clientes.

A continuación (ver figura 57) se presenta el diagrama final del proceso aplicado a la metodología CRISP-DM donde observamos la fase de despliegue. En esta etapa, se integran los resultados de los modelos de Cancelación Administrativa y Voluntaria en una base consolidada denominada **“BaseResultadoPrediccion.yxdb”**. Esta base permitirá utilizar la información obtenida en la fase de modelado para generar los últimos procesos, descargar los resultados del modelo seleccionado (Random Forest), y proceder a construir

el dashboard de la predicción del churn de clientes para facilitar la toma de decisiones sobre los posibles clientes que podrían desertar del servicio de internet.

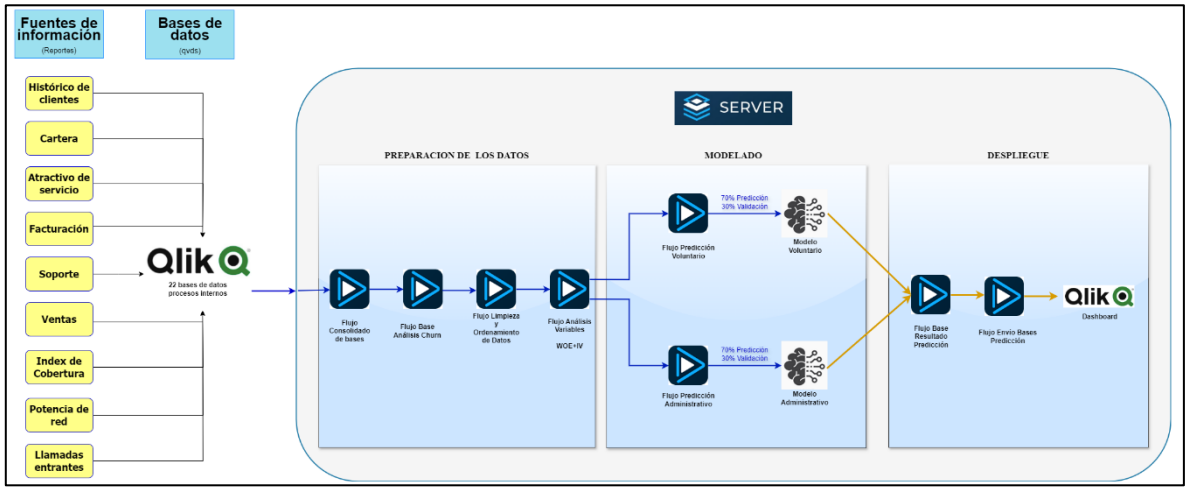


Figura 57: Diagrama de procesos - Fase de despliegue.

Realizado por: CHUQUER, William, 2024

Como mencionamos anteriormente, se genera el último flujo que nos permite enviar la base de predicción a uno de los servidores donde está alojada la herramienta Qlik Sense. Este archivo, inicialmente en formato .csv, se convertirá a .qvd, separando la información según el tipo de cancelación (administrativa o voluntaria). Antes de esta conversión, se realizan varios cálculos, como la generación de un score de cancelaciones, la evaluación de tareas, y la identificación del motivo de cancelación de un cliente, entre otros. Estos cálculos son cruciales para determinar la efectividad del modelo, basándose en datos históricos de deserción del servicio. La figura 58 muestra el flujo de trabajo, que abarca dos procesos principales: el mencionado anteriormente, y la generación de datos segmentados sobre la permanencia del cliente y los posibles factores de cancelación. Este enfoque permite un análisis más completo al momento de crear los qvds finales.

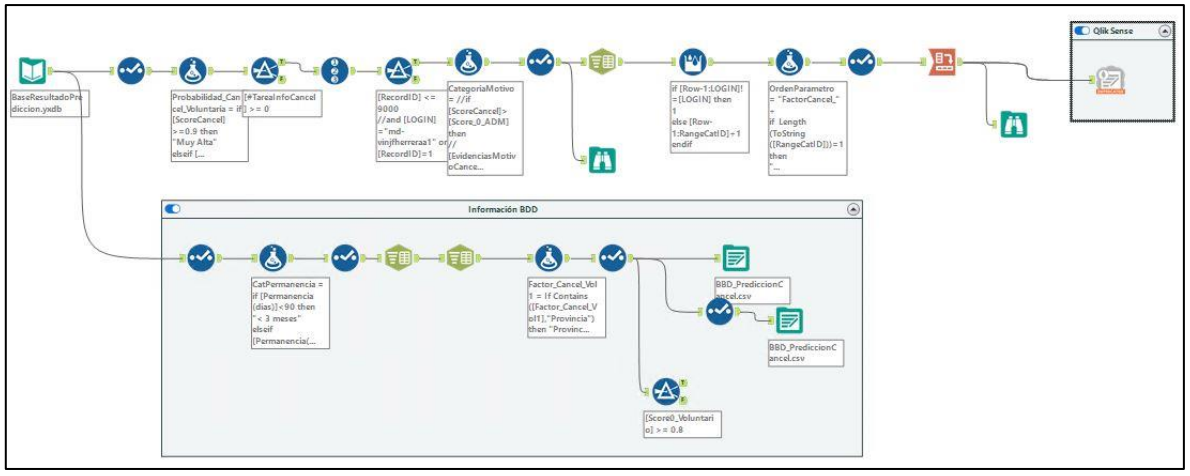


Figura 58: Flujo Fase de Despliegue - Predicción Churn de Clientes.

Realizado por: CHUQUER, William, 2024

Una vez que el flujo finaliza su proceso, se generará automáticamente la base de predicción denominada **"BDD_PrediccionCancel<fecha>.csv"** donde "<fecha>" se reemplazará con la fecha de ejecución del proceso y se almacenará en el servidor (ver figura 59). Además, esta base se transformará en dos archivos .qvd llamados **"INFO_PUNTO_PROBABILIDAD_CANCEL_ALL.qvd"** y **"INFO_PUNTO_PROBABILIDAD_CANCEL_ALL2.qvd"**, que se utilizarán para cargar la información en el reporte elaborado en Qlik Sense (ver figura 60).

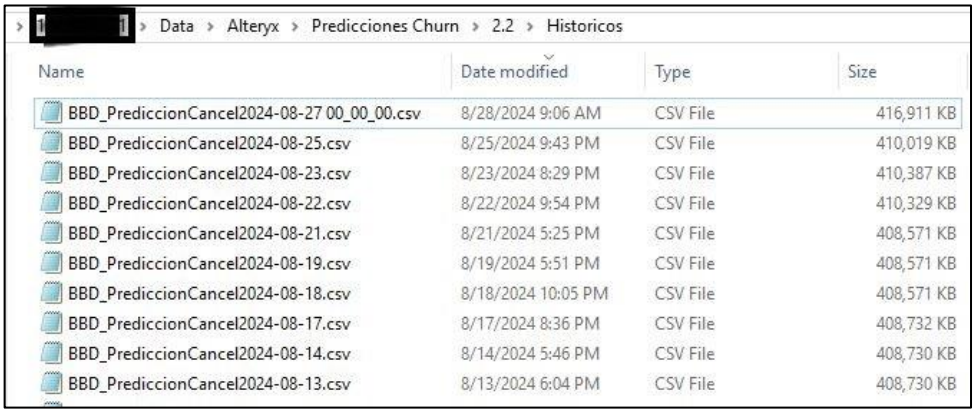


Figura 59: Bases Predicción Cancel.

Realizado por: CHUQUER, William, 2024

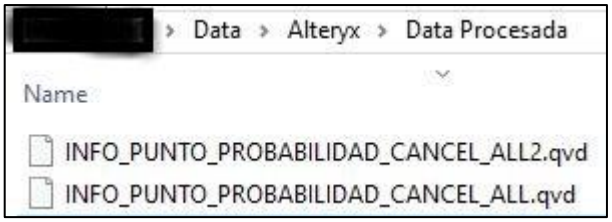


Figura 60: Qvds Predicción Cancel.

Realizado por: CHUQUER, William, 2024

Tras completar el proceso de flujos en la fase de despliegue, se procede a la construcción del dashboard, incorporando la información obtenida en este estudio. Este dashboard permitirá analizar e identificar patrones necesarios para predecir las posibles fugas de clientes (ver figura 61).

Figura 61: Dashboard Churn Rate - Probabilidad de Deserción Voluntaria y Administrativa.

Realizado por: CHUQUER, William, 2024

Consideraciones: En el capítulo 4 de resultados, se analizará el modelo ganador, Random Forest, destacando los resultados que lo posicionaron como el mejor durante el entrenamiento. También se presentará información relevante sobre la construcción del reporte, incluyendo los parámetros clave considerados en el análisis para identificar los factores de probabilidad mediante WOE + IV. Estos análisis contribuyeron significativamente a los procesos desarrollados en cada uno de los flujos aplicados en Alteryx.

CAPÍTULO IV: RESULTADOS

4.1. Clarificando el resultado

Cancelaciones administrativas:

De acuerdo con las métricas de evaluación, el modelo elegido para este estudio es el Random Forest que alcanzó una precisión del 99.48% en los datos de prueba, lo que demuestra su excelente capacidad para predecir la deserción de clientes. Al aplicar el modelo en el dataset de entrenamiento, se obtuvo la precisión del 96.73%. Estos resultados, consistentes en ambos dataset, indican que el modelo ha identificado patrones y características significativas para realizar la clasificación de los clientes que podrían desertar del servicio de internet.

Evaluar el rendimiento del modelo en datos de prueba es esencial para determinar su capacidad de generalización. Si el modelo muestra una precisión comparable en ambos dataset, tanto en el de entrenamiento como en el de prueba, esto sugiere que no está ajustado en exceso a los datos de entrenamiento, lo cual es un indicador favorable.

Dado lo anterior, podemos concluir que nuestro modelo es confiable, con una precisión del 99.5%. Esta confianza se refuerza con la evaluación previa de la matriz de confusión y la curva ROC, resultados que son satisfactorios para el negocio y alineados con el objetivo planteado al inicio de este estudio.

Presentamos a continuación las gráficas de Precision-Recall Curve, Curva ROC de los tres modelos entrenados y la variable de importancia mediante un Índice Gini del modelo apropiado.

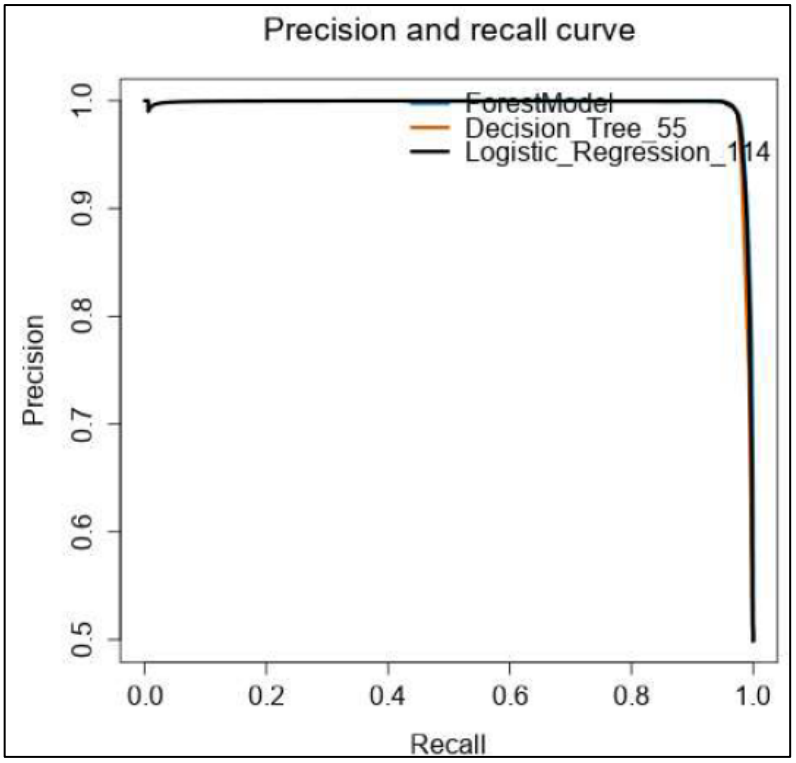


Figura 62: Curva de precisión y recuperación - Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

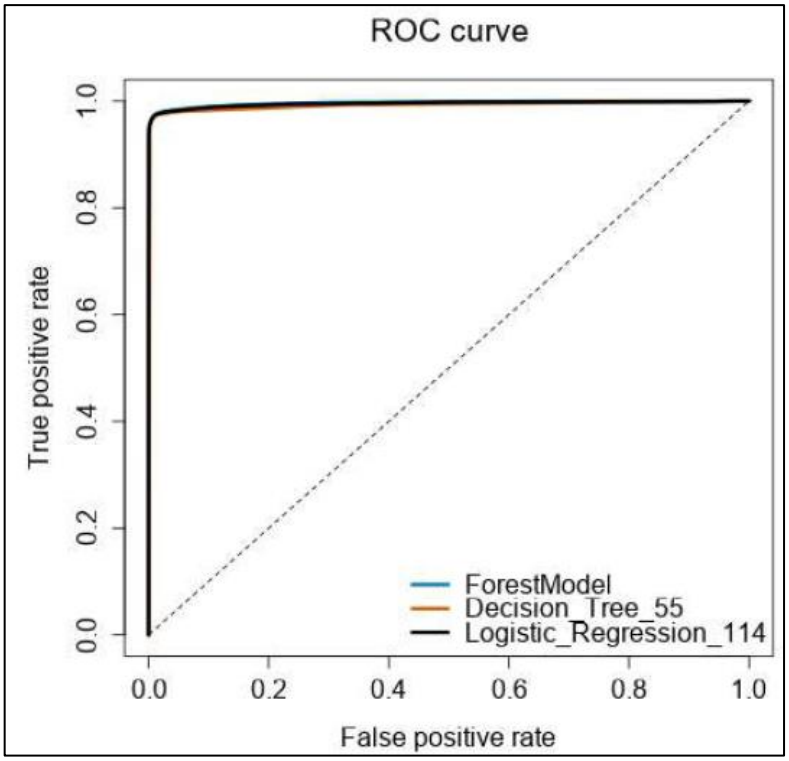


Figura 63: Curva ROC - Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

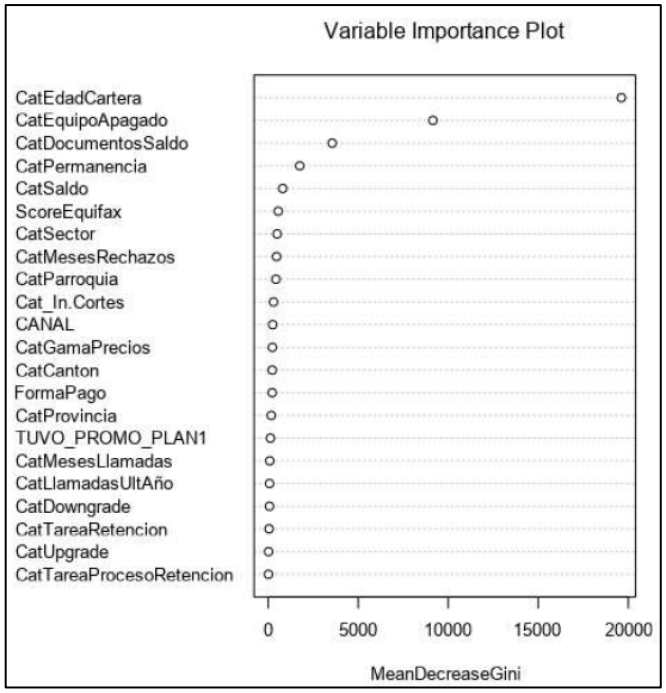


Figura 64: Índice de Gini - Cancelaciones Administrativas.

Realizado por: CHUQUER, William, 2024

Cancelaciones voluntarias:

De acuerdo con las métricas de evaluación, el modelo elegido para este estudio es el Random Forest que alcanzó una precisión del 81.08% en los datos de prueba, lo que demuestra su excelente capacidad para predecir la deserción de clientes. Al aplicar el modelo en el dataset de entrenamiento, se obtuvo la precisión del 92.11%. Estos resultados, consistentes en ambos dataset, indican que el modelo ha identificado patrones y características significativas para realizar la clasificación de los clientes que podrían desertar del servicio de internet.

Evaluar el rendimiento del modelo en datos de prueba es esencial para determinar su capacidad de generalización. Si el modelo muestra una precisión comparable en ambos dataset, tanto en el de entrenamiento como en el de prueba, esto sugiere que no está ajustado en exceso a los datos de entrenamiento, lo cual es un indicador favorable.

Dado lo anterior, podemos concluir que nuestro modelo es confiable, con una precisión del 81%. Esta confianza se refuerza con la evaluación previa de la matriz de confusión y la curva ROC, resultados que son muy buenos para el negocio y alineados con el objetivo planteado al inicio de este estudio.

Presentamos a continuación las gráficas de Precision-Recall Curve, Curva ROC de los tres modelos entrenados y la variable de importancia mediante un Índice Gini del modelo apropiado.

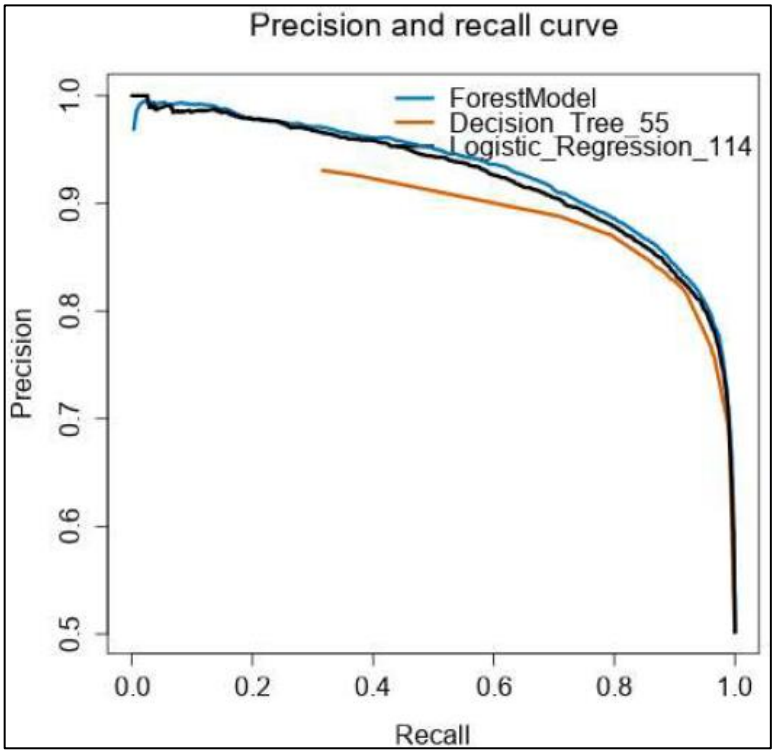


Figura 65: Curva de precisión y recuperación - Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

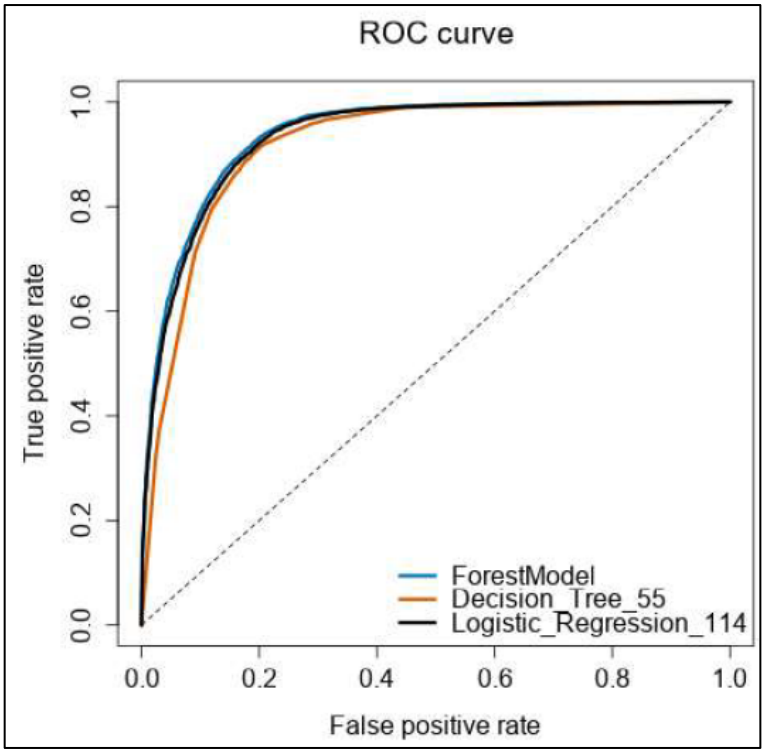


Figura 66: Curva ROC - Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

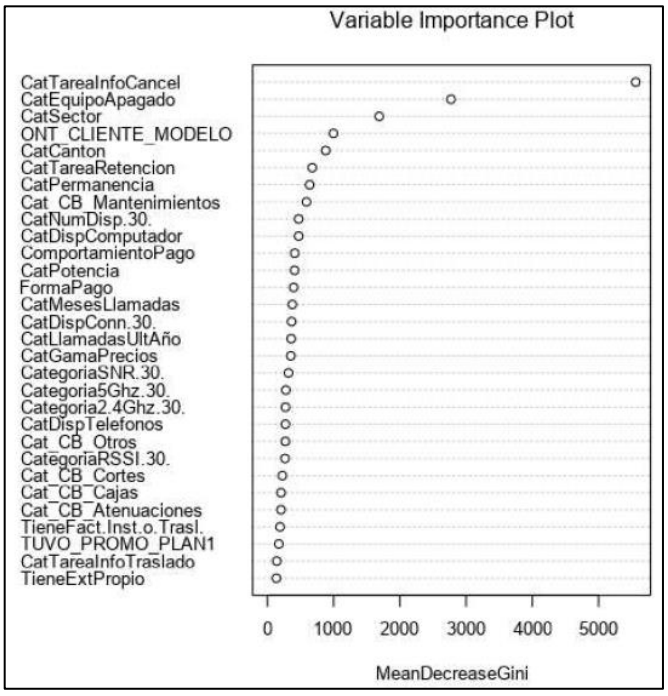


Figura 67: Índice de Gini - Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

Estudio aplicación de parámetros para ambos modelos de cancelaciones:

En el capítulo 3, utilizamos la técnica WOE + IV pero no se profundizó en su aplicación práctica. En esta sección de resultados (ver figura 68), mencionaremos que la importancia de parametrizar, categorizar y segmentar la información fue importante para garantizar que la probabilidad de predecir las cancelaciones, tanto voluntarias como administrativas, permitieron tener mayor precisión y efectividad en los datos.

La correcta aplicación de estas técnicas permitirá al modelo identificar patrones significativos y relevantes para el estudio, mejorando la capacidad de análisis y la utilidad de los resultados obtenidos. Estos factores de probabilidad gracias al uso de WOE + IV, permitieron evaluar y analizar las variables de cada LOGIN permitiendo identificar los factores que influyen en la cancelación de clientes. Los parámetros con WOE negativo se ordenan para ver cuál contribuye más a la cancelación, y se agrupan por motivo para identificar el motivo de cancelación más probable.

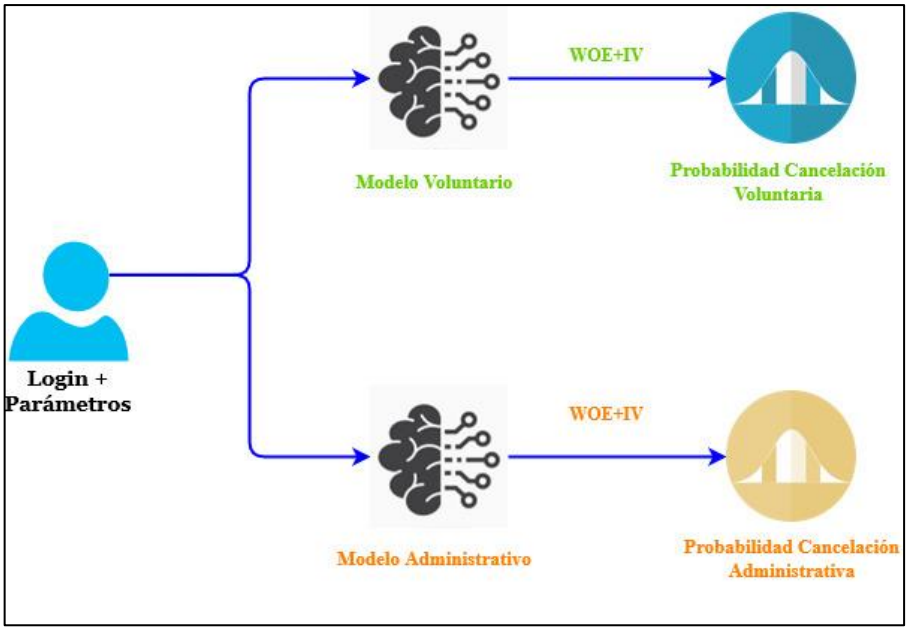


Figura 68: Interpretación de parametrización de los datos utilizando WOE + IV.

Realizado por: CHUQUER, William, 2024

Reporte Churn Rate de Clientes:

Como resultado final obtenido a este estudio, se realizó un reporte basado en la información del Churn Rate de Clientes, el cual facilita la identificación de patrones clave para la toma de decisiones y la detección de posibles factores de riesgo en el comportamiento de los clientes.

Por ejemplo, el reporte incluye un histograma que muestra la probabilidad de cancelación voluntaria de un cliente, categorizada desde Muy Baja hasta Muy Alta (ver figura 69). Al filtrar la categoría Muy Alta (ver figura 70), se identificaron 4940 logins con alta probabilidad de cancelar el servicio voluntariamente. Un mapa de calor (ver figura 71) revela que las zonas de mayor riesgo se concentran en las ciudades de Quito y Guayaquil. Este análisis permite comprender mejor el comportamiento de los clientes y, apoyado en la información obtenida, identificar posibles causas o factores que podrían influir en estas decisiones. Además, se creó una tabla detallada (ver figura 72) para un análisis más profundo de estos comportamientos.

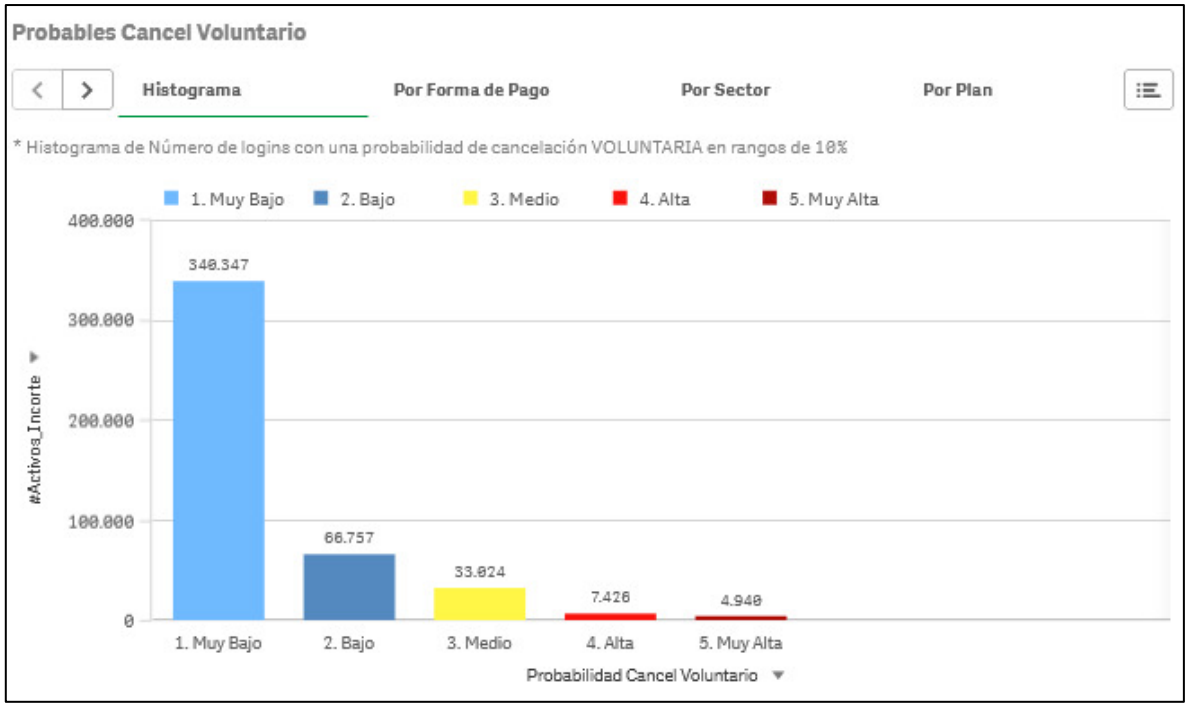


Figura 69: Histograma de Cancelaciones Voluntarias.

Realizado por: CHUQUER, William, 2024

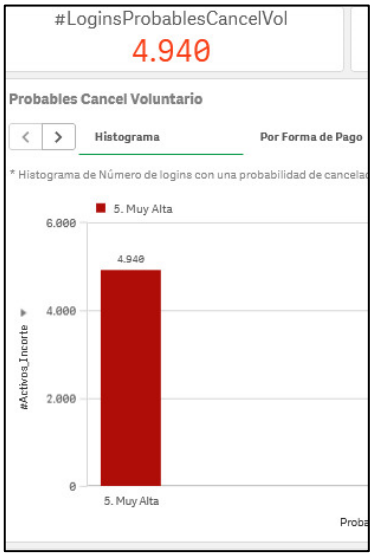


Figura 70: Filtro probabilidad muy alta de cancelación.

Realizado por: CHUQUER, William, 2024

Figura 71: Mapa Georeferencia de localidades con probabilidad de cancelación.

Realizado por: CHUQUER, William, 2024

Logins con probabilidad de cancelación Muy Alta, Alta y Media

Cancel Administrativa | Cancel Voluntaria

Total logins: 4940

LOGIN	Probabilidad Cancel Voluntario	Motivo Probable Cancelación	Factor_Cancel_Vol1
Totales			
md-ambjasilvac1	5. Muy Alta	Cambio de domiciolio	CatTareaInfoCancel:SI
md-durefpenarretat1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-gyedmdelgadoq1	5. Muy Alta	Cambio de domiciolio	CatTareaInfoCancel:SI
md-gyejplorencesc1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-gyeomcataguam2	5. Muy Alta	Cambio de domiciolio	CatTareaInfoCancel:SI
md-gyetmtrianat2	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-ibaaibenavidesp1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-maccjvalarezot1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-amncagalvezl2	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-amnjnvelasteguia1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-atrgfcorreap1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-uioadmontoyac1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-uiojpcastillov2	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-uiomdviterig3	5. Muy Alta	Cambio de domiciolio	CatTareaInfoCancel:SI
md-uiommproanov1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI
md-aanlaramirezm1	5. Muy Alta	Indisponibilidad del Servicio	CatTareaInfoCancel:SI

Figura 72: Tabla detalle de logins con probabilidad de cancelación.

Realizado por: CHUQUER, William, 2024

4.2. Discusión

En este capítulo, se presentan los resultados del estudio sobre la predicción de la deserción de clientes del servicio de internet utilizando los modelos de Regresión Logística, Árboles de decisión y Random Forest, este último mostró las mejores métricas de desempeño. El objetivo fue desarrollar un modelo preciso para identificar clientes propensos a abandonar la marca.

El modelo de Random Forest mostró un rendimiento superior con una precisión del 99.5% en las cancelaciones administrativas y 81% en las voluntarias. Estos resultados destacan la capacidad del modelo para capturar patrones complejos y no lineales en los datos, lo que le permitió generalizar mejor en comparación con otros modelos. Estas cifras fueron respaldadas por métricas de rendimiento sólidas, como la curva ROC y la matriz de confusión, que confirmaron la robustez del modelo, lo cual es una señal de que el modelo no sufre de sobreajuste (overfitting).

El modelo de Regresión Logística, aunque es efectivo en la clasificación binaria y proporciona coeficientes interpretables que permiten entender el impacto de cada variable, no pudo captar de manera tan efectiva las interacciones complejas entre las variables que el Random Forest sí logró identificar. Su desempeño fue razonable, pero quedó por debajo en términos de precisión y sensibilidad comparado con Random Forest.

Por otro lado, el modelo de Árboles de decisión, aunque es más sencillo de interpretar y proporciona un modelo de decisión claro, no alcanzó el mismo nivel de precisión. Su tendencia a sobreajustarse a los datos de entrenamiento limitó su capacidad para generalizar bien a nuevos datos. La simplicidad de este modelo puede ser ventajosa en algunos contextos, especialmente cuando se requiere una interpretación más directa, pero

en este caso, la mayor complejidad del Random Forest demostró ser más adecuada para el problema en cuestión.

La precisión debe interpretarse en función de la calidad de los datos utilizados. Se realizaron esfuerzos significativos en la recopilación y preparación de datos, lo que fortaleció el modelo de Random Forest. Se identificaron factores clave, como el tiempo de permanencia y el valor de inversión, que influyen en la deserción de clientes, proporcionando información valiosa para estrategias de retención. Y, aunque los resultados son prometedores, es importante reconocer ciertas limitaciones, como la variabilidad del modelo en diferentes contextos y la necesidad de actualizaciones periódicas para mantener su precisión. En resumen, el modelo Random Forest ha demostrado ser efectivo, y ofrece una base sólida para estrategias de retención, aunque es importante considerar su aplicabilidad en contextos específicos y actualizaciones continuas.

Para la empresa proveedora del servicio de internet en el Ecuador, el uso del modelo de Random Forest es altamente recomendable debido a su precisión y capacidad para predecir tanto cancelaciones administrativas como voluntarias. Esto le permite a la empresa tomar decisiones proactivas y desarrollar estrategias de retención más efectivas, enfocadas en los clientes con mayor riesgo de desertar.

No obstante, los otros modelos no deben ser descartados por completo. En particular, el Árbol de Decisión podría ser útil en escenarios donde se requiera una explicación más sencilla de los resultados o en casos en que se necesite una implementación rápida y menos compleja. Por su parte, la Regresión Logística podría ser útil para entender de manera más clara el impacto de variables específicas en la cancelación, lo cual podría complementar el análisis del Random Forest.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

El presente estudio, ha empleado el modelo de Random Forest para analizar la deserción de clientes en una empresa proveedora del servicio de internet para hogares en el Ecuador, donde podemos llegar a concluir lo siguiente:

- El modelo Random Forest ha demostrado ser altamente efectivo en la estimación de cancelaciones de clientes, con una tasa de precisión del 99.5% para cancelaciones administrativas y del 81% para cancelaciones voluntarias. Estos resultados subrayan la capacidad del modelo para identificar con precisión a los clientes que podrían abandonar el servicio de internet.
- La consistencia en la precisión tanto en los datos de entrenamiento como en los datos de prueba indica que el modelo no está sobreajustado y tiene una buena capacidad de generalización. Esto es crucial para la aplicación práctica del modelo en situaciones del mundo real.
- La aplicación de técnicas como WOE + IV ha sido fundamental para mejorar la precisión del modelo al identificar patrones significativos en los datos. Esto ha permitido una comprensión más profunda de los factores que contribuyen a la deserción de clientes.
- El análisis del Churn Rate de Clientes reveló que las zonas de mayor riesgo se concentran en Quito y Guayaquil. Además, se ha identificado logins con alta probabilidad de cancelación, proporcionando información valiosa para dirigir estrategias de retención de manera más efectiva.

- A pesar de los resultados positivos, el estudio reconoció limitaciones, como la variabilidad del modelo en diferentes contextos y la necesidad de actualizaciones periódicas. Estos factores deben ser considerados para mantener la eficacia del modelo a lo largo del tiempo.
- El estudio sobre la predicción de la deserción de clientes será fundamental para la empresa, ya que los resultados del análisis predictivo permitieron identificar patrones y características clave de los clientes propensos a cancelar el servicio, facilitando la implementación de estrategias de retención más efectivas. La capacidad de prever las cancelaciones tanto administrativas como voluntarias con alta precisión ofrece a la empresa una ventaja significativa para abordar proactivamente los riesgos de pérdida de clientes y mejorar la satisfacción del cliente.

5.2. Recomendaciones

En este estudio, los resultados obtenidos durante el tiempo del proyecto nos han permitido proponer las siguientes recomendaciones:

- Mediante los resultados del modelo, se recomienda desarrollar e implementar estrategias de retención focalizadas en las áreas de mayor riesgo identificadas (Quito y Guayaquil). Esto podría incluir ofertas personalizadas, programas de fidelización y mejoras en el servicio al cliente para mitigar el riesgo de cancelación.

- Asegurar que el modelo siga siendo relevante y efectivo, es crucial realizar un monitoreo continuo y actualizaciones periódicas del modelo. Esto incluirá la revalidación con datos nuevos y el ajuste de parámetros según sea necesario.
- Utilizar los hallazgos del análisis de Churn Rate para dirigir los recursos de manera más eficiente. Priorizar la atención a clientes con alta probabilidad de cancelación puede optimizar los esfuerzos de retención y reducir el churn.
- Continuar mejorando la calidad de los datos recolectados para asegurar que el modelo se base en información precisa y relevante. Esto incluye la actualización de datos y la incorporación de nuevas variables que puedan influir en el comportamiento del cliente.
- Capacitar al personal de la empresa en el uso y comprensión de los resultados del modelo los cuales pueden ayudar en la implementación efectiva de estrategias de retención y en la toma de decisiones basadas en datos.
- Considerar la evaluación de otros modelos predictivos y técnicas analíticas para comparar su desempeño con Random Forest. Esto puede ofrecer perspectivas adicionales y potencialmente mejorar las predicciones y estrategias de retención.

BIBLIOGRAFÍA

Agencia de Regulación y Control de las Telecomunicaciones. (2020). *Boletín Estadístico del Sector de las Telecomunicaciones.* Obtenido de https://www.arcotel.gob.ec/wp-content/uploads/2020/12/BOLETIN- NOVIEMBRE-2020-25-11-2020.pdf

Alvear, J. O. (2020). *Árboles de decisión y Random Forest*. Obtenido de https://bookdown.org/content/2031/arboles-de-decision-parte-i.html

Beatriz.Gil. (2019). *CRISP-DM: La metodología para poner orden en los proyectos*. Obtenido de https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Branch. (2022). Obtenido de https://branch.com.co/estadisticas-de-la-situacion-digital-en-ecuador-2021-2022/

Canal, P. (18 de nov de 2022). *IEBS*. Obtenido de Churn Rate: Qué es y cómo se calcula: https://www.iebschool.com/blog/que-es-churn-rate-marketing-digital/

CERES. (2023). Obtenido de https://www.redceres.com/post/netlife-se-posiciona-como-el-internet-mas-rapido-del-ecuador-reconocido-por-ookla-por-cinco-anos-co

Chakraborty, A. (05 de 08 de 2021). *Peso de la evidencia (WoE) y valor de la información (IV): ¿cómo usarlo en EDA y la construcción de modelos?* Obtenido de https://anikch.medium.com/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8

Chávez, V. (2018). *Árboles Decisión.* Obtenido de https://rpubs.com/elfenixsoy/arbol-veronica

Claudia L. Hernández G, y. M. (2009). *Hacia una metodología de gestión del conocimiento basada en minería de datos*. Obtenido de chrome-

extension://efaidnbmnnnibpcajpcglclefindmkaj/http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1&isAllowed=y

Ekos. (2023). *Por segundo año consecutivo, Netlife es reconocida como Empresa Socialmente Responsable*. Obtenido de https://ekosnegocios.com/articulo/por-segundo-ano-consecutivo-netlife-es-reconocida-como-empresa-socialmente-responsable

Gangeshwers, D. (2013). E-Commerce or Internet Marketing: A Business Review from Indian Context. International Journal of u- and e- Service, Science and Technology. 6, 187-194. 187-194.

González, L. (2019). *Matriz de Confusión*. Obtenido de https://aprendeia.com/matriz-de-confusion-machine-learning/

IBM. (2024). *¿Qué es el random forest?* Obtenido de https://www.ibm.com/mx-es/topics/random-forest

IBM. (2024). *¿Qué son las redes neuronales?* Obtenido de https://www.ibm.com/es-es/topics/neural-networks

INEC. (2023). *Tecnologías de la información y comunicación.* Obtenido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/TIC/2023/202307_Tecnologia_de_la_Informacion_y_Comunicacion-TICs.pdf

Information LAB. (2024). *The Information Lab Spain*. Obtenido de https://www.theinformationlab.es/blog/data-mining-

ejemplos/#:~:text=Empresas%20como%20Amazon%2C%20Apple%20o,con%20sus%20productos%20y%20servicios.

Jaén, M. E. (04 de Abril de 2019). *Fundamentos Estadísticos para Investigación de la Universidad de Murcia.* Obtenido de https://gauss.inf.um.es/feir/45/

Jélvez Caamaño, A. M. (2014). *SCIELO Venezuela.* Obtenido de http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1316-48212014000300004&lng=es&tlng=es

Kimball, R. &. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. En R. &. Kimball, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.*

Lab, T. I. (02 de Junio de 2021). *The Information Lab*. Obtenido de The Information Lab: https://www.theinformationlab.es/blog/que-es-alteryx-y-para-que-sirve/

Mera, J. (24 de Julio de 2023). *INESEM Bussines School*. Obtenido de https://www.inesem.es/revistadigital/informatica-y-tics/modelos-de-prediccion/

Nalda, V. (2024). *FutureSpace - Machine Learning: Los orígenes y la evolución*. Obtenido de https://www.futurespace.es/machine-learning-los-origenes-y-la-evolucion/

Narkhede, S. (2018). *Understanding AUC - ROC Curve - Towards Data Science*. Obtenido de https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

NPERF. (2023). *Barómetro de las conexiones fijas a Internet en Ecuador.* Obtenido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://media.nperf.com/files/publications/EC/Ecuador-Barometro-internet-fijo-S22022-S12023.pdf

NUNSYS. (28 de Mayo de 2024). *NUNSYS*. Obtenido de https://www.nunsys.com/producto-qlik-sense/

Pete Chapman, J. C. (2007). *Metodología CRISP-DM para minería de datos*. Obtenido de https://www.dataprix.com/es/book/export/html/107

Python, O. (2023). *Logos-World*. Obtenido de https://logos-world.net/python-logo/

Salesforce. (27 de 07 de 2023). *¿Qué es el churn rate (tasa de abandono de clientes) y cómo calcularlo?* Obtenido de https://www.salesforce.com/es/blog/churn-rate-tasa-abandono-clientes/

Sánchez, J. A. (2020). *¿Cómo aprenden las máquinas? Machine Learning y sus diferentes tipos*. Obtenido de https://datos.gob.es/es/blog/como-aprenden-las-maquinas-machine-learning-y-sus-diferentes-tipos#:~:text=Estos%20son%3A%20aprendizaje%20supervisado%2C%20aprendizaje,supervisado%20y%20aprendizaje%20por%20refuerzo.

Shin, T. (2020). *Comprensión de la Matriz de Confusión y Cómo Implementarla en Python*. Obtenido de https://www.datasource.ai/es/data-science-articles/comprension-de-la-matriz-de-confusion-y-como-implementarla-en-python

solutions, l. d. (28 de Mayo de 2024). *lis data solutions*. Obtenido de https://www.lisdatasolutions.com/es/que-es-alteryx/

Vanesa Berlanga Silvente, M. J. (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE. Revista d'Innovació i Recerca en Educació*, 66.

Venturini, S. (2016). *Cross-Validacion for Predictive Analytics*. Obtenido de http://www.milanor.net/blog/cross-validation-for-predictive-analytics-using-r/

ANEXOS

DESARROLLO DEL ANÁLISIS EXPLORATORIO DE DATOS – FASE PREPARACIÓN DE LOS DATOS

```
# -*- coding: utf-8 -*-
"""Analisis_Exploratorio_Datos.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1sWuCQwGwT1XUjHgTM0Ke77CvcNiXqz4V

# **Análisis Exploratorio de Datos**

**Lectura del Dataset**

Se hará un análisis exploratorio a fondo de cada uno de las variables de la base para entender a profundidad su origen, su objetivo, su calidad (comprobación) con el objetivo de obtener una base óptima para entrenar la versión inicial del modelo de churn.
"""

# Importar librerías

import numpy as np                # para arrays
import pandas as pd               # para dataframes
import matplotlib.pyplot as plt   # para gráficos
import seaborn as sns             # para gráficos rápidos
```

```
import os                       # para sistema operativo

# Cargamos el dataset Base_Analisis_Churn.csv
datos = pd.read_csv('Base_Analisis_Churn.csv')

print(datos.shape)
datos.head()

"""**Análisis de cada variable**"""

# Con data.info() podemos ver las variables categóricas
# (Dtype=object)
datos.info()

# Revisamos las primeras filas
datos.head()

# Revisamos las últimas filas
datos.tail()

# Estadísticas descriptivas para las columnas numéricas
datos.describe()

# Crear un DataFrame de resumen
resumen = pd.DataFrame({
    'Nombre Columna': datos.columns,
```

```
    'Tipo de Dato': datos.dtypes,
    'Valores Nulos': datos.isnull().sum(),
    'Valores Únicos': datos.nunique()
})

# Mostrar el resumen completo
print(resumen)

"""En esta sección se analizará la base previamente cargada. Se analizará cada una de las features
para entenderla mejor"""

# Ver la forma del dataset

print( f"Features (columnas): {datos.shape[1]}")

print( f"Registros (filas): {datos.shape[0]}")
print( f"Celdas: {datos.shape[0] * datos.shape[1]}")

print(f"Porcentaje de NaN: {datos.isna().mean().mean() * 100  :.2f}%")

# Seleccionar las columnas categóricas
categorical_columns = datos.select_dtypes(include=['object']).columns

# Resumen detallado para cada variable categórica
for col in categorical_columns:
    print(f'Columna: {col}')
```

```
    print(f'Conteo de Valores: {datos[col].count()}')
    print(f'Valores Únicos: {datos[col].nunique()}')
    print(f'Valor Más Frecuente (Top): {datos[col].mode()[0]}')
    print(f'Frecuencia del Valor Más Frecuente: {datos[col].value_counts().max()}')
    print('-'*40)

# Convertir cadenas vacías y espacios en blanco en NaN
datos.replace(r'^\s*$', np.nan, regex=True, inplace=True)

# Contar los valores NaN en cada columna
valores_nan = datos.isnull().sum()

# Filtrar solo las columnas que tienen NaN
valores_nan = valores_nan[valores_nan > 0]

# Mostrar resumen de columnas con valores NaN
print("Resumen de valores NaN por columna:")
print(valores_nan)

# Identificación de valores faltantes
valores_faltantes = datos.isnull().sum()
print(valores_faltantes[valores_faltantes > 0])

# Opcional: Visualización de valores faltantes
import seaborn as sns
import matplotlib.pyplot as plt
```

```
plt.figure(figsize=(10, 6))
sns.heatmap(datos.isnull(), cbar=False, cmap='viridis')
plt.show()

# Histograma de todas las variables numéricas
datos.hist(bins=20, figsize=(10, 15), layout=(13, 5), edgecolor='black')
plt.tight_layout()  # Ajusta el espaciado automáticamente
plt.show()

# Distribución de variables categóricas
categorical_columns = datos.select_dtypes(include=['object']).columns

for col in categorical_columns:
    plt.figure(figsize=(10, 5))
    sns.countplot(y=col, data=datos)
    plt.show()

# Overview de las features
for i in datos.columns:
    contador_nan = datos[i].isna().sum()
    contador_dash = (datos[i] == "-").sum()
    contador_cero = (datos[i] == "0").sum()  + (datos[i] == 0).sum()
    registros = datos.shape[0]

    print(
```

```
        # General
        f"Columna:      \033[1m{i.upper()}\033[0m",
        f"Tipo de Dato: {datos[i].dtype}",

        "",

        # Valores Nulos
        f"NaN:          {contador_nan}      {contador_nan/ registros * 100 :.2f}%",
        f"Dash:         {contador_dash}      {contador_dash / registros * 100 :.2f}%",
        f"Total Nulos:  {contador_nan + contador_dash}      {(contador_nan + contador_dash)/
registros * 100 :.2f}%",
        f"Ceros:        {contador_cero}      {contador_cero / registros * 100 :.2f}%",

        "",

        # Valores No Nulos
        f"Valores Únicos:   {datos[i].nunique()}",
        f"Distribución: \n\n{datos[i].value_counts()}",

        # Espacio
        "\n\n\n",

        sep = "\n"
        )
```

```
    # Estadística para valores numéricos
    if datos[i].dtype != "object":
        print(

            f"Máx.:     {datos[i].max()}",
            f"Media:    {datos[i].mean()}",
            f"Mediana:  {datos[i].median()}",
            f"Mín.:     {datos[i].min()}",
            "",
            sep = "\n"
        )

    # Graficar distribución
    if datos[i].nunique() > 100 and datos[i].dtype == "object":

        print("Gráfico complejo de procesar", "\n\n\n", )

    else:
        # Manejar valores no numéricos antes de graficar
        if datos[i].dtype == 'object':
            numeric_data = pd.to_numeric(datos[i], errors='coerce').dropna()  # Convertir a numérico,
ignorando errores
            if numeric_data.size > 0:  # Comprueba si quedan valores numéricos
                plt.figure(figsize=(9,2))
                sns.histplot(data=numeric_data, bins=10)
                plt.xticks(rotation=90)
```

```
            plt.show()
        else:
            print(f"Columna '{i}' no contiene valores numéricos para graficar.\n\n\n")
    else:
        plt.figure(figsize=(9,2))
        sns.histplot(data = datos[i], bins = 10)
        plt.xticks(rotation=90)
        plt.show()

# Seleccionar solo columnas numéricas
datos_numericos = datos.select_dtypes(include=['float64', 'int64'])

# Calcular la matriz de correlación
matriz_correlacion = datos_numericos.corr()

# Visualización de la matriz de correlación
plt.figure(figsize=(12, 8))
sns.heatmap(matriz_correlacion, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Matriz de Correlación de Variables Numéricas')
plt.show()

# Relación entre dos variables numéricas
sns.pairplot(datos, hue='ComportamientoPago')
plt.show()

# Distribución de churn por diferentes características categóricas
```

```
for col in categorical_columns:
    if col != 'ComportamientoPago':
        plt.figure(figsize=(10, 5))
        sns.barplot(x=col, y='ComportamientoPago', data=datos)
        plt.show()

# Boxplot para detectar outliers
plt.figure(figsize=(12, 6))
sns.boxplot(data=datos.select_dtypes(include=['float64', 'int64']))
plt.show()
```