PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

TRABAJO DE TITULACIÓN MAESTRÍA EN SISTEMAS DE INFORMACIÓN
MENCIÓN DATA SCIENCE

"CLASIFICACIÓN DE CLIENTES Y PREDICCIÓN DE DESERCIONES USANDO ALGORITMOS K-MEANS Y REGRESIÓN LOGÍSTICA"

PREVIO LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

Autor: Raúl Torres F.

Director: Msc. Jorge Calderón

Quito – Junio 2023

Dedicatoria

"A mis padres, Martha y Raúl por su amor incondicional, por enseñarme desde niño a valorar la vida, luchar por mis sueños y a ser feliz. A mis hermanos, Liliana y Diego, porque siempre han sido un ejemplo de superación, esfuerzo y trabajo, y a mi hijo Raúl Ismael por enseñarme cada día lo hermoso que es ser padre."

Agradecimiento

Quiero agradecer sinceramente a Dios, por siempre cuidarme, guiarme y por nunca haber soltado mi mano en los momentos más difíciles.

Índice de Gráficos

Índice de Tablas

CLASIFICACIÓN DE CLIENTES Y PREDICCIÓN DE DESERCIONES USANDO ALGORITMOS K-MEANS Y REGRESIÓN LOGÍSTICA

Resumen

El análisis de la deserción de clientes se ha vuelto un desafío común en muchas organizaciones, dado que comprender a fondo las razones detrás de la pérdida de clientes resulta crucial para el éxito y la continuidad de cualquier empresa. En este contexto, este trabajo propone el uso de dos enfoques complementarios: el algoritmo no supervisado K-Means y el algoritmo supervisado Regresión Logística.

Por un lado, el algoritmo K-Means permite explorar datos no etiquetados en busca de patrones y segmentos ocultos que puedan revelar información valiosa sobre la deserción de clientes. Por otro lado, la Regresión Logística se emplea para construir modelos que predigan la probabilidad de que un cliente abandone su servicio, basándose en variables relevantes del negocio.

La combinación de estos enfoques de análisis brinda una comprensión más profunda y precisa del comportamiento de los clientes, lo cual resulta fundamental para que las empresas puedan adaptar estrategias efectivas de retención y fidelización permitiéndoles tomar decisiones informadas y proactivas, implementando acciones que reduzcan la pérdida de clientes y aumenten la satisfacción y lealtad.

Summary

Customer churn analysis has become a common challenge in many organizations, as understanding the underlying reasons behind customer loss is crucial for the success and continuity of any company. In this context, this study proposes the use of two complementary approaches: the unsupervised algorithm K-Means and the supervised algorithm Logistic Regression.

On one hand, the K-Means algorithm allows for exploring unlabeled data to uncover hidden patterns and segments that can reveal valuable information about customer churn. On the other hand, Logistic Regression is employed to build models that predict the probability of a customer abandoning their service, based on relevant business variables.

The combination of these analytical approaches provides a deeper and more accurate understanding of customer behavior, which is essential for companies to adapt effective retention and loyalty strategies. This enables them to make informed and proactive decisions, implementing actions that reduce customer churn and increase satisfaction and loyalty.

1. Planteamiento del Problema

¿De qué forma las empresas de telecomunicaciones identifican, controlan y plantean estrategias para que sus clientes permanezcan con sus servicios contratados la mayor cantidad de tiempo posible? Esta es uno de los principales cuestionamientos que las empresas se preguntan día a día, esto debido a que los clientes grandes o pequeños son la fuente principal de sus ingresos con los cuales la empresa sustenta su continuidad y proyecta su crecimiento; tanto así que, si sus clientes dejan de consumir sus productos o servicios, la empresa inmediatamente empieza a perder ingresos causando inestabilidad económica y peligrando a futuro su continuidad.

En el caso específico de las empresas que ofrecen productos y servicios de telecomunicaciones, las cuales cuentan con una gran cantidad de datos tanto de sus clientes como del uso sus productos y servicios, se deben plantear la necesidad de implementar tecnologías de machine learning que les permita trabajar con los datos de sus usuarios con el fin de identificar el o los tipos de clientes que usan sus productos y servicios, para posterior determinar que clientes pueden terminar cancelando sus servicios o no.

Dada esta problemática, las empresas deben contar con herramientas de análisis de datos que les permita implementar algoritmos de machine learning, con los cuales puedan extraer un conocimiento propio de sus datos, lo cual otorgue a las empresas ventajas competitivas para identificar ciertas condiciones únicas de sus clientes. Permitiendo a las empresas plantear estrategias de fidelización, retención, aseguramiento de cartera y creación de productos y servicios optimizados que se ajusten a la realidad de sus clientes actuales y mercado futuro.

1.1. Objetivos

1.1.1. General

Desarrollar un modelo de segmentación con el algoritmo no supervisado K-Means y predicción de deserción de clientes mediante el algoritmo de Regresión Logística.

1.1.2. Específicos

- Recolección del conjunto de datos considerando la mayor cantidad de variables relevantes acerca del cliente y sus servicios.
- Análisis exploratorio descriptivo del conjunto de datos.
- Identificar las variables a ser usadas en los modelos.
- Diseñar un modelo que agrupe a los clientes por características en común con el fin de identificar patrones comunes para entender características específicas de cada grupo.
- Predecir la probabilidad de que un cliente pueda cancelar o no su servicio.

1.1.3. Justificación

El principal objetivo de una empresa no solamente es el obtener nuevos clientes para generar recursos, sino también, conseguir que sus clientes se mantengan la mayor cantidad de tiempo usando y adquiriendo nuevos productos o servicios, debido a esto, las

empresas deben plantearse de manera continua estrategias que les permitan no solamente ofrecer un buen servicio o productos de calidad, sino también, puedan conocer al cliente con el fin de cubrir sus expectativas considerando que en la actualidad los clientes tienen la posibilidad de influenciar sobre el resto de personas mediante el uso de las redes sociales con un gran alcance y en un menor tiempo, de tal forma que estos puedan convertirse en detractores o promotores de la marca impulsando o no a nuevas personas para adquirir sus productos o servicios. Por esta razón, las empresas deben enfocar sus esfuerzos en identificar a sus clientes en base a características comunes de comportamiento con el fin de estructurar estrategias específicas optimizando recursos tanto económicos como humanos e incrementando su efectividad; al final el entendimiento del cliente permite a las empresas reducir la deserción de sus clientes, crecer en el mercado y a la vez crear nuevos productos y servicios personalizados en base a la necesidad específica del cliente, todo esto permitirá a la empresa posicionarse gracias a su reputación de buen manejo de clientes.

El presente proyecto plantea el desarrollo de un modelo de clusterización el cual permitirá clasificar a los clientes considerando comportamientos en base a características en común, adicional sobre los datos clasificados, se plantea el desarrollo de un modelo de predicción que permita estimar la cantidad de deserciones.

2. Estado del Arte

Existen diversos estudios sobre análisis de datos para segmentar y determinar la probabilidad de abandono de clientes, los mismos que presentan puntos de vista particulares según su autor y uso de distintos algoritmos de machine learning, así también, se puede evidenciar que los trabajos han sido desarrollados en diferentes sectores comerciales.

2.1. Investigaciones Previas

(Alegre, 2020) El trabajo de fin de Máster desarrollado en la Universidad Complutense de Madrid denominado, “Predicción de abandono de clientes en una empresa de telecomunicaciones.”, basa su análisis en un conjunto de datos extraída del sistema CRM de una compañía de telecomunicaciones y servicios asociados de Estados Unidos, en donde como objetivo principal se plantea “Desarrollar un modelo de predicción de clientes de una compañía de telecomunicaciones, utilizando diferentes variables de tipo demográficas, de uso de los servicios y vinculación con la compañía.” para lo cual el autor, hace uso de técnicas de modelado como Regresión Logística, Redes Neuronales, Arboles de Decisión, Bagging, Random Forest, Gradient Boosting, Extreme Gradient Boosting, Support Vector Machine, Ensamblado y Validación Cruzada Repetida los cuales fueron evaluados con diferentes métricas de comparación como Accuracy, ROC y Tasa de Fallos en la predicción; por otra parte, uno de sus objetivos específicos indica “Segmentar el conjunto de clientes en clústeres que sean heterogéneos entre sí.” para lo cual el autor usa el algoritmo no supervisado K-Means el cual será usado como complemento al mejor modelo de predicción de Churn escogido de tal forma que agrupe a los clientes en base a características similares con el fin de aplicar estrategias de retención focalizadas en donde se pueda ofrecer productos o servicios optimizados a la necesidad de los clientes, establecer medios de comunicación específicos y brindar una atención diferenciada por tipo de cliente.

Por otra parte, el trabajo de titulación presentado en la Universidad de Bogotá Jorge Tadeo Lozano plantea el tema “Predicción de abandono de clientes en telecomunicaciones

mediante el aprendizaje automático", el cual plantea como objetivo principal "Implementar y evaluar un modelo de aprendizaje automático que permita identificar los clientes propensos a presentar Churn en el segmento pospago, a través de la recopilación de datos históricos.". Este trabajo compara ocho algoritmos de machine learning y determina el idóneo en base a la puntuación de precisión del área bajo la curva ROC en donde los resultados muestran que XGBClassifier, GradientBoosting Classifier, RandomForestClassifier y MLPClassifier se encuentran sobre en el 78%, Regresión Logística y LinearDiscriminantAnalysis en el 76% y el resto de los algoritmos GaussianNB y KNeighborsClassifier con el 74% y 73% respectivamente. (Falla, 2021).

2.2. Marco Teórico

2.2.1. Machine Learning

En español es conocido también como Aprendizaje Automático, el cual forma parte de las ciencias de la computación y también es un disciplina de las ramas de la inteligencia artificial la cual con el uso de algoritmos otorga a las computadoras la capacidad de entender su entorno mediante el análisis de datos.

> En Machine Learning, se asigna un programa de computadora para realizar algunas tareas y se dice que la máquina ha aprendido de su experiencia si su desempeño medible en estas tareas mejora a medida que gana más y más experiencia en la ejecución de estas tareas. Entonces, la máquina toma decisiones y hace predicciones / pronósticos basados en datos. (Ray, 2019, p. 35)

2.2.2. Aprendizaje Supervisado

Este tipo de algoritmo necesita de asistencia externa y realiza el análisis de los datos a partir de un conjunto de datos que es divido en dos partes, una para entrenamiento y otra para prueba con el objetivo de predecir valores futuros en base a lo aprendido en el proceso de entrenamiento. Mahesh (2019) menciona: "El conjunto de datos de entrenamiento tiene una variable de salida que debe predecirse o clasificarse. Todos los algoritmos aprenden un tipo de patrones del conjunto de datos de entrenamiento y aplica al conjunto de datos de prueba para predicción o clasificación" (p 381).

2.2.3. Aprendizaje no Supervisado

Este tipo de algoritmos no necesita contar con un conjunto de datos etiquetado. Bonaccorso (2017) afirma que: "Este enfoque se basa en la ausencia de cualquier supervisor y por lo tanto de medidas de error absoluto; es útil cuando es necesario aprender cómo se pueden agrupar un conjunto de elementos según su similitud" (p 12).

2.2.4. Regresión Logística

Los algoritmos de regresión logística devuelven un resultado binomial, es decir, su resultado muestra la probabilidad de que un evento ocurra o no. Ray (2019) menciona: "La regresión logística se utiliza para tratar un problema de clasificación. Da el resultado binomial ya que da la probabilidad de que ocurra un evento o no (en términos de 0 y 1) en función de los valores de las variables de entrada." (p 36-37).
Ray (2019) afirma: El algoritmo de regresión logística tiene las siguientes ventajas y desventajas de uso:

Ventajas

- Simplicidad de implementación.
- Eficiencia computacional.
- Eficiencia desde la perspectiva del entrenamiento.
- No se requiere escalado para las entidades de entrada.
- Utilizado para resolver problemas de escala industrial.
- Dado que el resultado de la regresión logística es una puntuación de probabilidad, para aplicarlo a la resolución de un problema empresarial es necesario especificar métricas de rendimiento personalizadas para obtener un punto de corte que pueda utilizarse para clasificar el objetivo.
- La regresión logística no se ve afectada por un pequeño ruido en los datos y la multicolinealidad.

Desventaja

- La capacidad para resolver problemas no lineales, ya que su superficie de decisión es lineal, propensa a sobre ajustarse, no funcionará bien a menos que se identifiquen todas las variables independientes. (p 37)

2.2.5. K-Means

Los algoritmos de clusterización son un tipo de algoritmos no supervisados que son utilizados para resolver problemas de agrupamiento. (Ray, 2019).

> K-means es uno de los algoritmos de aprendizaje no supervisado más simples que resuelven el conocido problema de agrupamiento. El procedimiento sigue una forma simple y fácil de clasificar un conjunto de datos dado a través de un cierto número de conglomerados. La idea principal es definir k centros, uno para cada grupo. Estos centros deben colocarse de manera astuta porque la ubicación diferente provoca resultados diferentes. Por lo tanto, la mejor opción es colocarlos lo más lejos posible de entre sí. (Mahesh, 2018, p.383)

Este tipo de algoritmos permite asociar en grupos específicos nuestros datos en base a características en común, matemáticamente esto nos indica que, “La agrupación se realiza minimizando la suma de los cuadrados de las distancias entre los datos y el centroide del clúster correspondiente. Por lo tanto, el propósito del agrupamiento de K-media es clasificar los datos.” (Teknomo, 2007, p.1).

2.2.6. Jupyter Notebook

Es una aplicación web que puede hacer uso de varios kernels y no necesariamente limitarse a uno en específico, permitiendo ejecutar proyectos de ciencia de datos y simulación numérica; a la vez permite al analista documentar el código escrito y finalmente visualizar los resultados.

Jupyter es un proyecto de código abierto que tiene como objetivo permitir el uso de diferentes lenguajes de programación en una plataforma computacional. De hecho, su nombre proviene de tres de los lenguajes de programación más populares para la

computación científica: Julia, Python y R. Está construido sobre la interfaz web de Jupyter Notebook, que reúne las posibilidades de ejecutar código, incluyendo texto junto con ecuaciones LaTeX (a través de MathJax), video y todo lo que se puede visualizar dentro de un navegador. (Cabrera y Diaz, 2018, p.1)

Entre los principales usos de una Jupyter Notebook están:

- Análisis de big data
- Modelización estadística
- Diseño, programación y entrenamiento de modelos basados en aprendizaje automático
- Visualización de datos

Adicional hay que mencionar que Jupyter Notebook es gratuito, de código abierto (licencia BSD modificada), trabaja en el explorador, permite colaboraciones mediante Jupyter Hub y soporta más de 50 lenguajes de programación, entre otros.

2.2.7. NumPy

NumPy es uno de los módulos de Python que trabaja con objetos de matrices multidimensionales al igual que con funciones matemáticas de alto nivel. NumPy org (2022) menciona:

> NumPy es el paquete fundamental para la computación científica en Python. Es una biblioteca de Python que proporciona un objeto de matriz multidimensional, varios objetos derivados (como matrices y matrices enmascaradas), y una variedad de rutinas para operaciones rápidas en matrices, incluyendo matemáticas, lógica, manipulación de formas, clasificación, selección, E/S, transformaciones discretas de Fourier, álgebra lineal básica, operaciones estadísticas básicas, simulación aleatoria y mucho más.

Por otra parte, hay que mencionar que NumPy no está incluido al momento de instalar Python, sino que debe ser instalado posterior a la instalación de Python; por otro lado, al ser escrito en su mayoría con lenguaje C otorga gran velocidad de ejecución al operar funciones matemáticas y numéricas. (González, 2023.)

2.2.8. Pandas

Es una de las aplicaciones más populares y usadas para el análisis y manipulación de datos de código abierto. Entre sus principales características están:

- Un objeto DataFrame rápido y eficiente para la manipulación de datos con indexación integrada.
- Herramientas para leer y escribir datos entre estructuras de datos en memoria y diferentes formatos: CSV y archivos de texto, Microsoft Excel, bases de datos SQL y el formato rápido HDFS.
- Alineación inteligente de datos y manejo integrado de datos faltantes: obtenga una alineación automática basada en etiquetas en los cálculos y manipule fácilmente los datos desordenados en una forma ordenada;
- Transformación y rotación flexibles de conjuntos de datos.

- Segmentación inteligente basada en etiquetas, indexación elegante y creación de subconjuntos de grandes conjuntos de datos
- Las columnas se pueden insertar y eliminar de las estructuras de datos para la mutabilidad del tamaño.
- Agregar o transformar datos con un potente grupo por motor que permite operaciones de división, aplicación y combinación en conjuntos de datos.
- Fusión y unión de conjuntos de datos de alto rendimiento.
- La indexación de ejes jerárquicos proporciona una forma intuitiva de trabajar con datos de alta dimensión en una estructura de datos de menor dimensión.
- Funcionalidad de serie temporal: generación de rango de fechas y conversión de frecuencia, estadísticas de ventanas móviles, cambio de fecha y retraso. Incluso cree compensaciones de tiempo específicas del dominio y únase a series de tiempo sin perder datos.
- Altamente optimizado para el rendimiento, con rutas de código críticas escritas en Python o C.
- Python con pandas se usa en una amplia variedad de dominios académicos y comerciales, que incluyen finanzas, neurociencia, economía, estadísticas, publicidad, análisis web y más. (Pandas.Pydata.org, 2023)

2.2.9. R

R es un software libre para computación estadística y gráficos que proporciona una amplia variedad de técnicas estadísticas (modelación lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupación, ...) y técnicas gráficas, y es altamente extensible, se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS.

2.2.10. R-Studio

R-Studio es un entorno de desarrollado integrado (IDE) el cual permite utilizar las funcionalidades de R de manera gráfica y estructurada. Loo y Jonge (2012) mencionan que RStudio integra el entorno R, un editor de texto muy avanzado, el sistema de ayuda de R, el control de versiones y mucho más en una sola aplicación. RStudio no realiza ninguna operación estadística; solo hace que sea más fácil para usted realizar tales operaciones con R. Lo más importante es que RStudio ofrece muchas funciones que facilitan mucho el trabajo reproducible.

2.2.11. Anonimización de datos

Es una técnica utilizada para proteger la privacidad del individuo quien es dueño de los datos que están siendo utilizados, en este sentido la Agencia Española Protección de Datos en su Guía Básica de Anonimización publicada en octubre de 2022 menciona:

> La anonimización consiste en la conversión de datos personales en datos que no se pueden utilizar para identificar a ningún individuo. La anonimización hay que considerarla como un proceso basado en el riesgo, que incluye tanto la aplicación de técnicas de anonimización como salvaguardas para evitar la reidentificación. (AEPD, 2022, p.6).

3. Metodología

Para cumplir con el objetivo del trabajo de titulación se elige la metodología CRISP-DM la cual incluye descripciones de las fases normales de un proyecto y además un ciclo de vida para minería de datos. El ciclo de vida del modelo CRISP-DM contiene seis fases, estas fases tienen dependencias entre sí, cabe mencionar que no existe una secuencia estricta entre cada una de ellas y en algunos casos se puede avanzar y retroceder entre fases según la necesidad.

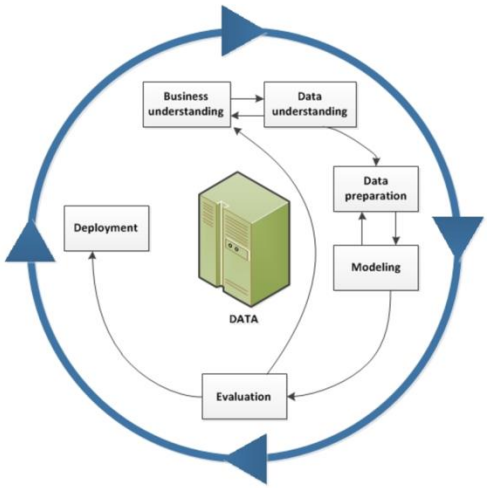


Figura 1: Ciclo de vida de minería de datos
Fuente: IBM (2021)

3.1. Entendimiento del Negocio

Hasta febrero de 2023, según los datos publicados en la página web de la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL), se puede observar que las empresas de telecomunicaciones de servicio móvil que proveen servicios son Telefónica 1, Telefónica 2 y Telefónica 3. Además, estas empresas han presentado un comportamiento histórico que se detalla a continuación:

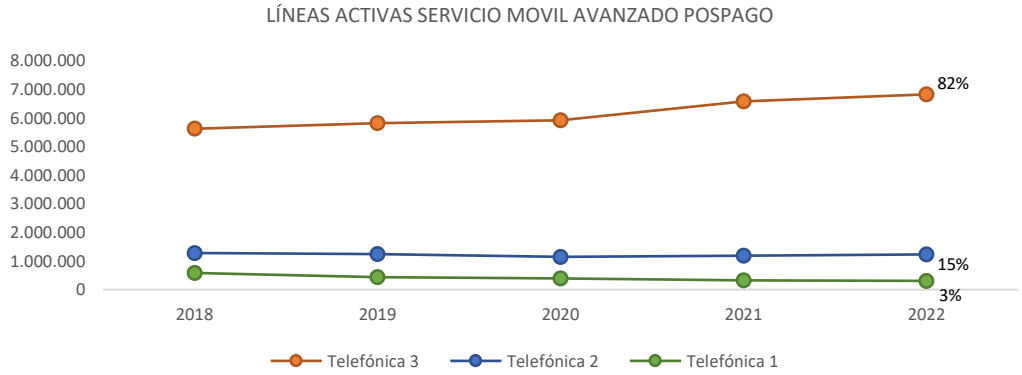


Figura 2: Líneas activas Servicio Móvil Avanzado Ecuador

En la Figura 2, se puede observar que la empresa Telefónica 3 tiene una mayor participación en el mercado de servicios móviles avanzados en Ecuador, con un porcentaje del 82%, seguida de Telefónica 2 con una participación media del mercado de 15%, mientras que Telefónica 1 tiene una participación menor en comparación con las otras dos empresas, con un porcentaje de 3%. Es interesante destacar que tanto la

Telefónica 2 como Telefónica 3 han mantenido su participación en el mercado desde el año 2018. Sin embargo, la Telefónica 1, además de tener una participación más baja, ha experimentado una disminución constante en su porcentaje de participación desde ese mismo año.

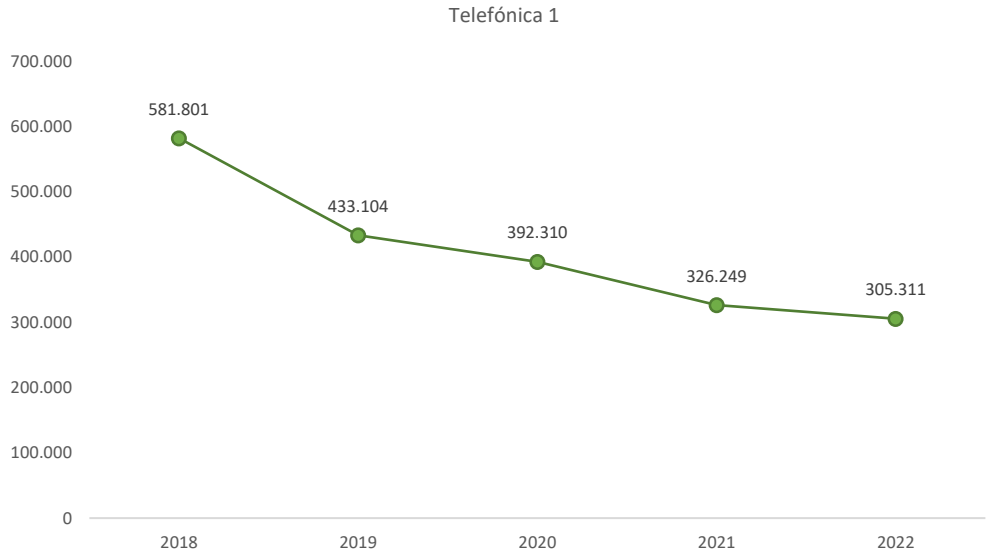


Figura 3. Líneas activas Servicio Móvil Avanzado Telefónica 1

En la figura 3 se observa que el comportamiento del parque de clientes de la Telefónica 1 muestra la necesidad de plantear una solución efectiva para disminuir la tasa de deserción y captar nuevos clientes, a fin de aumentar su participación en el mercado de servicios móviles avanzados. Por ello, es fundamental desarrollar un modelo de segmentación de clientes en base a características comunes de comportamiento y generar un modelo de predicción de deserción, lo que permitirá establecer estrategias de fidelización y retención preventiva.

De esta manera, se podrán implementar estas estrategias a reducir la tasa de deserción de clientes en el producto de servicios móviles avanzados.

3.2. Comprensión de los Datos

La fase de comprensión de datos de CRISP-DM implica estudiar más de cerca los datos disponibles de minería. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto. (IBM, 2021).

El objetivo en esta fase es obtener una comprensión inicial del conjunto de datos con el que se va a trabajar, para lo cual se realizan actividades como: recolección de datos relevantes, exploración de datos para identificar patrones, tendencia, así como, identificación de posibles relaciones entre variables mediante el uso de técnicas estadísticas y visuales.

3.2.1. Descripción de los Datos

El data set proviene de una base de datos estructurada que está compuesta por 20 variables y 304.572 registros, estas variables contienen información respecto al cliente, al producto

y al comportamiento de consumo del servicio contratado. Adicional, hay que destacar que una de las variables determina si el cliente ha cancelado su servicio o si aún lo mantiene activo.

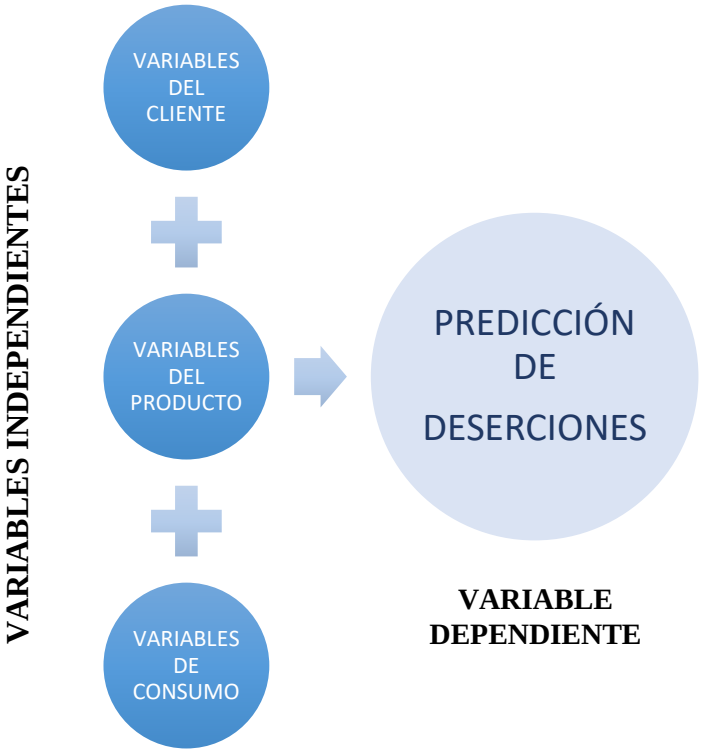


Figura 4. Estructura de Datos

Por otra parte, hay que mencionar que las variables que contienen información del cliente, producto y comportamiento del uso del servicio están conformadas por diferentes tipos de variables, a continuación, un detalle de los tipos de variables:

Tabla 1: Detalle del tipo de variables que conforman el data set

GRUPO DE VARIABLES	VARIABLE	TIPO DE VARIABE
PRODUCTO	Número de Teléfono	Numérica Discreta
	Permanencia del Servicio	Categórica Ordinal
	Estado del Servicio	Cualitativa Dicotómica
CLIENTE	Edad	Numérica Discreta
CONSUMO	Consumo de Facebook	Numérica Continua
	Consumo de WhatsApp	Numérica Continua
	Consumo de Twitter	Numérica Continua
	Consumo de YouTube	Numérica Continua
	Consumo de Deezer	Numérica Continua
	Consumo SnapChat	Numérica Continua
	Consumo de Spotify	Numérica Continua
	Consumo de Instagram	Numérica Continua

Fuente: Elaboración propia obtenida del data set

3.2.2. Exploración de los Datos

Durante esta fase, se explora los datos con la ayuda de tablas o gráficos que permiten visualizar propiedades de los datos, se realiza cálculos para identificar estadísticas descriptivas, análisis de relación entre las variables, análisis de distribución de las frecuencias e identificación de valores atípicos.

La Documentación sobre la metodología CRISP-DM de IMB menciona que: "Estos análisis pueden ayudarle a describir los objetivos de minería de datos generados durante la fase de comprensión comercial. También pueden ayudarle a formular hipótesis y dar forma a las tareas de transformación de datos que tienen lugar durante la preparación de los datos." (IBM, 2021).

a. Visualización de datos

Se realiza la exploración del data set con el objetivo de entender los patrones en los datos y potencialmente plantear una hipótesis.

i. Proporción de Servicios Activos vs Servicios Retirados

En la Figura 5, se puede observar que el porcentaje de servicios activos versus el porcentaje de servicios retirados no presenta un desbalance de la información, de tal forma que al construir los modelos de clasificación y predicción de deserción de clientes no presentará sesgos hacia la clase mayoritaria o dificultades para identificar correctamente las muestras de la clase minoritaria.

Figura 5. Porcentaje de Servicios Activos y Retirados

Raschka y Mirjalili (2019) afirman que, el desequilibrio de clases influye en un algoritmo de aprendizaje durante el ajuste del modelo. Dado que los algoritmos de aprendizaje automático generalmente optimizan una función de recompensa o costo que se calcula como una suma de los ejemplos de entrenamiento que ve durante el ajuste, es probable que la regla de decisión esté sesgada hacia la clase mayoritaria. En otras palabras, el algoritmo aprende implícitamente un modelo que optimiza las predicciones en función de la clase más abundante en el conjunto de datos, para minimizar el costo o maximizar la recompensa durante el entrenamiento.

ii. Cantidad de Servicios Activos y Servicios Retirados por Género

La Figura 6 muestra que para el género masculino se tiene una proporción del 38% de servicios activos y del 62% de servicios retirados, mientras que para el género femenino la proporción es del 55% de servicios activos y del 55% de servicios retirados. Estos datos sugieren que las mujeres tienen una tasa ligeramente mayor de servicios activos que los hombres, y que la tasa de servicios retirados es más alta para el género masculino al comparar con el género femenino.

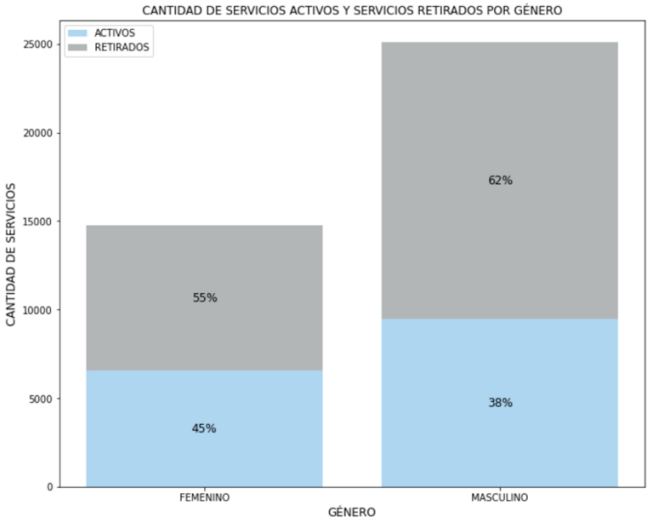


Figura 6. Cantidad de Servicios por Género

iii. Concentración de servicios activos por permanencia

La Figura 7 muestra que, se tiene 4 grupos de clientes activos en el tiempo, el primer grupo corresponde a clientes nuevos con una antigüedad entre 0 y 18 meses, el segundo grupo se ubica entre los 19 y 42 meses, el grupo con mayor cantidad de servicios activos se ubica entre los 43 y 102 meses, el cuarto grupo se ubica entre los 103 y 126 meses y finalmente el grupo más antiguo con una cantidad mínima de servicios activos superior a los 127 meses de permanencia.

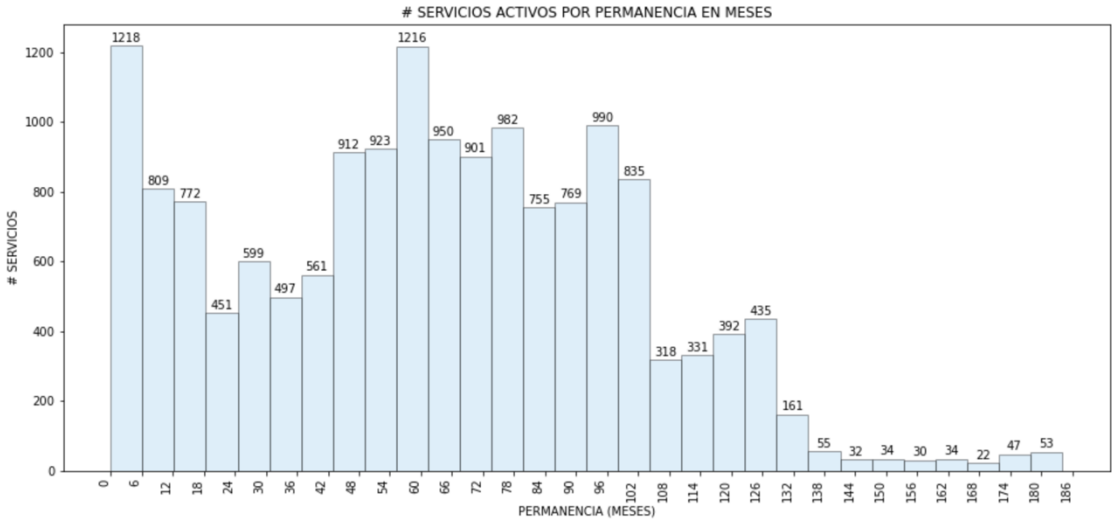


Figura 7. Cantidad de servicios activos por meses de Permanencia

Por otro lado, respecto a la cantidad de servicios retirados, en la Figura 8 se puede apreciar que, entre los 0 y 18 meses de permanencia existe una gran cantidad de servicios retirados, lo cual muestra que el esfuerzo realizado para la obtención de nuevos clientes no perdura en el tiempo, otro grupo se ubica entre los 19 y 60 meses y finalmente el último grupo con una cantidad menor de servicios retirados que sobrepasaron los 61 meses.

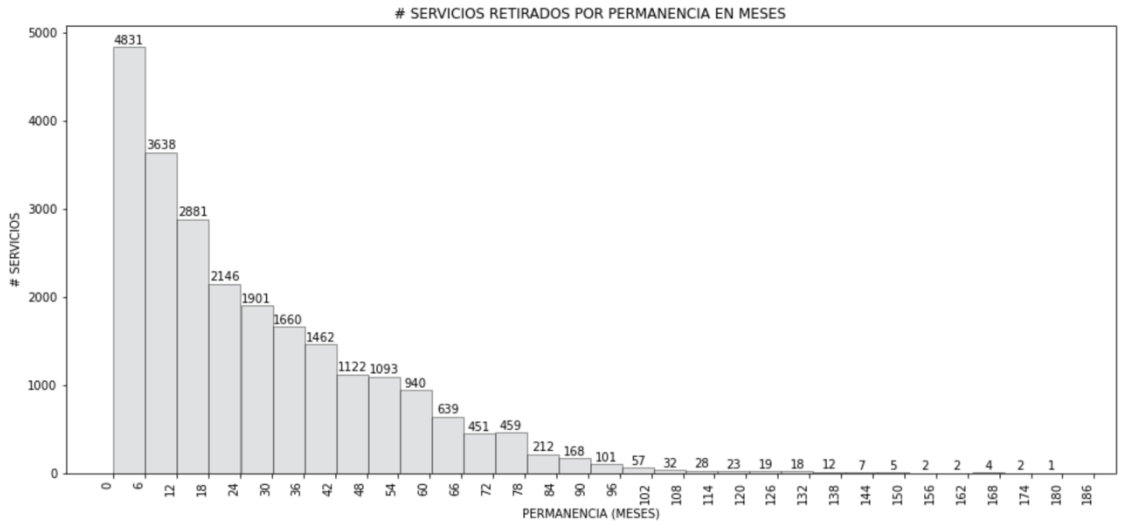


Figura 8. Cantidad de servicios retirados por meses de Permanencia

iv. Concentración de servicios activos por edad del cliente

En la Figura 9, se puede evidenciar que el rango de edad con una gran cantidad de servicios activos se encuentre entre los 30 y 45 años, un segundo grupo de servicios activos se encuentre entre los 46 y 69 años, por otro lado, hay otros dos grupos con una cantidad mínima de servicios activos que se encuentran en el rango entre 18 y 29 años y entre 70 100 años.

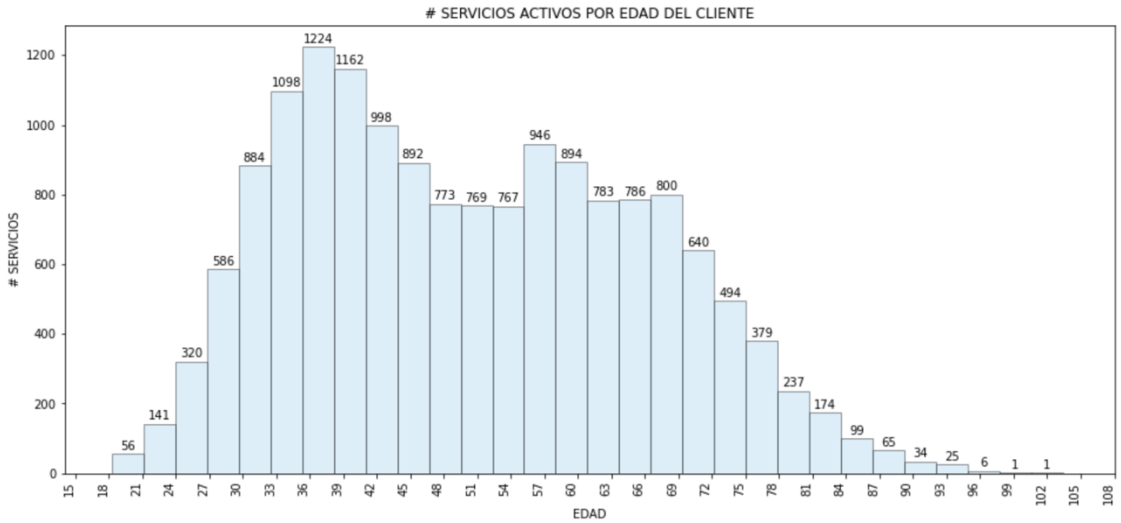


Figura 9. Cantidad de servicios activos por edad

Por otro lado, la Figura 10 muestra que, la mayor cantidad de servicios retirados se concentran en edades entre los 27 y 45 años y un segundo grupo entre los 46 y 60 años.

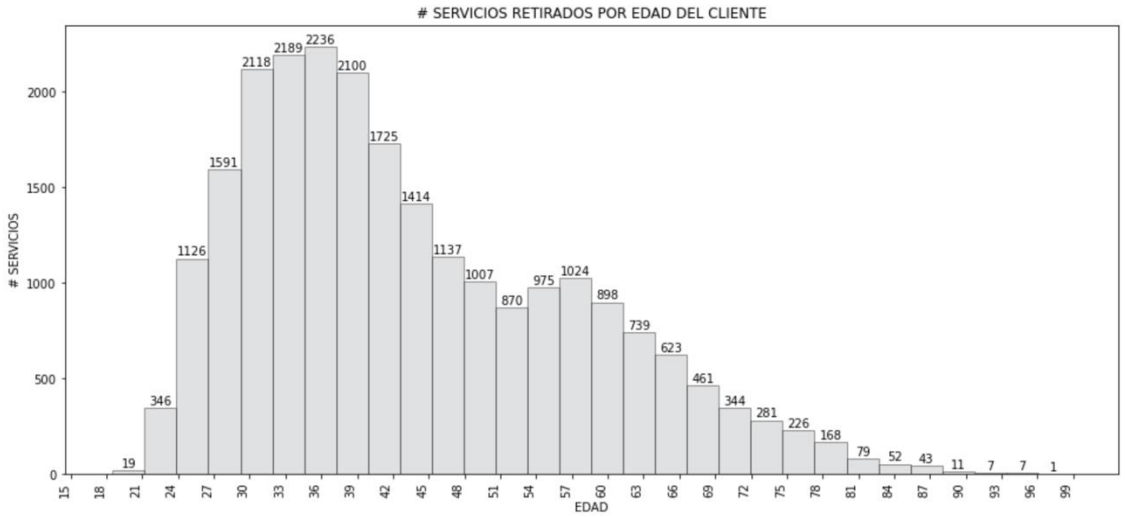


Figura 10. Cantidad de servicios retirados por edad

b. Análisis de correlación entre variables

En la Figura 11, se observa que la variable dependiente presenta:

- Una correlación negativa media con una de las variables independientes.
- Cinco correlaciones negativas débiles con sus variables independientes.

Por otro lado, entre sus variables independientes existe correlaciones negativas débiles lo cual no implica que existe multicolinealidad que pueda dificultar el análisis al implementar el algoritmo de regresión.

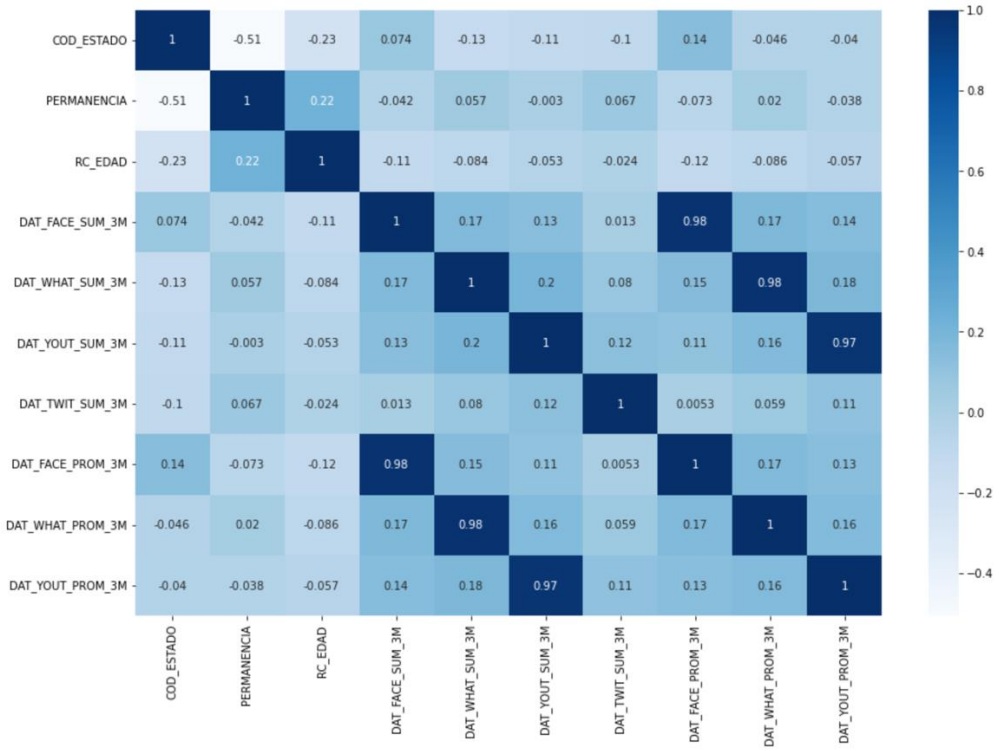


Figura 11. Correlación de variables

3.2.3. Preparación de Datos

La etapa de preparación de datos es una de las más importantes y laboriosas ya que en esta fase se llevan a cabo una serie de tareas esenciales para garantizar que los datos estén listos para el análisis. (IBM, 2021).

Chapman et al. (2000) mencionan que, las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

3.2.3.1. Limpieza de Datos

Durante esta etapa se realiza tareas que permite identificar y eliminar valores duplicados, valores faltantes y valores atípicos.

i. Identificación y eliminación de valores duplicados

En la Tabla 2 se puede evidenciar que el porcentaje de valores duplicados encontrados corresponde al 4% respecto al total de datos.

Tabla 2. Porcentaje de valores duplicados

	VALORES	**%**
ÚNICOS	38.537	96%
DUPLICADOS	1.463	4%
	40.000	**100%**

Fuente. Elaboración propia obtenida del data set

Considerando que el porcentaje de valores duplicados alcanza el 4% versus el total del data set, se procede a eliminar aquellos valores duplicados con el fin de evitar que al implementar K-Means el algoritmo asigne incorrectamente los puntos a los clústeres generando una mala agrupación, o afecte a la precisión al implementar el algoritmo de Regresión Logística.

ii. Identificación y eliminación de valores faltantes

Se identifica que el data set posee un 0.20% de valores faltantes en el grupo de variables independientes del cliente, a continuación, el detalle.

Tabla 3. Cantidad y porcentaje de valores faltantes o nulos

VARIABLES	**VALORES NULOS**	**%**
Género	148	0,38%
Edad	148	0,38%

Fuente. Elaboración propia obtenida del data set

Considerando que los valores faltantes corresponden a las variables del cliente como Edad y Género, se plantea el reemplazo de valores faltantes aplicando la media para la primera variable y la moda de los datos para la segunda y tercera variable, las cuales son técnicas simples, rápidas y útiles para el reemplazo de valores faltantes.

iii. Identificación y eliminación de valores atípicos

En la Tabla 4 se evidencia que, las variables independientes poseen valores atípicos, lo cual puede generar problemas al implementar K-Means debido a que los valores atípicos pueden ser asignados a grupos incorrectos y afectar la calidad de la segmentación, mientras que, al implementar Regresión Logística, los outliers pueden influir en los coeficientes de la regresión y afectar la precisión del modelo. Además, se evidencia que seis de las variables independientes tienen más del 10% de valores atípicos, y siete variables independientes poseen menos del 10% de valores atípicos versus el total de los datos.

Tabla 4: Porcentaje de valores atípicos

VARIABLES	% VALORES ATÍPICOS
DAT_TWIT_SUM_3M	14,87%
DAT_FACE_PROM_3M	12,38%
DAT_FACE_SUM_3M	12,25%
DAT_WHAT_SUM_3M	10,39%
DAT_YOUT_PROM_3M	10,31%
DAT_WHAT_PROM_3M	10,05%
DAT_YOUT_SUM_3M	9,88%
NUM_SERVICIO	4,24%
PERMANENCIA	0,67%
RC_EDAD	0,08%
COD_ESTADO	0,00%
RC_DES_SEXO	0,00%
COD_PRODUCTO	0,00%

Fuente: Elaboración propia obtenida del data set

Considerando esto se plantea el uso de la técnica del rango intercuartílico para eliminar valores atípicos del conjunto de datos, para esto se encuentra la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Los valores que se encuentran por debajo de Q1 - 1.5 veces el rango intercuartílico o por encima de Q3 + 1.5 veces el rango intercuartílico se consideran atípicos y son eliminados del conjunto de datos. Hay que mencionar que esta técnica es útil para detectar valores extremos que podrían sesgar los resultados del análisis y afectar la interpretación de los resultados. Como resultado, al eliminar los valores atípicos, se mejora la calidad de los datos y se obtienen resultados más confiables y representativos en el análisis estadístico.

iv. Anonimización de datos

La Guía básica de anonimización elaborada por la autoridad de Protección de Datos de Singapur (2022), acerca de la anonimización menciona que, consiste en la conversión de datos personales en datos que no se pueden utilizar para identificar a ningún individuo. La anonimización hay que considerarla como un proceso basado en el riesgo, que incluye tanto la aplicación de técnicas de anonimización como salvaguardas para evitar la reidentificación.

Considerando que el data set en análisis contiene datos respecto al servicio como su número identificador, es necesario aplicar una técnica de anonimización que permita

resguardar esta información, de tal forma que, se hace uso de la herramienta Fernet la cual fue desarrollada por la Autoridad Criptográfica de Python que permite cifrar datos y que no pueden ser manipulados o leídos sin la clave. El proceso de cifrado con Fernet consiste en:

1. Importar la librería Fernet

```
from cryptography.fernet import Fernet
```

Figura 11. Importación de librería Fernet

2. Generar la clave para el cifrado

```
key_fernet_tt = Fernet.generate_key()
```

Figura 12. Generar clave de cifrado

3. Crear un objeto el cual permite cifrar y descifrar los datos

```
obj_Fernet = Fernet(key_fernet_tt)
```

Figura 13. Creación de objeto para cifrado y descifrado

4. Cifrado de datos

```
df_datos['NUM_SERVICIO_CIFRADO'] = df_datos['NUM_SERVICIO'].apply(lambda x: obj_Fernet.encrypt(str(x).encode()))
```

Figura 14. Proceso de cifrado de datos

5. Datos cifrados

	NUM_SERVICIO_CIFRADO
0	b'gAAAAABkVbbJTxoFtCSJs0ULQ3jtj4YucXFSTMrTSnfa...
1	b'gAAAAABkVbbJcQf0GSWCK3QolGw2b73kpANwxKxAVhC9...
2	b'gAAAAABkVbbJATUVDp6gWPZqt0FBVg7EYNkXG1AHeTlj...
3	b'gAAAAABkVbbJ43h3kPb531T2F3V7jWMuGLWRzG2rhC_v...
4	b'gAAAAABkVbbJluODFSQgvXCdgW5iRFFSxxltlQVixInf...

Figura 15. Datos cifrados

v. Selección de variables

En la implementación de los algoritmos K-Means y Regresión Logística, se seleccionaron cuidadosamente las variables relevantes relacionadas con el cliente, el producto y el comportamiento del cliente con el servicio. Estas variables se dividieron en dos grupos: aquellos clientes que han desertado previamente y aquellos que mantienen activos sus servicios. Durante el proceso de selección de variables, se adoptó un enfoque combinado que incorporó tanto métodos objetivos como el conocimiento experto en el dominio. Reconociendo la importancia de aprovechar la experiencia de los expertos, se llevó a cabo una colaboración en la cual se recopiló la retroalimentación y las ideas de los profesionales con un profundo conocimiento del negocio. Su comprensión detallada y su experiencia específica ayudaron a identificar los factores que pueden influir en la deserción de clientes de manera significativa. Esta combinación de enfoques condujo a una selección de variables sólida y fundamentada, mejorando así la capacidad de los

modelos para predecir y comprender la deserción de clientes con mayor precisión y efectividad. Para mayor referencia, la Tabla 5 muestra la estructura del conjunto de datos final que se utilizará en la etapa de modelado.

Tabla 5: Selección de variables

GRUPO DE VARIABLES	VARIABLE	TIPO
PRODUCTO	ESTADO	DEPENDIENTE
	PERMANENCIA	INDEPENDIENTE
	DATOS_FACEBOOK_3_MESES	INDEPENDIENTE
	DATOS_WHATSAPP_3_MESES	INDEPENDIENTE
	DATOS_YOUTUBE_3_MESES	INDEPENDIENTE
	DAT_TWITTER_3_MESES	INDEPENDIENTE
	DAT_FACEBOOK_PROMO_3_MESES	INDEPENDIENTE
	DAT_WHATSAPP_PROMO_3_MESES	INDEPENDIENTE
	DAT_YOUTUBE_PROMO_3_MESES	INDEPENDIENTE
CLIENTE	EDAD	INDEPENDIENTE

Fuente: Elaboración propia obtenida del data set

vi. Normalización de variables

Previo el paso a la etapa de modelado, es importante normalizar los datos de las variables que se utilizarán. Esto debido a que las variables tienen diferentes escalas, lo que puede afectar la precisión y el rendimiento del modelo. A continuación, en la Figura 16 se tienen el resumen estadístico de las variables que se utilizarán en el análisis en donde se evidencia la variación de escalas entre ellas.

	ESTADO	COSTO	PERMANENCIA	EDAD	DATOS_WHATSAPP_3_MESES	DATOS_YOUTUBE_3_MESES	DATOS_SPOTIFY_3_MESES
count	390875.000000	390875.000000	390875.000000	390875.000000	390875.000000	390875.000000	390875.000000
mean	1.919051	7.690230	47.120571	44.863151	1.847389	0.924055	0.120371
std	0.996720	9.218144	39.035386	13.919926	3.911047	3.940890	0.770622
min	1.000000	0.000000	0.000000	18.080000	0.000000	0.000000	0.000000
25%	1.000000	0.000000	11.000000	34.210000	0.000000	0.000000	0.000000
50%	1.000000	0.000000	38.000000	41.870000	0.330000	0.180000	0.000000
75%	3.000000	14.530000	76.000000	53.820000	2.140000	0.850000	0.010000
max	3.000000	99.990000	229.000000	105.040000	437.810000	418.110000	111.950000

Figura 16. Identificación de escalas diferentes en las variables

Por consiguiente, se realiza la normalización de las variables con el método de normalización min-max, el cual escala los datos de tal forma que el valor mínimo se transforma a 0 y el valor máximo se transforma en 1, mientras que los valores intermedios se escalan de forma proporcional entre 0 y 1.

3.3. Modelado

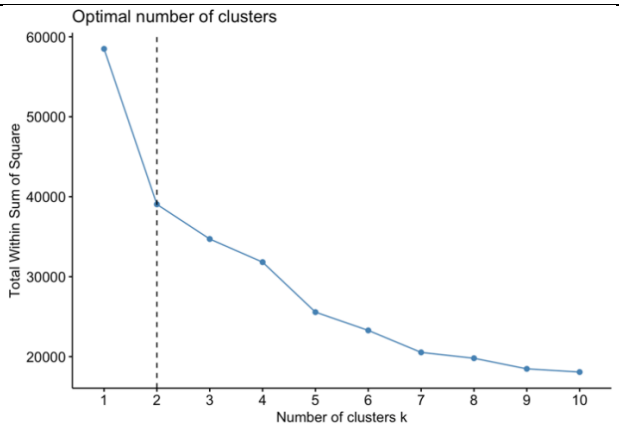
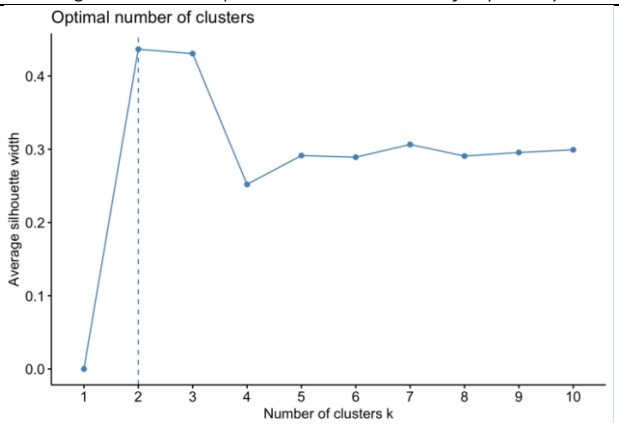
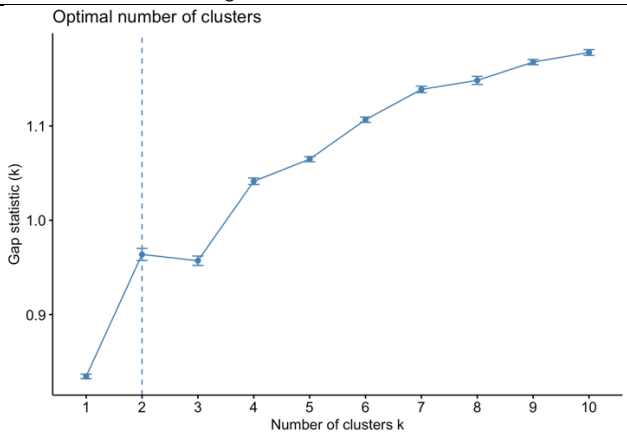
3.3.1. K-Means

(Scikit - Learn) en su capítulo de Clustering menciona que: "el algoritmo K-Means agrupa los datos tratando de separar muestras en "k" grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo." De tal forma que, esta técnica al ser del tipo no supervisada permite agrupar conjuntos de datos no

etiquetados en "k" grupos, descubriendo patrones o estructuras intrínsecas en los datos, basándose en la similitud de características comunes entre ellos.

En tal sentido, una de las acciones indispensables para la implementación del algoritmo K-Means, es conocer la cantidad óptima de clústeres que agruparán a los datos bajo análisis. Para esto, se plantea el uso de los métodos WSS, Silueta y Gap Stat, mediante los cuales se puede identificar la cantidad idónea de clústeres a ser usados en la implementación de K-Means. En la Tabla 5, se evidencia el número de clústeres que cada uno de los métodos sugiere como óptimos.

Tabla 6. Número idóneos de clústeres

MÉTODO	GRÁFICA	CLÚSTERES
WSS	 *Figura 17. WSS (Within-Cluster Sum of Squares)*	2
Silhouette	 *Figura 18. Silhouette*	2
Gap Stat	 *Figura 20. Gap Stat*	2

Fuente: Elaboración propia obtenida del data set

La identificación del número idóneo de clústeres con los métodos WSS, Silueta y Gap Stat, sugieren “2” clústeres. Tomando en cuenta que, los tres métodos sugieren “2” clústeres, se implementa K-Means considerando los “2” clústeres sugeridos. Una vez que se implementa K-Means con los clústeres sugeridos, se obtiene la proporción entre el porcentaje de la suma de cuadrados entre clústeres (between_SS) versus la relación con la suma de cuadrados total (total_SS), la cual revela que el modelo tiene la capacidad de explicar el 33.2% de la variabilidad total de los datos en la formación de clústeres.

```
Within cluster sum of squares by cluster:
[1] 14917.79 24126.06
 (between_SS / total_SS =  33.2 %)
```

Figura 21. Proporción de la variabilidad K-Means

Por otro lado, al observar la Figura 22 se evidencia que al graficar los “2” clústeres los datos están agrupados de manera distintiva y claramente diferenciada dentro de cada clúster definido.

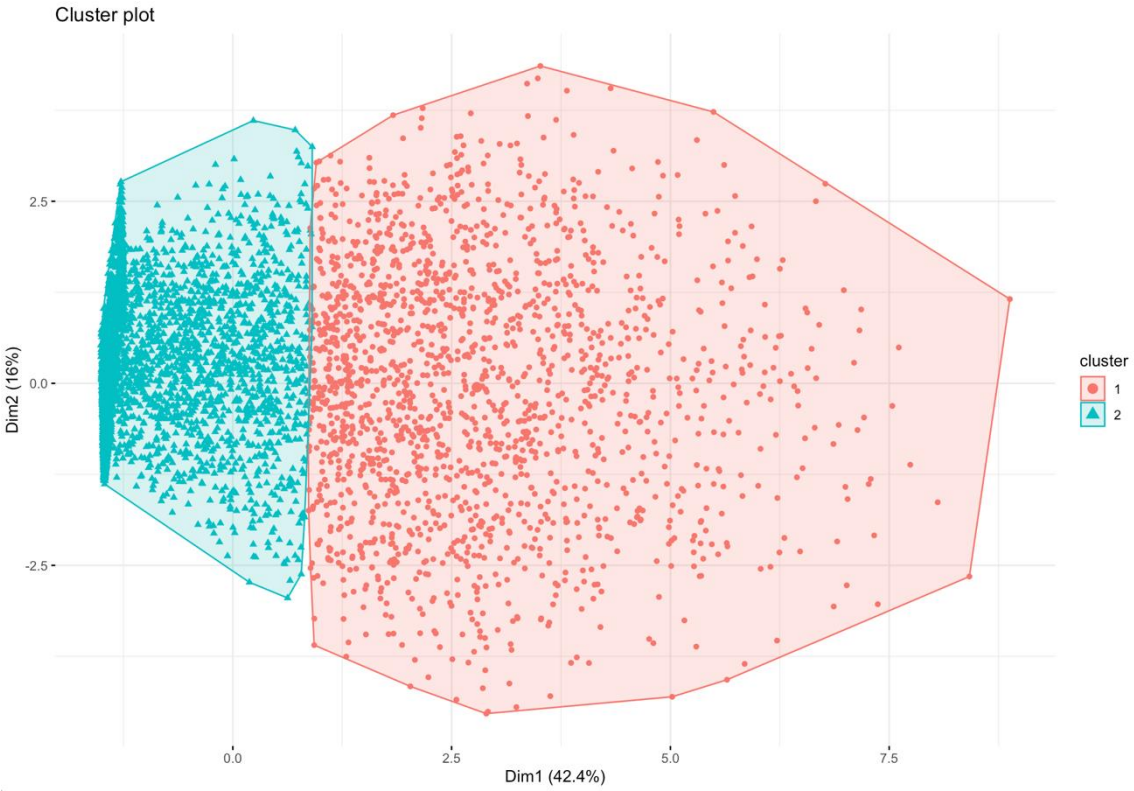


Figura 22. Clusterización de Datos

3.3.1.1. Evaluación y Resultados K-Means

El resultado final de la agrupación de datos mediante K-Means, permite conocer de forma específica como se agrupan los datos en base a un comportamiento en particular de cada grupo formado, para entender un poco más acerca de esto, a continuación, en la Tabla 7 y Figura 22, se representan las medias de cada una de las variables que conforman los clústeres.

Tabla 7. Medias por cada Clúster

VARIABLES	CLÚSTERES	
	CLÚSTER 1	CLÚSTER 2
PERMANENCIA	0,14	-0,05
EDAD	-0,06	0.02
DATOS_FACEBOOK_3_MESES	1,09	-0,41
DATOS_WHATSAPP_3_MESES	1,08	-0,41
DATOS_YOUTUBE_3_MESES	1,15	-0,43
DAT_TWITTER_3_MESES	0,62	-0,23
DAT_FACEBOOK_PROMO_3_MESES	1,05	-0,39
DAT_WHATSAPP_PROMO_3_MESES	1,12	-0,42
DAT_YOUTUBE_PROMO_3_MESES	1,13	-0,42

Fuente: Elaboración propia obtenida del data set

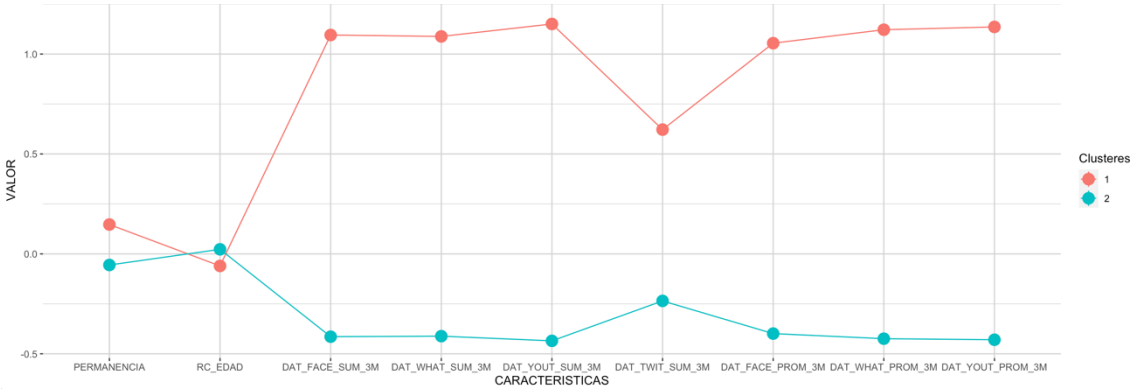


Figura 22. Análisis de Clústeres según sus variables

Al analizar la representación de las medias y variables que conforman los clústeres, se identifica lo siguiente:

Clúster 1

- La media de Permanencia es mayor respecto al Clúster 2.
- La media de Edad es menor respecto al Clúster 2.
- La media de Consumo de Datos Móviles en los últimos tres meses en Aplicaciones es mayor respecto al Clúster 2.
- La media de Consumo de Datos Móviles en los últimos tres meses en Twitter es mayor respecto al promedio de las otras aplicaciones como Facebook, WhatsApp y YouTube en el mismo Clúster.
- La media de Consumo de Datos Móviles Promocional en los últimos tres meses en Aplicaciones es mayor respecto al Clúster 2.

Clúster 2

- La media de Permanencia es menor respecto el Clúster 1.
- La media de Edad es mayor respecto el Clúster 1.
- La media de Consumo de Datos Móviles en los últimos tres meses en Aplicaciones es menor respecto al Clúster 1.
- La media de Consumo de Datos Móviles en los últimos tres meses en Twitter es menor respecto al promedio de las otras aplicaciones como Facebook, WhatsApp y YouTube en el mismo Clúster.
- La media de Consumo de Datos Móviles Promocional en los últimos tres meses en aplicaciones es menor respecto al Clúster 1.

La segmentación de clientes generada por el algoritmo K-Means brinda la oportunidad de implementar estrategias que aborden las necesidades actuales de cada grupo en específico, desarrollando productos o servicios adaptados a sus requerimientos particulares, lo cual permitirá incrementar la retención de los clientes y fomentar su permanencia a largo plazo.

3.3.2. Regresión Logística

Brage (2020) sobre la Regresión Logística afirma que: "Se trata de una de las técnicas más conocidas y utilizadas para modelar una variable respuesta dicotómica en función de un conjunto de variables predictoras, que pueden ser continuas o categóricas." (p 1).

Por otro parte, respecto a las ventajas que brinda el uso de la técnica de regresión logística se puede mencionar que, es simple de implementar, ofrece eficiencia computacional desde el punto de vista de entrenamiento, no requiere que los datos sean escalados y comúnmente es usado en aplicaciones empresariales (Ray, 2019).

Para el caso particular del análisis de deserción de clientes, la técnica de regresión logística permite predecir considerando sus variables predictoras si un cliente tiene la probabilidad de retirar su servicio o no, en este caso la variable dependiente representa al servicio cuando sigue activo con el valor de "1" y "0" cuando el servicio fue cancelado.

Previo la implementación del modelo de regresión logística, se divide al conjunto de datos en dos partes, un porcentaje es asignado para el conjunto de entrenamiento el cual se utiliza para ajustar el modelo, y otro asignado para el conjunto de pruebas que se utiliza para evaluar la capacidad de generalización el modelo. En la Figura 23 y Figura 24, se aprecia como quedan estructurados tanto el conjunto de entrenamiento como el conjunto de prueba.

```
'data.frame':   26975 obs. of  6 variables:
 $ ESTADO                : int  1 1 1 1 0 0 (
 $ PERMANENCIA           : int  70 2 51 124 4
 $ EDAD                  : num  59.3 71 78.3
 $ DATOS_WHATSAPP_3_MESES: num  0.26 0 0.01 (
 $ DATOS_YOUTUBE_3_MESES : num  1.02 0.09 0.1
 $ DATOS_SPOTIFY_3_MESES : num  1.54 0 0.01 (
```

Figura 23. Conjunto de entrenamiento

```
'data.frame':   11562 obs. of  6 variables:
 $ ESTADO                : int  0 1 1 0 1 1
 $ PERMANENCIA           : int  14 91 95 38
 $ EDAD                  : num  87.4 76 87.1
 $ DATOS_WHATSAPP_3_MESES: num  0 0.16 0.24
 $ DATOS_YOUTUBE_3_MESES : num  0 0.05 0.37
 $ DATOS_SPOTIFY_3_MESES : num  0 0 0 0 0.17
```

Figura 24. Conjunto de prueba

La división del conjunto de datos se realiza considerando una práctica común, donde se asigna el 70% de los datos al conjunto de entrenamiento y el 30% al conjunto de pruebas.

De esta manera, se garantiza una cantidad suficiente de datos para entrenar el modelo y también para evaluar su rendimiento. Una vez dividido el conjunto de datos, se procede con el ajuste del modelo utilizando el conjunto de entrenamiento. Para este propósito, se emplea la función "glm" de RStudio, la cual se aplica considerando la distribución de tipo binomial, de la siguiente manera.

```
modelo_regresion <- glm(COD_ESTADO ~ ., data = data_entrenamiento, family = "binomial")
```

Figura 25. Ajuste del Modelo de Regresión Logística – Binomial

Posterior a realizar el ajuste del modelo, se obtiene información relevante sobre los coeficientes estimados utilizando la función "summary(modelo_regresion)", a continuación, en la Figura 26 se muestra el resultado.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.183e+00  8.138e-02 -51.409  < 2e-16 ***
PERMANENCIA      2.961e-02  6.451e-04  45.907  < 2e-16 ***
RC_EDAD          3.699e-02  1.373e-03  26.933  < 2e-16 ***
DAT_FACE_SUM_3M  2.805e+00  2.166e-01  12.952  < 2e-16 ***
DAT_WHAT_SUM_3M  9.367e+00  3.592e-01  26.080  < 2e-16 ***
DAT_YOUT_SUM_3M  2.805e+00  2.849e-01   9.846  < 2e-16 ***
DAT_TWIT_SUM_3M  4.329e-01  1.314e-01   3.293  0.00099 ***
DAT_FACE_PROM_3M -6.454e+00  4.492e-01 -14.367  < 2e-16 ***
DAT_WHAT_PROM_3M -1.911e+01  7.264e-01 -26.313  < 2e-16 ***
DAT_YOUT_PROM_3M -5.554e+00  5.713e-01  -9.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 26. Coeficientes Modelo de Regresión Logística

El resultado muestra que las variables predictoras tienen valores $p < 0.05$, lo que indica una alta significancia estadística. Por consiguiente, a estos coeficientes estimados se los considera significativos y sugieren que tienen una relación estadísticamente significativa con la variable objetivo.

3.3.2.1. Evaluación Regresión Logística

Para la evaluación del modelo de regresión logística, se considera el uso de los siguientes indicadores.

a. Área bajo la curva (AUC) – ROC (Receiver Operating Characteristic)

El análisis de Churn en este estudio ha revelado un valor significativo del Área bajo la Curva ROC (AUC) del 95%. El AUC es una medida numérica que cuantifica la habilidad de un modelo de clasificación binaria para distinguir entre clientes que se mantienen y aquellos que abandonan un servicio. Un AUC de 95% indica un rendimiento sólido y una capacidad sustancial para realizar una clasificación precisa. Cuanto más cercano esté el AUC a 1, mayor será la capacidad del modelo para clasificar correctamente a los clientes. Este resultado destaca la alta probabilidad de que el modelo utilizado en este estudio clasifique correctamente a un cliente seleccionado aleatoriamente como retención o Churn. Este hallazgo es de gran relevancia, ya que brinda una base sólida para la toma de decisiones y la implementación de estrategias efectivas de retención de clientes. En la

Figura 27, se representa la relación existente entre la tasa de verdaderos positivos y la tasa de falsos positivos a medida que se ajusta el umbral de clasificación del modelo.

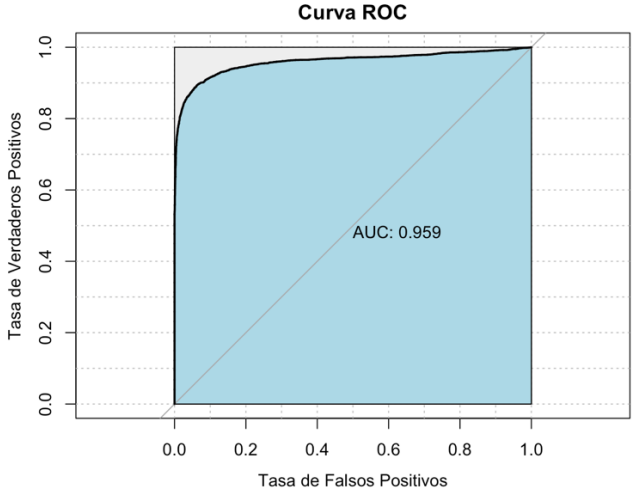


Figura 27. Curva ROC

b. Precisión - Exactitud

Como parte de la evaluación del modelo de regresión logística implementado, se obtiene el valor de la precisión y exactitud que para este modelo es del 92% para los dos indicadores, esto indica que es capaz de clasificar correctamente el 92% de los casos positivos y negativos de Churn en el conjunto de datos de prueba. Esto implica que el modelo tiene una capacidad razonable para distinguir entre los clientes que se mantienen y los que abandonan el servicio. Sin embargo, es importante tener en cuenta que existen margen para mejoras y optimizaciones en el modelo, ya que hay un 23% de casos que no se clasificaron correctamente.

c. Matriz de confusión

El análisis de la matriz de confusión del modelo de regresión logística aplicado en este análisis muestra información importante respecto al rendimiento en la clasificación binaria del modelo, a continuación, en la Tabla 8 se muestra los resultados.

Tabla 8. Matriz de Confusión

	Clase 0 (Predicción)	Clase 1 (Predicción)
Clase 0 (Real)	6.706	199
Clase 1 (Real)	711	3.946

Fuente: Elaboración propia obtenida del data set

La interpretación de los resultados de la Tabla 8, es la siguiente:

- En la fila "0" y la columna "0" se encuentran los casos en los que la predicción del modelo es "0" y el valor real también es "0". En este caso, hay 6.706 observaciones clasificadas correctamente como "0" (verdaderos negativos).
- En la fila "0" y la columna "1" se encuentran los casos en los que la predicción del modelo es "1", pero el valor real es "0". Aquí, hay 199 observaciones clasificadas incorrectamente como "1" cuando en realidad son "0" (falsos positivos).

- En la fila "1" y la columna "0" se encuentran los casos en los que la predicción del modelo es "0", pero el valor real es "1". Hay 711 observaciones clasificadas incorrectamente como "0" cuando en realidad son "1" (falsos negativos).
- En la fila "1" y la columna "1" se encuentran los casos en los que la predicción del modelo es "1" y el valor real también es "1". Aquí, hay 3.946 observaciones clasificadas correctamente como "1" (verdaderos positivos).

Se observa que existe bu un buen número de verdaderos negativos y verdaderos positivos, lo cual es deseable. Sin embargo, también hay una cantidad considerable de falsos positivos y falsos negativos, lo cual indica que el modelo tiene cierto margen de mejora.

4. Conclusiones

- El uso del algoritmo no supervisado K-Means sobre un conjunto de datos no etiquetado, ha permitido identificar la existencia de dos grupos de clientes que tienen comportamientos particulares, permitiendo comprender mejor las necesidades y preferencias de cada uno de ellos, esto otorga la ventaja para diseñar estrategias personalizadas para cada segmento, adaptando productos y servicios a características y demandas de cada uno de ellos, con la posibilidad de incrementar la satisfacción y fidelidad de estos clientes, y a la vez minimizando la cantidad de deserciones.
- K-Means al ser una técnica eficiente y escalable puede ser aplicable a grandes volúmenes de datos de clientes, lo cual permite a las organizaciones analizar de manera rápida y efectiva los datos de deserción de clientes, identificar patrones relevantes y tomar decisiones basadas en evidencia.
- La implementación de la técnica supervisada Regresión Logística proporciona coeficientes que indican la dirección y la magnitud de la influencia de cada variable en la probabilidad de deserción de clientes, por lo que las empresas podrían usar la mayor cantidad de variables predictoras significativas para que el modelo pueda mejorar su rendimiento de predicción y de esta forma tomar medidas para mitigar el impacto de deserciones.
- La implementación de la técnica supervisada de Regresión Logística brinda coeficientes que ofrecen información valiosa sobre la influencia de cada variable en la probabilidad de deserción de clientes, por consiguiente, esta particularidad permite al analista de datos incrementar variables predictoras significativas que permiten mejorar su rendimiento de predicción aprovechando al máximo las variables predictoras e implementar estrategias de retención más efectivas.

5. Recomendaciones

- El análisis de deserción de clientes no puede ser encaminado únicamente en el uso de técnicas como K-Means o Regresión Logística, por lo que es recomendable optar por el uso adicional de técnicas de machine learning como árbol de decisión, bosques aleatorios, máquina de soporte vectorial o redes neuronales.
- Implementar un proceso ETL (Extracción, Transformación y Carga) que mejore la eficiencia y la calidad de la información, así también, contar con un sistema transaccional que optimice los procesos de producción y almacenamiento de datos mejorando la calidad desde la fuente de origen.
- Debido a que la dinámica empresarial y las preferencias de los clientes pueden cambiar en el tiempo, es fundamental establecer un cronograma de evaluación periódica sobre los modelos de análisis de deserción de clientes una vez que estén en producción, de tal forma que, permita determinar si es necesario realizar ajustes sobre los modelos, como la eliminación o inclusión de variables, con el objetivo de mejorar constantemente su rendimiento.

Referencias Bibliográficas

Alegre, F. (2020). Predicción de abandono de clientes en una empresa de telecomunicaciones (Tesis de posgrado). Universidad Complutenses de Madrid, Madrid, España.

Autoridad Nacional de Protección de Datos de Singapur. (2022). *Guía Básica de Anonimización.* Recuperado de https://www.aepd.es/es/documento/guia-basica-anonimizacion.pdf

Bonaccorso, Giuseppe. (2017). *Machine Learning Algorithms – A reference guide to popular algorithms for data Science and machine learning.* Recuperado de https://books.google.com.ec/books?id=_-ZDDwAAQBAJ&printsec=copyright&redir_esc=y#v=onepage&q&f=false

Brage, M. (2020). *Análisis de datos categóricos: regresión logística y multinomial* (Trabajo de Fin de Grado). Universidad de La Laguna, Tenerife, España.

Cabrera, E. y Díaz, E. (2018). Guide to Jupyter Notebooks for educational purposes. Universidad Complutense de Madrid, Madrid, España. Recuperado de https://www.um.es/documents/378246/2964900/Normas+APA+Sexta+Edición.pdf/27f8511d-95b6-4096-8d3e-f8492f61c6dc

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Cryptography. *Fernet cifrado simétrico.* Recuperado de https://cryptography.io/en/latest/fernet/

Falla, J. (2021). Predicción De Abandono De Clientes En Telecomunicaciones Mediante El Aprendizaje Automático (Tesis de posgrado). Universidad de Bogotá Jorge Tadeo Lozano, Bogotá, Colombia.

Gonzales, L. (2023). *Librería NumPy.* Recuperado de https://aprendeia.com/libreria-de-python-numpy-machine-learning/

IBM (2021). *Conceptos básicos de ayuda de CRISP-DM.* Recuperado de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Loo, Mark P.J. van der, and Edwin de Jonge. 2012. Learning RStudio for R Statistical Computing. Birmingham, UK: Packt publishing.

Mahesh, Bata. (2018). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR). *9*(1). 381-385. Recuperado de https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

NumPy. (2022). *NumPy User Guide.* Recuperado de https://numpy.org/doc/stable/user/whatisnumpy.html

Pandas. (2023). *About Pandas.* Recuperado de https://pandas.pydata.org/about/index.html

R-Project. *What is R?*. Recuperado de https://www.r-project.org/about.html

Raschka, S., y Mirjalili. V. (2019). *Python Machine Learning*. Recuperado de https://falksangdata.no/wp-content/uploads/2022/07/python-machine-learning-and-deep-learning-with-python-scikit-learn-and-tensorflow-2.pdf

Ray, Susmita. (Febrero de 2019). A Quick Review of Machine Learning Algortihms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). Conferencia llevada a cabo en Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.

Scikit – Learn. *Clustering.* Recuperado de https://scikit-learn.org/stable/modules/clustering.html#k-means

Teknomo, K. (2007). K-Means Clustering Tutorial. Recuperado de http://sigitwidiyanto.staff.gunadarma.ac.id/Downloads/files/38034/M8-Note-kMeans.pdf