UNIVERSIDAD MAYOR DE SAN SIMÓN

FACULTAD DE CIENCIAS Y TECNOLOGÍA

DIRECCIÓN DE POSGRADO

**DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES**

**SEGUNDA VERSIÓN**

**PROYECTO PRESENTADO PARA OBTENER EL GRADO DE LICENCIATURA EN INGENIERÍA DE SISTEMAS**

**MODALIDAD DOBLE TITULACIÓN**



**Cochabamba – Bolivia**

**2024**

MODELO DE ANÁLISIS PREDICTIVO PARA EL ABANDONO DE CLIENTES EN UNA EMPRESA DE ALOJAMIENTO WEB CON HERRAMIENTAS DE MACHINE LEARNING.

Por

Jeyson Oliver Bartolome Urrelo

El presente documento, Trabajo de Grado es presentado a la Dirección de Posgrado de la Facultad de Ciencias y Tecnología en cumplimiento parcial de los requisitos para la obtención del grado académico de Licenciatura en Ingeniería de Sistemas, modalidad Doble Titulación, habiendo cursado el Diplomado “Estadística Aplicada a la Toma de Decisiones” propuesta por el Centro de Estadística Aplicada (CESA) en su segunda versión.

ASESOR/TUTOR

Lic. Fernandez Guzman Helder Octavio

COMITÉ DE EVALUACIÓN

Ing. Ronald Patiño Tito, M.Sc. (Presidente)

Ing. Guillen Salvador Roxana, M.Sc. (Coordinador)

Ing. Espinoza Orosco José, M.Sc. (Tribunal)

Ing. Marcelo Antezana Camacho (Tribunal)



**Aclaración**

**Este documento describe el trabajo realizado como parte del programa de estudios de Diplomado “Estadística Aplicada a la Toma de Decisiones” en el Centro de Estadística Aplicada CESA y la Dirección de Posgrado de la Facultad de Ciencias y Tecnología. Todos los puntos de vista y opiniones expresadas en el mismo son responsabilidad exclusiva del autor y no representan necesariamente las de la institución.**

**Resumen**

La industria de alojamiento web es fundamental en la era digital, brindando servicios esenciales para empresas y particulares que buscan presencia en línea. La retención de clientes es crucial, ya que la competencia está presente y la satisfacción del cliente influye directamente en la reputación y el éxito a largo plazo de la empresa. El abandono de clientes puede resultar costoso y afectar la estabilidad financiera.

Este proyecto se centra en el análisis predictivo del abandono de clientes en una empresa de alojamiento web en el Estado Plurinacional de Bolivia. La metodología abarca desde la recopilación de datos hasta la aplicación de modelos de aprendizaje supervisado. Se utilizó la base de datos de la empresa “Boliviahost”, la cual presta servicios a nivel nacional e internacional, abarcando el período de 2017 a 2023, con un enfoque especial en la ubicación geográfica de Bolivia.

El desarrollo de esta investigación sigue la estructura metodológica del proceso CRISP-DM (Cross-Industry Standard Process for Data Mining). La primera fase comprendió el entendimiento del negocio, se procedió con la recopilación exhaustiva y la interpretación de los datos disponibles. La etapa de análisis exploratorio de datos permitió identificar patrones preliminares y entender mejor la distribución de las variables.

Posteriormente, se llevó a cabo el procesamiento de datos para prepararlos adecuadamente para el desarrollo de modelos predictivos. La implementación de algoritmos de Machine Learning, que incluyó el árbol de decisión, el bosque aleatorio, la Regresión Logística, Knn y Naive Bayes, se alineó con la fase de modelado del proceso CRISP-DM. Cada modelo se ajustó y evaluó cuidadosamente para garantizar su idoneidad en la predicción del abandono de clientes en la empresa de alojamiento web.

Los resultados cuantitativos destacan la notable eficacia de los modelos de árbol, especialmente el árbol de decisión y el bosque aleatorio. En términos de métricas de evaluación, el árbol de decisión sobresale con un Roc Auc del 92%, una precisión del 92%, un recall del 87%, y un F1 Score del 90%. De manera similar, el Árbol de decisión muestra un desempeño sólido con un Roc Auc del 91%, una precisión del 90%, un recall del 90%, y un F1 Score del 89%. Estos resultados resaltan la capacidad de predecir el abandono de clientes, superando a los modelos Knn, Naive Bayes y Regresión Logística.

Este estudio ofrece una visión integral de los factores detrás del churn y destaca la eficacia de los modelos de árbol. Las recomendaciones incluyen mejorar la satisfacción del cliente en servicios específicos, enfocar estrategias de retención en áreas geográficas clave y el uso de modelos de aprendizaje profundo para abordar la problemática de manera más precisa y avanzada.

**Palabras clave**

Abandono de cliente, alojamiento web, aprendizaje supervisado, análisis predictivo.

*Este proyecto está dedicado con profundo amor y gratitud a dos personas especiales en mi vida, Jheymi y mi querida hermanita Lilian.*

*En memoria de mi querida hermanita Lili, cuyo amor y determinación continúan inspirándome. Aunque no estés físicamente presente, siempre serás parte de mi vida y este trabajo es un tributo a tu eterno legado de fortaleza y amor.*

*A Jheymi, quien iluminó mi vida con amistad y sabiduría. Tu influencia perdura en este trabajo, recordando la importancia de la perseverancia y la amistad sincera.*

**Agradecimientos**

*A Dios, por ser la fuente de fortaleza en mi vida, le expreso mi sincero agradecimiento. Su amor incondicional y guía han sido mi sustento en cada paso del camino.*

*A mi universidad, Mayor de San Simon, por brindarme el entorno propicio para el aprendizaje, la exploración y el desarrollo académico*

*A todo el equipo de profesionales del diplomado “Estadística aplicada a la toma de decisiones”, por la transmisión de conocimientos y la invaluable información proporcionada, la cual fue fundamental para la realización y comprensión de este proyecto*

*Al Licenciado Fernandez Guzman Helder Octavio, por su valiosa y desinteresada colaboración, su orientación y su dedicación, que contribuyeron de manera significativa al desarrollo y éxito de este proyecto..*

*A mis queridos padres y hermanos, cuyo constante apoyo, amor y aliento han sido la fuente de mi fortaleza y motivación a lo largo de esta travesía académica.*

*A mis amigos, cuya amistad y ánimo fueron pilares esenciales durante los momentos desafiantes y gratificantes de este proyecto.*

*A todos ustedes, mi más profundo agradecimiento por haber sido parte integral de este importante capítulo académico. Su contribución ha dejado una huella duradera en mi formación profesional y personal.*

**Tabla de contenidos**

[1. Introducción 1](#_heading=h.30j0zll)

[1.1. Antecedentes 1](#_heading=h.1fob9te)

[1.2. Justificación 3](#_heading=h.3znysh7)

[1.3. Planteamiento del problema 3](#_heading=h.2et92p0)

[1.4. Objetivo general 4](#_heading=h.tyjcwt)

[1.4.1. Objetivos específicos 4](#_heading=h.3dy6vkm)

[2. Marco teórico 6](#_heading=h.4d34og8)

[2.1. Análisis Predictivo 6](#_heading=h.1oyxqrjlmf2l)

[2.2. Alojamiento Web 6](#_heading=h.2z5hcdbaovlp)

[2.3. Tasa de rotación de clientes (Churn) 6](#_heading=h.2s8eyo1)

[2.4. Metodología CRISP-DM 7](#_heading=h.4kbfotsutfkq)

[2.5. Aprendizaje automático 7](#_heading=h.sbk64o3z0llr)

[2.5.1. Tipos de Aprendizaje 8](#_heading=h.ytepduxho2oc)

[2.5.1.1. Aprendizaje supervisado 8](#_heading=h.9s59yc5im0ax)

[2.5.1.2. Aprendizaje no supervisado 9](#_heading=h.4i7t8kdvze0y)

[2.6. Modelos de aprendizaje supervisado 9](#_heading=h.f0yduigjlmoi)

[2.6.1. Modelos Lineales 9](#_heading=h.uwe6wa5hd4z5)

[2.6.1.1. Regresión Logística 9](#_heading=h.y5b1jaxcao99)

[2.6.2. Modelos de árbol 10](#_heading=h.ay2buqh97qj1)

[2.6.2.1. Bosque aleatorio 10](#_heading=h.i2rgg78n409e)

[2.6.2.2. Árbol de decisión 10](#_heading=h.gknxj7rvvjrk)

[2.7. Otros modelos de clasificación 11](#_heading=h.hs4n4swkq1z)

[2.7.1. Naive Bayes 11](#_heading=h.w918cijzm7yc)

[2.7.2. Vecinos más cercanos 11](#_heading=h.3c8kpwulssud)

[2.8. Variables descriptivas 11](#_heading=h.ufnu8lbdn0vd)

[2.9. Variables destino o etiqueta 11](#_heading=h.wpj0eqezat8x)

[2.10. La prueba chi-cuadrado 12](#_heading=h.aiku0q2z4rzm)

[2.11. Descripción estadística 12](#_heading=h.s62lkdfxo5g1)

[2.11.1. Media 12](#_heading=h.a4cgmq2am9re)

[2.11.2. Mediana 13](#_heading=h.cw1atnlov4tc)

[2.11.3. Moda 13](#_heading=h.gieefa158xk1)

[2.11.4. Minimo y Maximo 13](#_heading=h.hiprf34xpt21)

[2.11.5. Rango 13](#_heading=h.wjiydlwq7xlm)

[2.11.6. Varianza 13](#_heading=h.1ncf4527fzeu)

[2.11.7. Desviación Estándar 13](#_heading=h.3qr6pn7qxq9d)

[2.11.8. Correlación 14](#_heading=h.rhhkbsv4rhjk)

[2.12. Python en ciencia de datos 14](#_heading=h.n5p7gsuv7mpy)

[2.12.1. Manipulación de datos con Python 14](#_heading=h.czkik7w05yht)

[2.12.1.1. NumPy 14](#_heading=h.ly2i65z4gzfr)

[2.12.1.2. Pandas 15](#_heading=h.2wx416c0yftr)

[2.12.1.3. Matplotlib 15](#_heading=h.5nd6s75bty32)

[2.12.2. Aprendizaje automático con Python 15](#_heading=h.94xehnhoayw4)

[2.12.2.1. Scikit-learn 15](#_heading=h.g81ujbj9ilmv)

[2.12.3. Métricas de evaluación con Python 15](#_heading=h.6lw8l9xahgta)

[2.12.3.1. Exactitud(Accuracy) 15](#_heading=h.ymucirmq39iw)

[2.12.3.2. Precisión 15](#_heading=h.gr0lmrr26aja)

[2.12.3.3. Sensibilidad (recall) 16](#_heading=h.nz6lawqapa8)

[2.12.3.4. F1-score 16](#_heading=h.nc1njo194q34)

[2.12.3.5. Curvas Roc 16](#_heading=h.kswcui9xdp9)

[2.12.3.6. Matriz de confusión 17](#_heading=h.o6q2r59xio0)

[2.12.3.7. Validación cruzada por k iteraciones 17](#_heading=h.tsahzi3hzd6)

[2.12.3.8. Hiperparámetros y búsqueda en cuadrícula en Aprendizaje Automático 18](#_heading=h.eyn9bkx157fy)

[2.12.3.9. Curva de aprendizaje 18](#_heading=h.1pao42aom45j)

[3. Metodología. 20](#_heading=h.3rdcrjn)

[3.1. Área de estudio 20](#_heading=h.26in1rg)

[3.2. Flujograma metodológico 20](#_heading=h.35nkun2)

[3.3. Fuentes de información 22](#_heading=h.44sinio)

[3.3.1. Fuentes de información Secundaria 22](#_heading=h.p24g9w20gidp)

[3.4. Entendimiento del negocio 22](#_heading=h.2jxsxqh)

[3.4.1. Objetivo del negocio 22](#_heading=h.ppooeqej3esz)

[3.4.2. Situación Actual 22](#_heading=h.v3up5qqz7trf)

[3.4.3. Recursos y Restricciones 22](#_heading=h.j1ic32bbxkik)

[3.4.4. Plan de Proyecto Inicial 22](#_heading=h.pxcxeinata5g)

[3.5. Recopilación de datos 23](#_heading=h.jsv2v19n003c)

[3.6. Entendimiento de los datos 23](#_heading=h.se6833xo2qh6)

[3.7. Análisis Exploratorio de Datos 25](#_heading=h.xfa09ycwuibn)

[3.7.1. Descripción Estadística: 25](#_heading=h.b37alih9rr7u)

[3.7.2. Análisis Univariable 27](#_heading=h.am82wpzergq6)

[3.7.3. Análisis Bivariado 30](#_heading=h.kx7tstwda3mc)

[3.7.3.1. Correlación entre variables cuantitativas 31](#_heading=h.npt4foa0pin1)

[3.7.3.2. Prueba Chi-cuadrado para Variables Cualitativas 32](#_heading=h.jvqik3wksaxg)

[3.8. Procesamiento de los datos 33](#_heading=h.t4pt0yvd9iso)

[3.8.1. Tratamiento de valores nulos 33](#_heading=h.mg2nt0em2yzb)

[3.8.2. Tratamiento de valores atípicos 33](#_heading=h.7ei38pv8xzhc)

[3.8.3. Transformación de variables cualitativas 33](#_heading=h.8e4evddpcnil)

[3.9. Desarrollo de Modelo 33](#_heading=h.ycm0nz63ucly)

[3.9.1. Análisis del Problema 33](#_heading=h.z8yojf5jfsa1)

[3.9.2. Exploración de Algoritmos 33](#_heading=h.zecr7w2rrgu9)

[3.9.3. Implementación Inicial 33](#_heading=h.2tbh1faeoqd9)

[3.9.4. Evaluación 35](#_heading=h.l9nrglpnsvua)

[3.10. Ajuste de Modelos 36](#_heading=h.ku77b7bp47sa)

[3.10.1. Ajuste del modelo Knn 36](#_heading=h.xk7b6zh5cxqj)

[3.10.2. Ajuste del modelo Regresión Logística 37](#_heading=h.8rf7uphdgzoc)

[3.10.3. Ajuste del modelo Naive Bayes 37](#_heading=h.6hr0o7rste67)

[3.10.4. Ajuste del modelo Árbol de decisiones 37](#_heading=h.sm3r72q6ftkn)

[3.10.5. Ajuste del modelo Bosque aleatorio 38](#_heading=h.9ktpb7ytpxqj)

[3.11. Técnicas de Optimización 39](#_heading=h.heim21ldaw91)

[3.12. Validación Cruzada 39](#_heading=h.vi2gchgjgieq)

[**4. Resultados y Discusión 40**](#_heading=h.mwj3ee7o5awq)

[4.1. Análisis descriptivo de los datos 40](#_heading=h.147n2zr)

[4.2. Entrenamiento del modelo 42](#_heading=h.l9o46uqi4652)

[4.2.1. Resultados de la precisión de los modelos 42](#_heading=h.fgg7n57h37yt)

[4.3. Resultados de métricas de evaluación de Árbol de decisiones 43](#_heading=h.g1a4rla8ib9a)

[4.3.1. Curva de aprendizaje de Árbol de decisiones 45](#_heading=h.r1rz6f9arur0)

[4.4. Resultados de métricas de evaluación de Knn 46](#_heading=h.4u2y3yv14s2s)

[4.4.1. Curva de aprendizaje de Knn 47](#_heading=h.7d6o5ch5goji)

[4.5. Resultados de métricas de evaluación de Regresión logística 48](#_heading=h.5qogo8lvrpm7)

[4.5.1. Curva de aprendizaje de Regresión logística 49](#_heading=h.ks3p70q8gcld)

[4.6. Resultados de métricas de evaluación de Naive Bayes 50](#_heading=h.8jsqmeblirk2)

[4.6.1. Curva de aprendizaje de Naive Bayes 51](#_heading=h.gch023u339kq)

[4.7. Resultados de métricas de evaluación Bosque aleatorio 52](#_heading=h.n4kyhtcnjl0h)

[4.7.1. Curva de aprendizaje de Bosque aleatorio 53](#_heading=h.c10ap6jzybka)

[4.8. Discusión de resultados 55](#_heading=h.ihv636)

[5. Conclusiones 58](#_heading=h.41mghml)

[6. Recomendaciones 61](#_heading=h.2grqrue)

[Referencias bibliográficas 63](#_heading=h.vx1227)

[Anexos 66](#_heading=h.3fwokq0)

[Anexo 1. Base de datos 66](#_heading=h.29hop3xzimu8)

[Anexo 2. Consulta Sql 67](#_heading=h.faq9pnsl4cx)

[Anexo 3. Análisis y limpieza de Datos 68](#_heading=h.nvq0v7yc7ql)

[Anexo 4. Modelado 69](#_heading=h.oo5cy3ck3ya3)

[Anexo 5. Cd 70](#_heading=h.1v1yuxt)

**Lista de figuras**

[Figura 1-1: Conexiones al servicio de distribución de señal por gestión 4](#_heading=h.6f7fgm5jnjs3)

[Figura 2-1: Gráfico de Roc 16](#_heading=h.73oeg0akkyrj)

[Figura 2-2: Gráfica Matriz de confusión 17](#_heading=h.4q15ho6jcnd7)

[Figura 2-3: Gráfica Curva de aprendizaje 19](#_heading=h.8lrmge8mpbhz)

[Figura 3-1: Área de estudio 20](#_heading=h.lnxbz9)

[Figura 3-2: Flujograma metodológico 21](#_heading=h.1ksv4uv)

[Figura 3-3: Modelo entidad relación 24](#_heading=h.xtsx2pciw2vd)

[Figura 3-4: Histogramas para la variable ‘meses\_activo’ 27](#_heading=h.fijuoa9y2xux)

[Figura 3-5: Gráfico de boxplot para las variables ‘meses\_activo’ 27](#_heading=h.jnupakxo7it4)

[Figura 3-6: Histograma para las variables “Total” 28](#_heading=h.5r0s09noi5fu)

[Figura 3-7: Gráfico de boxplot para las variables “Total” 28](#_heading=h.z7m22f42twdm)

[Figura 3-8: Gráfica de barras para la variable “state” 29](#_heading=h.8uew4zsc54p9)

[Figura 3-9: Gráfica de barras para la variable “domainstatus” 29](#_heading=h.nn5hgqimiyxj)

[Figura 3-10: Gráfica de barras para la variable “domainstatus” 30](#_heading=h.llr3lpyzraus)

[Figura 3-11: Gráfica de barras para la variable “billingcycle” 30](#_heading=h.v3fly7trqcaj)

[Figura 3-12: Gráfica bivariado entre variables cuantitativas 31](#_heading=h.xud6xrlduuax)

[Figura 3-13: Gráfico de correlación entre las variables ‘total’ y ‘meses\_activo’ 31](#_heading=h.xemw3kmku1fp)

[Figura 4-1: Estado de las cuentas 40](#_heading=h.23ckvvd)

[Figura 4-2: Servicios con el mayor número de cuentas dadas de baja 41](#_heading=h.23ckvvd)

[Figura 4-3: Porcentaje de cuentas dadas de baja por departamento 41](#_heading=h.23ckvvd)

[Figura 4-4: Comparación entre modelos preliminares y ajustados 42](#_heading=h.4d2raes1uz6b)

[Figura 4-5: Precisión entre modelos 43](#_heading=h.enwe65xv1d4r)

[Figura 4-6: Matriz de confusión árbol de decisiones 44](#_heading=h.1ugoqqocyavw)

[Figura 4-7: Métricas de evaluación Arbol de decisiones 45](#_heading=h.23ckvvd)

[Figura 4-8: Curva de aprendizaje Árbol de decisión 46](#_heading=h.x32qupa3f30z)

[Figura 4-9: Matriz de confusión Knn 47](#_heading=h.23ckvvd)

[Figura 4-10: Métricas de evaluación knn 47](#_heading=h.aqs5ns41j85n)

[Figura 4-11: Curva de aprendizaje para Knn 48](#_heading=h.1b9x5l1wjvwj)

[Figura 4-12: Matriz de confusión regresión logística 49](#_heading=h.d4i5razhtu24)

[Figura 4-13: Métricas de evaluación regresión logística 49](#_heading=h.23ckvvd)

[Figura 4-14: Curva de aprendizaje regresión logística 50](#_heading=h.hoih11g9ghtu)

[Figura 4-15: Matriz de confusión Naive Bayes 51](#_heading=h.23ckvvd)

[Figura 4-16: Métricas de Naive Bayes 51](#_heading=h.23ckvvd)

[Figura 4-17: Curva de aprendizaje para Naive Bayes 52](#_heading=h.ru21kb7zdbc2)

[Figura 4-18: Matriz de confusión Bosque aleatorio 53](#_heading=h.7523dqt2mhd4)

[Figura 4-19: Métricas de evaluación random forest 53](#_heading=h.qro2idphk90i)

[Figura 4-20: Curva de aprendizaje Bosque aleatorio 54](#_heading=h.wubchxy7tn7k)

[Figura 4-21: Comparación entre métricas de evaluación 56](#_heading=h.kv33nkewxrss)

[Figura 4-22: Comparación entre porcentajes de churn 57](#_heading=h.thq1jlc1ox6l)

**Lista de tablas**

[Tabla 3-1: Descripción de las tablas de la base de datos 16](#_heading=h.tikz7q24bd2v)

[Tabla 3-2: Descripción de los campos de la tabla resultado de la ejecución de la consulta SQL 17](#_heading=h.2bn6wsx)

[Tabla 3-3: Descripción de los tipos de variable 17](#_heading=h.igltis9sng4)

[Tabla 3-4: Descripción Estadística 19](#_heading=h.7ew08hukxmxv)

[Tabla 3-6: Prueba Chi-cuadrado para Variables Cualitativas 24](#_heading=h.uc04ehnwov7f)

[Tabla 4-1: Evaluación de modelos de aprendizaje automático 43](#_heading=h.j7o7pjzay5ho)

[Tabla 4-2: Métricas de evaluación entre modelos comunes 44](#_heading=h.9hq1v8goh5rt)

[Tabla 4-3: Evaluación de modelos de aprendizaje automático 44](#_heading=h.1res3gl3yes5)

[Tabla 4-4: Comparación entre modelos 45](#_heading=h.dlqivz9yup28)

# Introducción

En un mundo empresarial cada vez más competitivo y dinámico, la pérdida de clientes se ha convertido en un desafío crucial que impacta el éxito y la prosperidad de las organizaciones. La capacidad de comprender y anticipar por qué los clientes eligen abandonar un servicio es esencial para lograr un crecimiento sostenible y mantener una ventaja competitiva sólida. En Bolivia, empresas de sectores clave como banca, telecomunicaciones y supermercados se esfuerzan por cultivar la lealtad de sus clientes y evitar la pérdida, lo cual impacta directamente en sus ingresos y la pérdida de participación de mercado (Periódico Opinión, 2017).

Hoy en día, esperamos que las aplicaciones web y móviles estén disponibles para todo el mundo. Esto es posible gracias a las empresas de alojamiento web, que se encargan de poner en funcionamiento y administrar estas aplicaciones en servidores especiales. De esta manera, se facilita que puedas acceder a estas aplicaciones desde cualquier dispositivo, ya sea una computadora, un teléfono móvil o una tablet (Amazon web service, s.f.).

Esta investigación se enfoca en el análisis predictivo del abandono de clientes, con el objetivo de construir un modelo predictivo altamente preciso que permita identificar a aquellos clientes con una mayor probabilidad de abandonar el servicio de alojamiento web. Para alcanzar este objetivo, se emplearán técnicas avanzadas de Machine Learning y análisis estadístico aplicadas a los datos históricos de los clientes. Estos métodos consideran diversos factores, como los patrones de compra, las interacciones con el servicio al cliente y otras variables clave.

## Antecedentes

El alojamiento web o hosting se refiere al servicio que proporciona espacio en servidores conectados a Internet para alojar y servir sitios web. Con la llegada de “World Wide Web”, comúnmente conocida como la Web, tuvo su génesis en 1989 en el CERN (Centro Europeo para la Investigación Nuclear) en Ginebra, Suiza. En ese tiempo, el investigador británico Tim Berners-Lee se encontraba inmerso en la búsqueda de una solución eficaz para abordar el problema de la proliferación y heterogeneidad de la información en la incipiente red de Internet. Integrando servicios ya existentes, como el ampliamente utilizado Gopher en esa época, Berners-Lee desarrolló la arquitectura fundamental que ha evolucionado hasta convertirse en la Web contemporánea. Partiendo de ese programa, en octubre Tim Berners-Lee comienza a desarrollar un editor y navegador gráfico de hipertexto para “NeXTStep”, eligiendo “WorldWideWeb” como nombre del programa y “World Wide Web” como nombre del proyecto. En noviembre se instala el primer servidor del mundo web llamado “NeXTcube” y se publica la primera página web (Luján Mora, 2002, p.17-18).

En 1998, el escenario de los servidores presenció otro avance con el surgimiento del Sun Ultra II, conocido como el primer servidor de Google. Inicialmente albergó el motor de búsqueda Backrub de Larry Page y Sergey Brin en la Universidad de Stanford, marcando el inicio de lo que eventualmente se convertiría en Google. El Sun Ultra II presentaba dos procesadores de 200 MHz y 256 MB de RAM (De Leon, 2017).

El alojamiento web en la actualidad ha experimentado una notable evolución, existe una amplia variedad de servicios de alojamiento web disponibles, que van desde opciones compartidas hasta servidores dedicados y servicios en la nube. Estas soluciones ofrecen diferentes niveles de recursos, rendimiento y seguridad para satisfacer las diversas necesidades de los usuarios.

El análisis predictivo, en la actualidad, ha adquirido una relevancia sin precedentes en diversos campos y sectores del mundo. Esta disciplina, respaldada por los avances en inteligencia artificial y aprendizaje automático, ha demostrado su capacidad para transformar datos en conocimientos y tomar decisiones informadas sobre eventos futuros. Desde el ámbito empresarial hasta la medicina, abarcando el marketing, el turismo y otras disciplinas, el análisis predictivo se ha vuelto una herramienta esencial para anticipar tendencias, descubrir patrones ocultos y mejorar la toma de decisiones. Este impacto se refleja claramente en un estudio reciente realizado por (Valle Díaz & Huamán Romaní, 2022), en el cual se empleó el análisis predictivo para examinar indicadores macroeconómicos del sector turismo en el país de Perú después del periodo post Covid-19.

En el contexto boliviano, usar de manera innovadora el análisis predictivo puede ser clave para adelantarse a las necesidades de los clientes y mejorar servicios, ganando así ventaja competitiva. Es esencial que las empresas comprendan plenamente cómo esta innovación puede mantener su posición en el mercado.

Por ese motivo acudimos a la revisión de documentación de diferentes fuentes para comprender los antecedentes que puedan dar soporte al estudio.

El estudio “Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión” llevado a cabo por (Contreras Morales et al., s.f.) emplea modelos de aprendizaje supervisado para explicar de manera satisfactoria las condiciones que deben cumplirse para que ocurra el abandono voluntario de clientes en empresas de telecomunicaciones que ofrecen servicios post pago de televisión digital.

Se realizó un estudio que expone varios modelos de aprendizaje supervisado, utilizando la metodología de minería de datos 'Descubrimiento de Conocimiento en Bases de Datos (KDD)', con el propósito de prever la deserción de clientes en una empresa de telecomunicaciones según (Barrientos & Ríos, 2013).

Adicionalmente, en el trabajo de (Falla Arango, 2021) hace uso de una variedad de modelos de aprendizaje supervisado. Entre ellos, destaca el modelo XGBoost, para anticipar el abandono de clientes en una compañía de telecomunicaciones.

## Justificación

La empresa se verá beneficiada al contar con un modelo de análisis predictivo que permitirá desarrollar acciones para mejorar la retención de clientes, fortalecerá su posición en el mercado y aumentará su ventaja competitiva. Al priorizar la satisfacción del cliente y mantener relaciones a largo plazo, la empresa podrá construir una base de clientes leales y comprometidos, lo que generará un flujo constante de ingresos y una mayor estabilidad económica.

El análisis predictivo con Machine Learning permitirá a la empresa procesar grandes cantidades de datos de manera eficiente, identificando patrones de comportamiento y tendencias en la retención de clientes. Estos modelos avanzados proporcionarán una visión más clara y precisa de los factores que influyen en la retención, permitiendo a la empresa tomar decisiones basadas en datos sólidos.

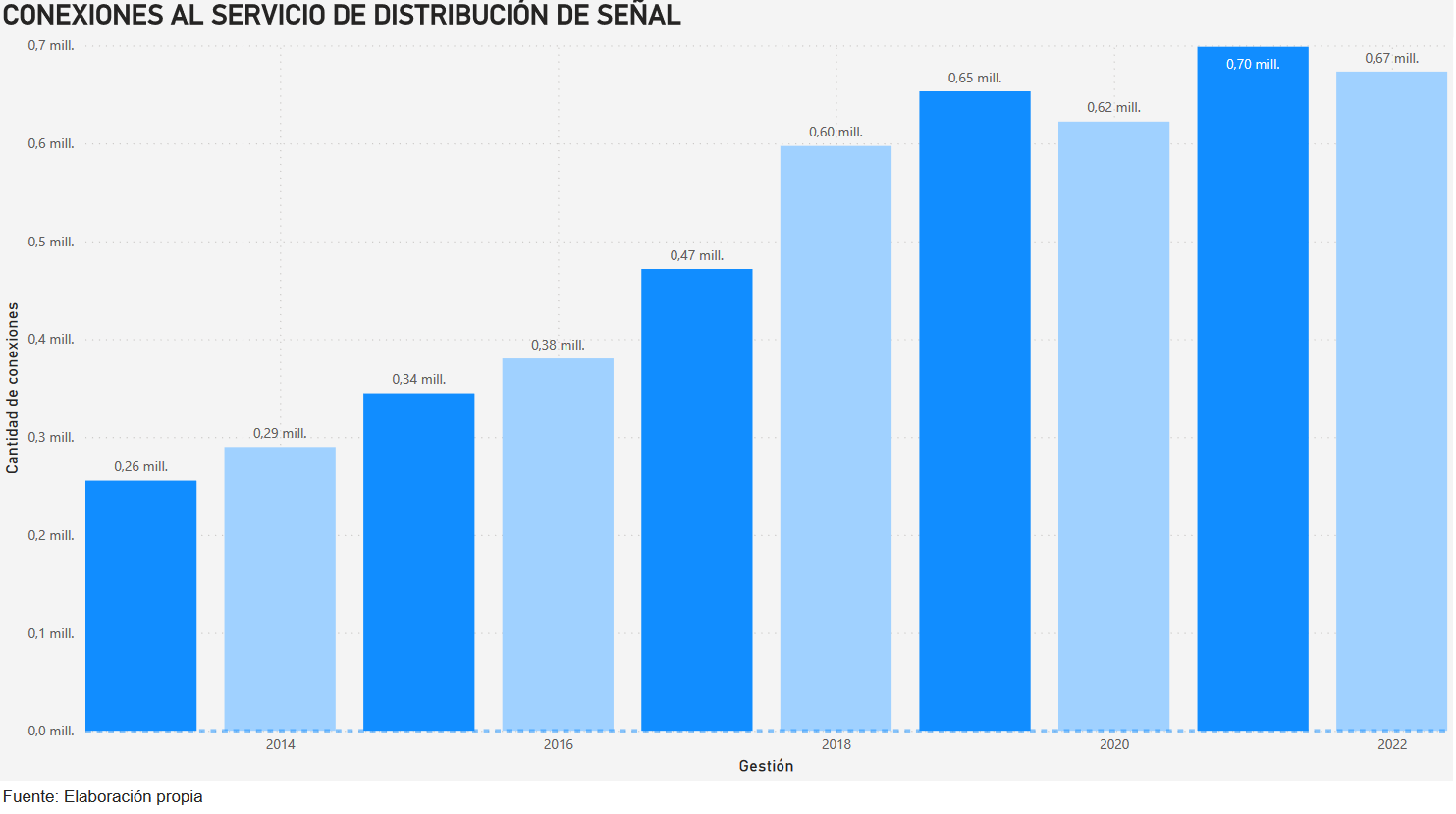
Reforzar la retención de clientes es una estrategia económicamente eficiente, al reducir los costos vinculados a la obtención de nuevos clientes. Además, fomenta relaciones duraderas, propiciando un incremento de los ingresos a largo plazo mediante la construcción de vínculos fuertes y sostenibles con los clientes.

## Planteamiento del problema

El abandono de clientes representa una pérdida significativa para la empresa, tanto en términos de ingresos como de reputación. Adquirir nuevos clientes puede ser costoso y requiere esfuerzos considerables en marketing y publicidad. Por lo tanto, resulta esencial para las empresas de alojamiento web entender los factores que influyen en la retención de clientes y desarrollar estrategias efectivas para mantenerlos satisfechos y comprometidos.

En el ámbito digital, tanto las empresas de alojamiento web como los proveedores de telefonía, internet y televisión por cable son, ante todo, empresas de servicios. Aunque cada una ofrece servicios diferentes, comparten el desafío común de retener a sus clientes en un entorno en constante cambio. Deben ajustarse a las preferencias cambiantes de los consumidores, especialmente en medio de la creciente competencia de servicios en línea. En consecuencia, comparten la tarea fundamental de comprender y anticipar las necesidades de los clientes, garantizando así su lealtad en sus respectivas industrias de servicios. Esta conexión proporciona un marco útil para examinar la problemática que abordaremos.

A medida que se ha desarrollado el panorama de conexiones del servicio de Televisión por Cable (TVC), se observa crecimiento en las conexiones al servicio de distribución de señal por gestión hasta la gestión 2019, en la gestión 2020 se produce una reducción, sin embargo y en gran medida por las ofertas empaquetadas, la cantidad de conexiones en la gestión 2021 se incrementó en un 12,27%, 76.348 conexiones más con relación a la gestión anterior y finalmente en la gestión 2022 se observa una nueva reducción de 3,62%, con 673.392 conexiones como se visualiza en la figura 1-1 (BET-22, 2023).

La disminución en este mercado se relaciona con el creciente aumento de los servicios de transmisión en línea, lo cual ha llevado a que los clientes migren hacia otras opciones. Estos servicios, como Disney Plus, Netflix, Amazon Prime Video, entre otros, están atrayendo a una parte de la audiencia (BET-22, 2023).

###### Figura 1-1: Conexiones al servicio de distribución de señal por gestión

**Fuente: BET-22, (2023)**

El desafío radica en cómo identificar a los clientes que tienen mayor riesgo de abandonar el servicio antes de que lo hagan. Los métodos tradicionales de análisis y seguimiento pueden no ser suficientemente efectivos para anticipar el comportamiento del cliente en un mercado en constante cambio.

## Objetivo general

Desarrollar un modelo predictivo utilizando herramientas de machine learning para detectar y anticipar el abandono de clientes en una empresa de alojamiento web.

### Objetivos específicos

* Revisar y comparar modelos existentes de predicción de abandono de clientes en la industria de tecnologías de información, utilizadas en investigaciones previas.
* Recopilar y analizar datos de clientes, patrones de compra y la duración de su relación con el servicio, para comprender su comportamiento a lo largo del tiempo.
* Implementar modelos de predicción de abandono utilizando técnicas de Machine Learning con el fin de anticipar potenciales casos de clientes que podrían abandonar la empresa de alojamiento web.
* Evaluar la efectividad de los modelos de predicción en términos de retención de clientes.

# Marco teórico

En este capítulo se exponen de forma muy puntual los conceptos, metodologías y otros elementos esenciales que serán aplicados en el desarrollo del proyecto. Estos conceptos no solo actuarán como los pilares teóricos de nuestro enfoque, sino que también servirán como las directrices que darán forma y sustancia a nuestro análisis y toma de decisiones.

## Análisis Predictivo

El análisis predictivo comprende un conjunto de técnicas estadísticas dentro de la analítica avanzada, con el objetivo de anticipar eventos futuros basándose en sucesos pasados. Más que una predicción adivinatoria, el análisis predictivo ofrece la oportunidad de mejorar procesos actuales que influyen en esos eventos. Respaldado por diversas técnicas estadísticas como la modelización, el aprendizaje automático y la minería de datos, el análisis predictivo proporciona un grado de confianza a sus análisis, apoyando a la toma de decisiones (López Navas et al., 2020).

## Alojamiento Web

El alojamiento web es un servicio brindado por un proveedor de servicios de Internet (ISP) que permite a los usuarios tener un espacio en un servidor para almacenar y gestionar información de manera segura. Este servicio no se limita solo a alojar páginas web, sino que también permite hospedar archivos diversos, como imágenes, videos, documentos y correos electrónicos, facilitando la administración completa de contenido en línea (Gutierrez, 2015).

## Tasa de rotación de clientes (Churn)

Tasa de rotación de clientes o churn se refiere al proceso de abandono o discontinuo de un contrato o relación comercial por parte de un cliente con una empresa. Este proceso puede involucrar a los clientes que deciden dejar de utilizar los servicios de una empresa y, en su lugar optan por cancelar su contrato o cambiar a otro proveedor. El churn puede ocurrir por diversas razones, como insatisfacción con la calidad del servicio, altos costos,falta de competitividad en los planes de precios, problemas de soporte, entre otros motivos (Lazarov & Capota, 2007).

Existen diferentes tipos de churn los cuales son: churn activo, donde el cliente toma la decisión consciente de abandonar el contrato y cambiar a otro proveedor, el churn rotacional, donde el cliente deja el contrato sin necesariamente cambiar a un competidor y el churn pasivo, dondees la empresa la que decide poner fina al contrato (Lazarov & Capota, 2007).

En la era digital actual, el análisis de datos desempeña un papel crucial en la gestión del churn. Las empresas pueden utilizar herramientas analíticas para identificar patrones de comportamiento de los clientes que puedan indicar una mayor probabilidad de churn, permitiendo así una intervención proactiva.

## Metodología CRISP-DM

CRISP-DM(Cross Industry Standard Process for Data Mining) por sus siglas en inglés es el estándar abierto más usado para los proyectos de aprendizaje automático y minería de datos. Es un modelo genérico, flexible y cíclico diseñado para ser adaptado por empresas o individuos para cubrir sus necesidades en procesos de minería de datos, es decir aquellos en donde es primordial descubrir patrones o información útil que se encuentra oculta en los datos.

Esta metodología proporciona un enfoque sistemático para abordar los desafíos específicos que enfrentamos al anticipar y prevenir el abandono de clientes y está conformado por las siguientes fases:

* Conocimiento del negocio: La primera fase consiste en entender el negocio y organización a la cual se va a ayudar a suplir alguna necesidad por medio del aprendizaje automático.
* Conocimiento de los datos: En esta fase el analista debe tener conocimiento de las diferentes fuentes de datos con las cuales cuenta la organización y los tipos de datos disponibles.
* Preparación de datos: Consiste en convertir los datos en algún formato específico de tal manera que sean lo suficientemente aptos para poder ser usado por un algoritmo de aprendizaje automático.
* Modelado: se utilizan diversos algoritmos de aprendizaje automático, desarrollamos modelos predictivos que pueden identificar patrones de comportamiento que indican el riesgo de abandono. Exploramos y probamos diferentes algoritmos para encontrar el más adecuado para nuestro contexto.
* Evaluación: El modelo es sometido a diferentes tareas de evaluación o desempeño a fin de verificar que sea capaz de hacer predicciones con una buena exactitud.
* Despliegue: Esta fase cubre todo el proceso de integrar un modelo de aprendizaje automático dentro de un sistema de información organizacional (Pineda Pertuz, 2022, p. 45-46).

## Aprendizaje automático

Es una rama de la inteligencia artificial que busca que un programa de computadora aprenda de un conjunto de datos con los cuales se entrenó, y buscará identificar un patrón con el que puede realizar predicciones sobre nuevos datos. A diferencia de la programación tradicional donde en primera instancia se procesan un conjunto de datos de entrada y por medio de una serie de reglas se genera una salida, en aprendizaje automático. Los datos y las salidas son los datos iniciales que mediante un proceso de entrenamiento produce las reglas, las cuales comúnmente reciben el nombre de modelo (Pineda Pertuz, 2022, p.35-36).

Este proceso de aprendizaje se lleva a cabo mediante un conjunto de etapas clave:

* Recopilación de Datos: En primer lugar, se recopilan y preparan datos relevantes para el problema que se desea resolver. Estos datos pueden incluir información histórica, características de entrada y resultados deseados.
* Limpieza de Datos: Los datos recopilados pueden contener valores atípicos o información redundante. En esta etapa, se realiza la limpieza de datos para asegurarse de que estén en un formato adecuado y sean de calidad.
* Entrenamiento del Modelo: Utilizando el conjunto de datos limpios, se entrena un modelo de aprendizaje automático. Durante el entrenamiento, el modelo busca patrones y relaciones en los datos para poder hacer predicciones o tomar decisiones.
* Validación y Evaluación: El modelo entrenado se evalúa utilizando datos no utilizados previamente (conjunto de validación o prueba) para medir su rendimiento y asegurarse de que sea capaz de generalizar adecuadamente.
* Ajuste y Optimización: En función de los resultados de la evaluación, se pueden realizar ajustes en el modelo o se pueden explorar diferentes algoritmos y enfoques para mejorar su rendimiento.

Es importante destacar que el Aprendizaje Automático abarca una variedad de técnicas y algoritmos, incluyendo la regresión, clasificación, clustering, redes neuronales, entre otros.

### Tipos de Aprendizaje

#### Aprendizaje supervisado

El objetivo principal del aprendizaje supervisado es aprender un modelo, a partir de datos de entrenamiento etiquetados, que nos permite hacer predicciones sobre datos futuros o no vistos. Aquí el término supervisado se refiere a un conjunto de muestras donde las señales de salida deseadas o etiquetas ya se conocen (Raschka & Mirjalili, 2019).

se distinguen dos tipos de aprendizaje supervisado:

* Regresión: En este tipo de aprendizaje, el objetivo es encontrar una función que relacione las características de entrada con una variable objetivo continua y, a partir de esta relación, predecir valores específicos.
* Clasificación: El objetivo principal es predecir a cuál de las categorías pertenece una nueva instancia de datos, basándose en el aprendizaje de patrones y relaciones a partir de un conjunto de datos etiquetado previamente.

#### Aprendizaje no supervisado

En el Aprendizaje No Supervisado, a diferencia del supervisado, no se proporcionan etiquetas o resultados aquí solo le damos las características al algoritmo, nunca las etiquetas. Queremos que agrupe los datos que le dimos según sus características. El algoritmo solo sabe que como los datos comparten ciertas características, de esa forma asume que pueda que pertenezcan al mismo grupo (Sandoval, 2018, p. 37).

Para el desarrollo de nuestro proyecto, se emplea el enfoque del aprendizaje supervisado de clasificación. En este tipo de aprendizaje, el objetivo primordial es entrenar un modelo utilizando datos de entrenamiento previamente etiquetados. Esto nos permitirá hacer predicciones precisas sobre datos futuros o no vistos. El término 'supervisado' se refiere a que nuestro conjunto de muestras de entrenamiento ya incluye las etiquetas conocidas. Este enfoque será esencial para resolver el problema específico que abordaremos en nuestro proyecto.

## Modelos de aprendizaje supervisado

Son el resultado del proceso de aprendizaje automático, y típicamente es producido por un algoritmo a partir del entrenamiento con una parte de los datos. El modelo constituye un conocimiento base con el cual es posible realizar predicciones al aplicarlo sobre nuevos datos de entrada (Pineda Pertuz, 2022, p. 40).

### Modelos Lineales

Estos tratan de encontrar un línea que se ajuste bien a la nube de puntos que se disponen. Aquí destacan desde modelos muy conocidos y usados como la regresión lineal también conocidos y usados como la regresión de mínimos cuadrados, la logística el cual es la adaptación de la línea a problemas de clasificación cuando son variables discretas o categóricas. Estos dos modelos tienen el problema del overfit esto significa que se ajustan demasiado a los datos disponibles, con el riesgo que esto tiene para nuevos datos que pudieran llegar. Al ser modelos relativamente simples no ofrecen resultados muy buenos para comportamientos más complicados (Sandoval, 2018, p. 38).

#### Regresión Logística

El concepto de regresión hace referencia a la ley experimental o fórmula matemática que traduce la relación entre variables correlacionadas. Generalmente cuando se quiere poner una variable en función de otra u otras, se acude al bien conocido recurso de regresión lineal simple o múltiple. Esta función utiliza normalmente el método de mínimos cuadrados y funciona fluidamente desde el punto de vista aritmético.

La regresión logística es un método en el que se calcula la probabilidad de que ocurra un evento específico utilizando variables independientes que se consideran relevantes o influyentes. Esta técnica implica derivar una función logística a partir de las variables independientes, lo que a su vez permite clasificar a los individuos en uno de los dos grupos definidos por los dos posibles valores de la variable dependiente. En resumen, la regresión logística es una herramienta utilizada para modelar relaciones entre variables y para realizar clasificaciones en problemas de dos categorías (Fiuza Pérez & Rodríguez Pérez, 2000).

### Modelos de árbol

Son modelos precisos, estables y más sencillos de interpretar básicamente porque construyen unas reglas de decisión que puede representar como un árbol. A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas. En estos modelos destacan los árboles de decisión y los random forest al ser más precisos y elaborados (Sandoval, 2018, p. 38).

#### Bosque aleatorio

Es una mejora de los modelos de árboles de decisión, utiliza la aleatorización para crear múltiples árboles de decisión. Estos árboles se combinan para formar un "bosque", y la aleatorización se aplica al muestrear datos con reemplazo y seleccionar predictores de manera aleatoria en cada nodo de decisión. La fortaleza del bosque aleatorio radica en su capacidad para descubrir interacciones y efectos no lineales entre los predictores sin necesidad de predefinir los. El modelo resultante se utiliza para predecir nuevos casos mediante clasificación o regresión. Aunque la interpretación es más compleja debido a la cantidad de árboles, se emplean medidas de importancia de variables para evaluar la influencia de cada predictor. La precisión se mide mediante la tasa de error promedio de los árboles individuales utilizando casos no utilizados en el entrenamiento (Rigatti, 2017).

#### Árbol de decisión

Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observación y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema. Constituyen probablemente el modelo de clasificación más utilizado y popular. Un árbol gráficamente se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada uno de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas salen de cada uno de estos nodos y se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver (Barrientos Martínez et al., 2009).

## Otros modelos de clasificación

### Naive Bayes

Los clasificadores bayesianos son clasificadores estadísticos. Pueden predecir la pertenencia de una clase, estimando la probabilidad de que una muestra pertenezca a una clase específica. Estos clasificadores se fundamentan en el teorema de Bayes. En el caso de los clasificadores bayesianos, se asume que el impacto del valor de un atributo en una clase determinada es independiente de los valores de los otros atributos (Naive Bayesian Classifier, 2007).

Podemos calcular la probabilidad de un suceso A, sabiendo además que ese A cumple cierta característica que condiciona su probabilidad. El teorema de Bayes entiende la probabilidad de forma inversa al teorema de la probabilidad total. El teorema de la probabilidad total hace inferencia sobre un suceso B, a partir de los resultados de los sucesos A. Por su parte, Bayes calcula la probabilidad de A condicionado a B (López, s.f.).

### Vecinos más cercanos

La clasificación de vecinos más cercanos, también conocida como K-nearest neighbors(KNN por sus siglas en inglés), se basa en la idea de que los patrones más próximos a un patrón objetivo X, para el cual buscamos la etiqueta, brindan información útil sobre la etiqueta. Knn asigna la etiqueta de clase de la mayoría de los patrones K más cercanos en el espacio de datos. Para lograr esto, debemos ser capaces de definir una medida de similitud en el espacio de datos (Kramer, 2013, p. 13-23).

## Variables descriptivas

Son los elementos que describen las instancias almacenadas en el conjunto de datos.Estas variables o atributos tiene un tipo de dato asociado. (Pineda Pertuz, 2022, p. 39)

La selección cuidadosa de estas variables puede mejorar la calidad y la eficacia de los modelos, permitiendo predicciones más precisas y útiles en una amplia variedad de aplicaciones, como la predicción de ventas, la detección de enfermedades o la recomendación de contenido personalizado.

## Variables destino o etiqueta

Son las características para etiquetar las instancias, aspecto necesario en algunos tipos de algoritmos para realizar la tarea de entrenamiento del modelo. También son designadas para representar aquellos atributos que se quieran predecir.

Las Variables Destino son especialmente importantes en ciertos tipos de algoritmos de aprendizaje supervisado, donde el modelo se entrena utilizando ejemplos previamente etiquetados. Durante el proceso de entrenamiento, el modelo aprende a relacionar las características de entrada con los valores de las Variables Destino. Una vez que el modelo ha sido entrenado de manera efectiva, puede utilizar esta relación aprendida para hacer predicciones precisas sobre nuevas instancias que no han sido etiquetadas previamente(Pineda Pertuz, 2022, p.39).

## La prueba chi-cuadrado

La prueba chi-cuadrado consiste en tomar una muestra y observar si hay diferencias significativas entre las frecuencias observadas y las especificadas por la ley teórica del modelo que se contrasta, también denominadas “frecuencias esperadas”. O sea, el test o prueba chi-cuadrado, comparar dos tablas, una observada (de frecuencias obtenidas en nuestro estudio), con otra teórica (de frecuencias esperadas y calculadas teóricamente).

Fue Karl Pearson quien propuso el siguiente estadístico:

Donde:

Oi: Frecuencias observadas. Es el número de casos observados clasificados en una determinada celda de la tabla de contingencia.

Ei: Frecuencias esperadas o teóricas. Es el número de casos esperados correspondientes a cada celda de la tabla de contingencia. Se puede definir como aquella frecuencia que se observaría si ambas variables fueran independientes.

Como en cualquier prueba de contraste estadístico, se intenta rechazar la hipótesis nula y aceptar en consecuencia, la hipótesis alternativa. La hipótesis nula se corresponde con la independencia de las variables, o lo que es lo mismo, que las diferencias entre las frecuencias observadas y esperadas son muy pequeñas, y en consecuencia, el estadístico chi cuadrado también obtendrá un valor muy pequeño (Rigatti, 2017, p-32).

## Descripción estadística

### Media

La estimación más elemental para la localización de un conjunto de datos es el valor medio o promedio. La media se obtiene sumando todos los valores y dividiéndolos por el número total de valores. Esta medida proporciona un centro representativo que facilita la comprensión de la tendencia central de los datos. Sin embargo, es importante tener en cuenta que la media puede verse afectada por valores atípicos, lo que significa que su valor puede ser distorsionado por observaciones extremas en el conjunto de datos (García Herrero et al., 2018, p. 40).

### Mediana

La mediana es el valor que se sitúa en el centro de un conjunto de datos, dividiéndolo en dos mitades iguales, donde la mitad de los datos se encuentra por encima y la otra mitad por debajo de este valor central. A diferencia de la media, la mediana es una medida de la posición central que no se ve afectada por valores extremos. Esta característica la hace particularmente útil en situaciones donde los datos presentan sesgo o distribuciones asimétricas, ya que la mediana proporciona una estimación más robusta del centro de la distribución (García Herrero et al., 2018, p. 41).

### Moda

El valor que aparece con mayor frecuencia en un conjunto de datos se conoce como moda. Identificar la moda es fundamental para reconocer los valores más frecuentes en la distribución de los datos, proporcionando así una descripción útil y concisa de la frecuencia de ocurrencia de diferentes valores en el conjunto (García Herrero et al., 2018, p. 41).

### Minimo y Maximo

Dentro de un conjunto de datos, el término "máximo" se refiere al valor más alto, mientras que el término "mínimo" hace referencia al valor más bajo. Estos puntos son cruciales para establecer los límites que delinean el rango de distribución de los valores en nuestros datos

### Rango

El rango es el intervalo de datos que existe entre el valor máximo y mínimo de un conjunto de datos. Este parámetro nos proporciona una visión general del grado de dispersión de nuestros datos, aunque es importante señalar que la varianza y la desviación estándar ofrecen medidas más precisas y detalladas de la dispersión estadística (García Herrero et al., 2018, p. 41).

### Varianza

La varianza se define como el promedio de las desviaciones al cuadrado respecto al valor medio de un conjunto de datos. Esta medida representa otra forma de evaluar la dispersión de los datos y se emplea en conjunto con la desviación estándar para cuantificar la variabilidad del conjunto en su totalidad. La varianza proporciona una perspectiva adicional sobre la extensión de los valores con respecto a la media, contribuyendo así a una comprensión más completa de la distribución de los datos (García Herrero et al., 2018, p. 41).

### Desviación Estándar

La desviación estándar, también conocida como promedio de las desviaciones al cuadrado respecto al valor medio, es una medida clave que ofrece información detallada sobre la variabilidad de un conjunto de datos. Si la desviación estándar es baja, entonces indica que los datos están agrupados cerca de la media del conjunto de datos. Y al contrario, entre más alta sea la desviación estándar, indica que los valores se extienden sobre un rango más amplio de datos (García Herrero et al., 2018, p. 41).

### Correlación

La correlación es un indicador que cuantifica la relación proporcional entre dos variables cuando una de ellas varía de manera sistemática en relación con los valores correspondientes de la otra. Cuando la correlación es alta entre las variables A y B, sugiere que si los valores de A aumentan o disminuyen, los valores de B también tienden a hacer lo mismo. Es importante destacar que, aunque una correlación significativa entre dos variables puede ser evidente, no implica directamente una relación causal entre ellas (García Herrero et al., 2018, p. 31).

## Python en ciencia de datos

Python es un lenguaje de programación de código abierto e interpretado, esto último significa que cada línea de código que escribamos es leída por un intérprete que las va ejecutando. Esta condición le otorga más rapidez en la ejecución de los programas, puesto que estos no tiene que compilarse previamente ahorrando el tiempo de compilación, Además, Python es un lenguaje multiplataforma, lo cual hace que cualquier programa que hagamos puede correr sin ningún problema en diferentes plataformas o sistemas operativos, as mismo este software de desarrollo dispone de una sintaxis muy simple lo que hace que su curva de aprendizaje no sea tan pronunciada. Cuenta con una una basta cantidad de librerías actualizadas permanentemente por su gran comunidad de desarrolladores (Saldaña, 2011).

### Manipulación de datos con Python

Python es la herramienta para el análisis de datos, ya que cuenta con un vasto ecosistema de bibliotecas y herramientas especializadas diseñadas específicamente para tareas de procesamiento, visualización y modelado de datos. Este ecosistema incluye bibliotecas altamente reconocidas como Pandas, NumPy y Matplotlib, que brindan a los analistas y científicos de datos las herramientas necesarias para manipular datos, realizar cálculos complejos y visualizar resultados de manera efectiva. Python, se ha consolidado como una alternativa para aquellos que buscan transformar datos en conocimiento y tomar decisiones informadas en una amplia gama de aplicaciones y sectores industriales (Pineda Pertuz, 2022, p.39).

### NumPy

NumPy es una biblioteca fundamental para la informática científica en Python. Proporciona un objeto de matriz multidimensional y diversas rutinas para realizar operaciones eficientes en matrices. A diferencia de las secuencias estándar de Python, las matrices NumPy tienen un tamaño fijo y requieren elementos del mismo tipo de datos. Son especialmente útiles para operaciones matemáticas avanzadas y manipulaciones eficientes en grandes conjuntos de datos. NumPy es ampliamente utilizado en la comunidad científica y matemática basada en Python. (McKinney, 2013)

### Pandas

La biblioteca pandas combina la eficiencia de cálculos de listas de NumPy con las habilidades de manipulación de datos presentes en hojas de cálculo y bases de datos relacionales, como SQL. Proporciona funcionalidades de indexado que facilitan la redimensión, segmentación, agregación y selección de subconjuntos de datos. Dado que la manipulación, preparación y limpieza de datos son habilidades cruciales en el análisis de datos, pandas se convierte en una herramienta esencial. (McKinney, 2013)

### Matplotlib

Matplotlib es la biblioteca de Python más reconocida para generar gráficos y visualizaciones de datos bidimensionales. Inicialmente creada por John D. Hunter, cuenta con un equipo activo de desarrolladores que la mantienen. Su diseño se orientó hacia la creación de gráficos aptos para su publicación. Aunque existen otras bibliotecas de visualización para programadores Python, Matplotlib sigue siendo ampliamente utilizada y se integra de manera adecuada con el resto del ecosistema. En mi opinión, es una elección sólida como herramienta de visualización por defecto (McKinney, 2013).

### Aprendizaje automático con Python

### Scikit-learn

Scikit-learn es una destacada biblioteca de aprendizaje automático diseñada para programadores de Python, que ha evolucionado desde su inicio en 2007 hasta convertirse en la principal herramienta de uso general en este campo. Con la contribución de más de 2000 personas, scikit-learn abarca diversos submódulos para modelos de clasificación, regresión, agrupamiento, reducción de dimensionalidad, selección de modelos y preprocesamiento de datos. Ofrece implementaciones de algoritmos como SVM, vecinos más cercanos, bosque aleatorio, regresión logística, entre otros. Scikit-learn, junto con otras librerías como pandas, ha sido esencial para posicionar a Python como un lenguaje productivo en la ciencia de datos (McKinney, 2013).

### Métricas de evaluación con Python

#### Exactitud(Accuracy)

La exactitud mide el porcentaje de casos que el modelo ha acertado. Se calcula como la proporción entre las predicciones correctas y el número total de predicciones realizadas. Además, es la métrica por defecto en los algoritmos de Python, utilizando su biblioteca SciKit-Learn, y puede ser evaluada mediante la función score(). También está disponible la función accuracy\_score() del submódulo metrics para obtener la exactitud en problemas de clasificación (Pineda Pertuz, 2022, p-165).

#### Precisión

La precisión es una métrica que mide la relación entre las predicciones clasificadas correctamente como positivas y el número total de predicciones correctas, tanto positivas como negativas. En Python, puedes calcular la precisión utilizando la librería SciKit-Learn, empleando la función precision\_score() del submódulo metrics. Esta métrica es esencial para evaluar la capacidad del modelo de clasificar correctamente las instancias positivas, proporcionando una perspectiva valiosa sobre la calidad de las predicciones afirmativas realizadas por el modelo (Pineda Pertuz, 2022, p-166).

#### Sensibilidad (recall)

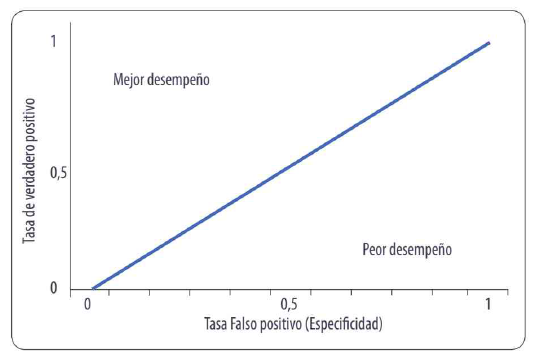
La sensibilidad, también conocida como recall, corresponde al número de elementos identificados correctamente como positivos respecto al total de positivos verdaderos. En Python, puedes calcular la sensibilidad utilizando la librería SciKit-Learn, haciendo uso de la función recall\_score() del submódulo metrics. Esta métrica es crucial para evaluar la capacidad del modelo de identificar de manera efectiva las instancias positivas, proporcionando información valiosa sobre la sensibilidad del algoritmo en la detección de casos afirmativos (Pineda Pertuz, 2022, p-166).

#### F1-score

La métrica de F1-score combina las medidas de precisión y sensibilidad en un solo valor, lo cual puede ser especialmente útil en casos donde se busca evaluar de manera conjunta el rendimiento de estas métricas. En Python, puedes calcular el F1-score utilizando la librería SciKit-Learn, mediante la función f1\_score() del submódulo metrics. Esta métrica proporciona una evaluación balanceada entre la capacidad del modelo para realizar predicciones precisas y su habilidad para identificar correctamente las instancias positivas, ofreciendo una perspectiva integral del desempeño del algoritmo (Pineda Pertuz, 2022, p-167).

#### Curvas Roc

Las Curvas Características Operativas del Receptor (ROC) son una herramienta valiosa para comparar distintos clasificadores a partir de los puntajes de sus predicciones. Estos puntajes pueden interpretarse como probabilidades, variando entre 0 y 1. La estructura de la curva ROC se visualiza en la figura 2-1



###### Figura 2-1: Gráfico de Roc

**Fuente: Pineda Pertuz, (2022)**

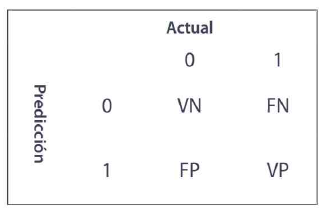
* El eje x representa el incremento de la tasa de falsos positivos.
* El eje y representa la tasa de verdaderos positivos.

En general, las curvas por debajo de cierto umbral tienen un rendimiento más bajo, mientras que las que se encuentran por encima exhiben un mejor desempeño. El objetivo es crear modelos cuyo rendimiento se sitúe dentro de los segmentos [0,0] - [0,1] y [0,1] - [1,1].

En Python, puedes generar curvas ROC utilizando la librería SciKit-Learn, específicamente mediante la función roc\_curve() del submódulo metrics. Este enfoque visual facilita la comparación entre modelos y permite tomar decisiones informadas sobre qué clasificador puede ser más adecuado para tu problema (Pineda Pertuz, 2022, p-168).

#### Matriz de confusión

Una matriz de confusión es una herramienta visual utilizada para evaluar el rendimiento de un modelo de aprendizaje supervisado. En Python, la librería SciKit-Learn proporciona una manera efectiva de generar esta matriz mediante el método confusion\_matrix() del submódulo metrics. Esta herramienta permite describir la distribución de los valores verdaderos y las predicciones, mostrando la cantidad de aciertos y errores en un problema de clasificación. Es especialmente útil en el contexto de una clasificación binaria, donde se asigna el valor 0 a la clase negativa y 1 a la clase positiva. La matriz de confusión presenta una representación detallada en la figura 2-2



###### Figura 2-2: Gráfica Matriz de confusión

**Fuente: Pineda Pertuz, (2022)**

* Verdaderos Positivos (VP): El modelo predijo correctamente la salida como positiva.
* Falsos Negativos (FN): El modelo predijo incorrectamente la salida como negativa.
* Falsos Positivos (FP): El modelo predijo incorrectamente la salida como positiva.
* Verdaderos Negativos (VN): El modelo predijo correctamente la salida como negativa.

Estos elementos permiten una evaluación más completa del desempeño del algoritmo de clasificación al analizar tanto los aciertos como los errores en la predicción de cada clase (Pineda Pertuz, 2022, p-164).

#### Validación cruzada por k iteraciones

La validación cruzada por ‘k’ iteraciones es un procedimiento esencial en el aprendizaje automático que implica dividir aleatoriamente un conjunto de datos en k porciones o pliegues. En cada iteración, k-1 de estos pliegues se utilizan para entrenar el modelo, mientras que el pliegue restante se reserva para realizar la prueba del modelo. Este proceso se repite ‘k’ veces, garantizando que cada pliegue haya actuado tanto como conjunto de entrenamiento como conjunto de prueba.

Esta técnica, implementada con el método “StratifiedKFold” en la biblioteca de Python, es particularmente efectiva para conjuntos de datos desbalanceados, ya que mantiene la proporción de clases en cada pliegue. El número de pliegues generados en este método es igual a la cantidad de iteraciones establecidas, lo que permite una exhaustiva evaluación del modelo en diversas configuraciones. Además, esta repetición sistemática brinda una estimación robusta del rendimiento del modelo, ya que cada instancia de datos tiene la oportunidad de estar en ambos conjuntos de entrenamiento y prueba a lo largo de las iteraciones (Pineda Pertuz, 2022, p-125).

#### Hiperparámetros y búsqueda en cuadrícula en Aprendizaje Automático

Este proceso de ajuste de hiperparámetros los cuales son configuraciones externas al modelo, como la tasa de aprendizaje, número de épocas, tamaño del lote y arquitectura, que deben ajustarse antes del entrenamiento. Estos parámetros influyen en el rendimiento y comportamiento del modelo y su elección adecuada es esencial para lograr un rendimiento óptimo. La búsqueda en cuadrícula, realizada por GridSearchCV, explora exhaustivamente el espacio de hiperparámetros, evaluando cada combinación posible. Aunque este método es completo, puede volverse computacionalmente costoso en espacios de búsqueda extensos.

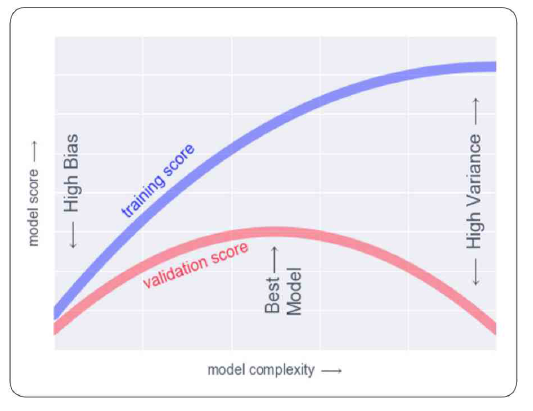
Por otro lado, RandomizedSearchCV ofrece una alternativa eficiente al seleccionar aleatoriamente un conjunto específico de combinaciones de hiperparámetros para evaluar. Esta estrategia es particularmente beneficiosa cuando el espacio de búsqueda es amplio y explorar todas las combinaciones sería impráctico.

Ambos enfoques, Grid Search y Randomized Search, forman parte de las herramientas esenciales que proporciona Scikit-learn para optimizar el rendimiento de modelos de aprendizaje automático. La elección entre estos métodos dependerá del tamaño del espacio de búsqueda y los recursos computacionales disponibles, permitiendo así una adaptabilidad efectiva a las necesidades específicas de ajuste de hiperparámetros en cada proyecto (Pineda Pertuz, 2022, p-134).

#### Curva de aprendizaje

En machine learning, uno de los desafíos cruciales es encontrar un punto óptimo que equilibre el sesgo y la varianza. Un sesgo elevado indica un problema de subajuste, lo que significa que el modelo no puede ajustarse adecuadamente a los datos. Por otro lado, una varianza alta es sinónimo de sobreajuste, donde el modelo se ajusta demasiado a los datos de entrenamiento y tiene dificultades para generalizar con nuevos datos.

En relación a la complejidad del modelo, observamos que un modelo complejo o sobre ajustado puede tener una alta precisión al determinarse con los datos de entrenamiento, pero esta precisión disminuye al evaluarlo con datos de validación o prueba. Además, se destaca que el punto más alto en la curva de validación es donde se encuentra el modelo óptimo ver figura 2-3. Este punto representa un equilibrio adecuado entre sesgo y varianza, proporcionando un modelo que generaliza bien a nuevos datos y evita tanto el subajuste como el sobreajuste (Pineda Pertuz, 2022, p-127).



###### Figura 2-3: Gráfica Curva de aprendizaje

**Fuente: Pineda Pertuz, (2022)**

## 

# Metodología.

El marco metodológico de este proyecto define cómo abordaremos el análisis predictivo del abandono de clientes en una empresa de alojamiento web. Presentaremos las estrategias para recopilar, entendimiento de los datos, procesamiento de los datos y diseño de modelos de Machine Learning. Esta metodología es esencial para alcanzar nuestros objetivos.

## Área de estudio

El enfoque principal de nuestra área de estudio se centra en el Estado Plurinacional de Bolivia, apoyándonos en información proporcionada por la empresa “Boliviahost” que abarca el territorio boliviano. Bolivia se sitúa en el centro de América del Sur, entre los 57º26′ y 69º38′ de longitud occidental del meridiano de Greenwich y los paralelos 9º38′ y 22º53′ de latitud sur (Ine, 2020).

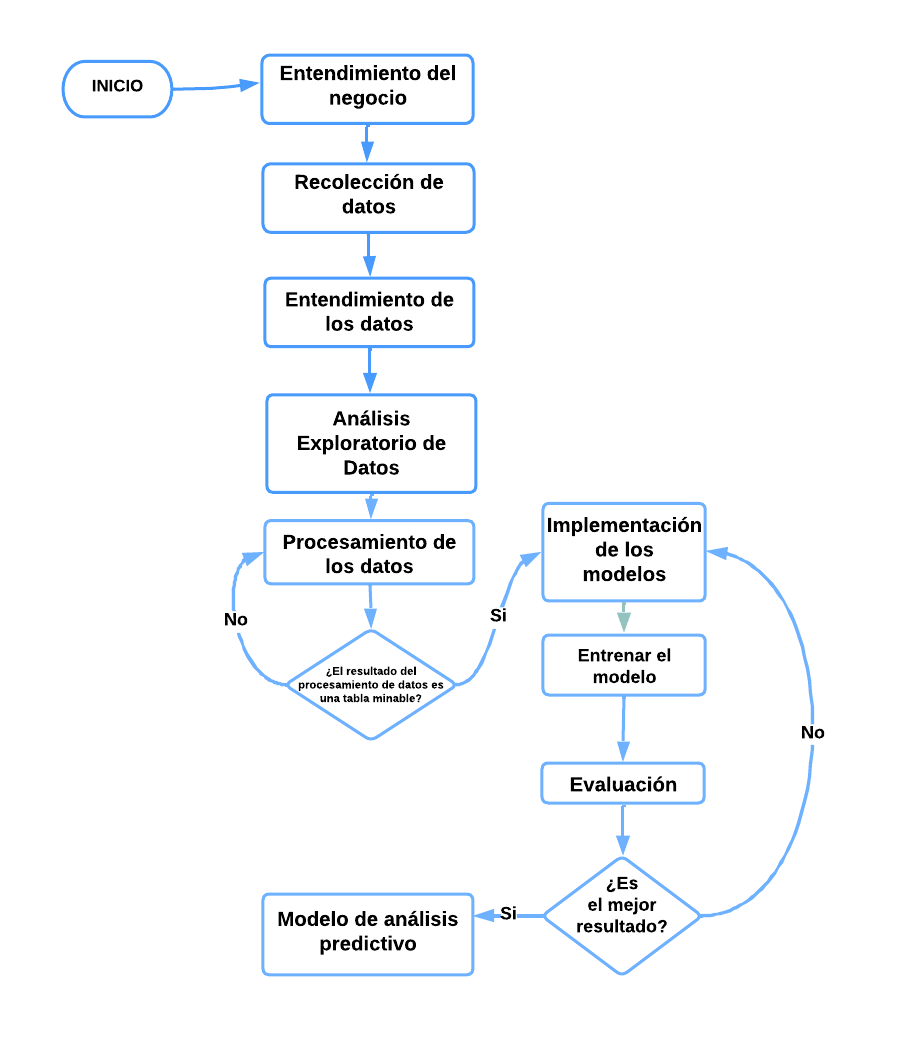
Para una visualización precisa de la ubicación, consulte la Figura 3-1, que proporciona una representación gráfica de la geolocalización de las áreas de interés en este proyecto de investigación.

###### Figura 3-1: Área de estudio

**Fuente: Elaboración propia, 2023**

## Flujograma metodológico

El desarrollo del siguiente proyecto sigue los siguientes pasos como se muestra en la figura 3-2.



###### Figura 3-2: Flujograma metodológico

**Fuente: Elaboración propia en función a CRISP-DM,(2023)**

## Fuentes de información

### Fuentes de información Secundaria

El conjunto de datos pertenece a la base de datos Whmcs(Web Host Manager Complete Solution) de la empresa “Boliviahost” para el período de 2017 a 2023. Estos datos se encuentran en formato “Sql-script” (Anexo 1), el cual cuenta con un total de seis tablas que contiene información de los parámetros de servicio, precio del servicio, tipos de suscripción, forma de pago, estado del servicio, duración del cliente con el servicio en meses así también como datos demográficos con relación al servicios ver tabla 3-1.

## Entendimiento del negocio

En esta fase, se busca comprender el negocio, sus objetivos y se establecen las bases fundamentales para el desarrollo del proyecto.

### Objetivo del negocio

* Mejorar la retención de clientes.
* Reducir la tasa de abandono en los servicios .

### Situación Actual

La empresa “BoliviaHost” brinda servicios de alojamiento web de manera remota, no cuenta con oficinas en Bolivia. Su alcance abarca tanto clientes nacionales como internacionales. Actualmente, cuenta con un equipo de 2 profesionales en el área de finanzas, 1 persona en el área de ventas, 5 profesionales en el área de servicio y 1 experto en soporte técnico.

### Recursos y Restricciones

* **Recursos**: Base de datos histórica de clientes, software especializado para el analizar y extracción de datos histórica de clientes.
* **Restricciones:** Limitaciones en la cantidad de datos.

### Plan de Proyecto Inicial

* Fase 1: Análisis exploratorio de datos.
* Fase 2: Desarrollo de modelos predictivos.
* Fase 3: Evaluación de modelos.
* Fase 4: Implementación de estrategias de retención.

## Recopilación de datos

La recopilación de datos se llevó a cabo a través de la página web "Boliviahost.com", donde los propios usuarios ingresan información mediante formularios específicos disponibles en el sitio. Estos formularios han sido diseñados para recopilar datos relevantes necesarios para ofrecer el servicio de alojamiento web. Es importante destacar que los datos recopilados son de naturaleza privada y confidencial. Se subraya que la recepción de estos datos se realizó mediante un acuerdo previo con la empresa, estableciendo las condiciones y términos de seguridad.

## Entendimiento de los datos

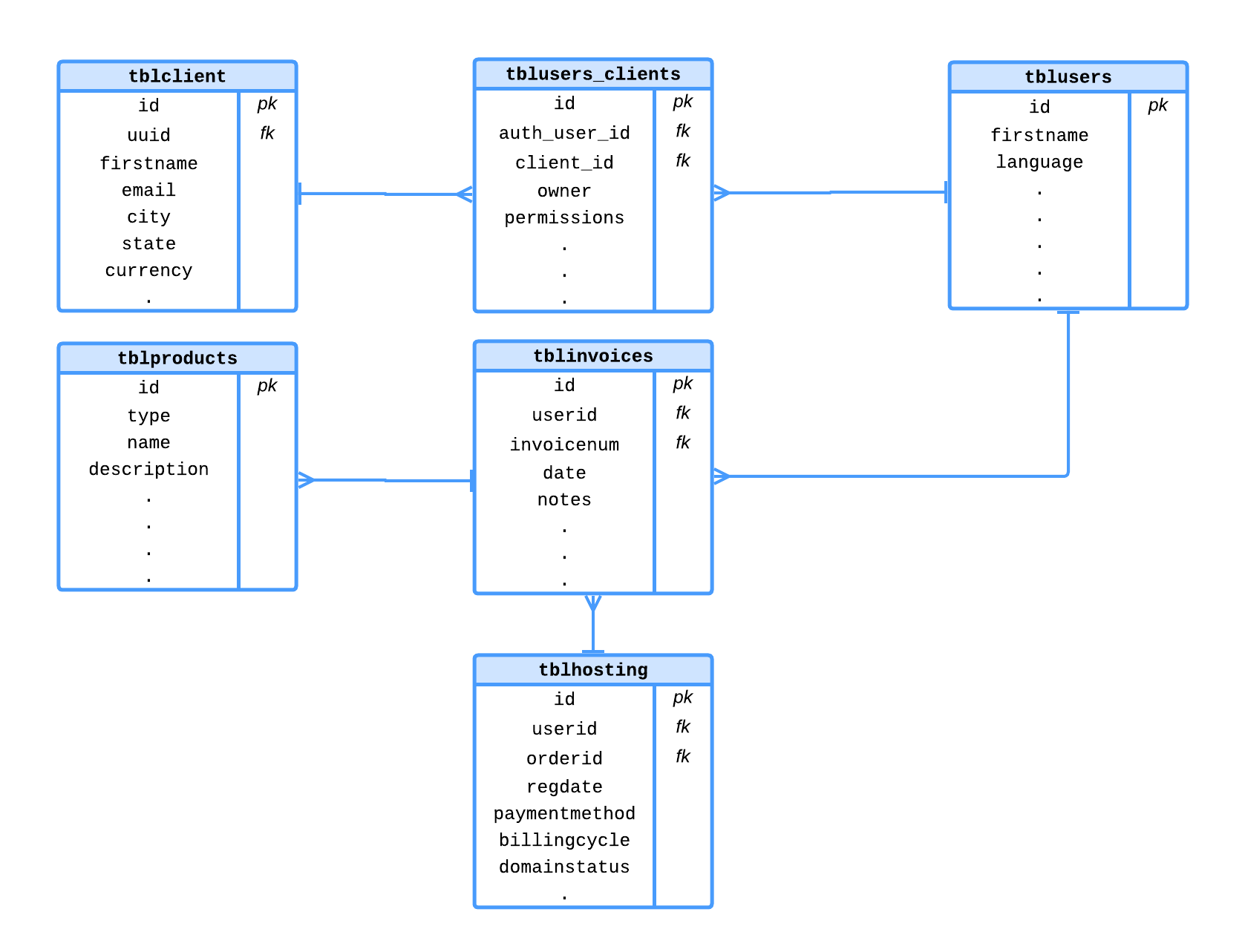
En esta etapa, se llevó a cabo un análisis exhaustivo de los datos disponibles, utilizando herramientas de administración de bases de datos para comprender su estructura y relaciones.

En una primera fase, se procedió a la identificación de las tablas que albergan la información pertinente para el proyecto. Posteriormente, se llevó a cabo un análisis detallado de la estructura de cada tabla, cuyos resultados pueden observarse en la Tabla 3-1. Para una visualización más clara de las relaciones entre las distintas tablas, se proporciona una representación gráfica en la Figura 3-3.

| Nombre de tabla | Descripción |
| --- | --- |
| tblclients | Tabla el cual almacena a los clientes |
| tblusers | Tabla que almacena a los usuarios |
| tblusers\_clients | Tabla que almacena la relación entre el cliente y el usuario |
| tblproduct | Tabla que almacena los servicios |
| tblhosting | Tabla que almacena la relación entre el servicio y el cliente |
| tblinvoices | tabla que almacena la factura del cliente |

##### Tabla 3-1: Descripción de las tablas de la base de datos

**Fuente: Elaboración propia,(2023)**



###### Figura 3-3: Modelo entidad relación

**Fuente: Elaboración propia, (2023)**

Finalmente se extrajeron los datos más relevantes para el proyecto. Para ello, se utilizó una consulta SQL para unir todas las tablas relevantes. La consulta se detalla en el Anexo 2. Esta tabla resultante consta de 1,849 registros, contiene los campos especificados en la tabla 3-2.

| Nombre de tabla | Descripción | Tipo |
| --- | --- | --- |
| client\_id | identificador del cliente | integer |
| service\_id | identificador del servicio | integer |
| name\_services | nombre del servicio | String |
| regdate | fecha de adquisición del servicio | date |
| active\_months | cantidad de meses con el servicio | integer |
| domain\_status | estado del servicio | String |
| paymentmethod | método de pago | String |
| billingcycle | frecuencia de pago | String |
| total | precio del servicio | Double |
| state | ciudad | String |

##### Tabla 3-2: Descripción de los campos de la tabla resultado de la ejecución de la consulta SQL

**Fuente: Elaboración propia,(2023)**

## Análisis Exploratorio de Datos

En esta etapa, se llevó a cabo un análisis exhaustivo con el propósito detectar tendencias, patrones y correlación de los datos inicialmente se identifican los tipos de variables ver tabla 3-3. (Anexo 3)

| Nombre de Variable | Tipo | Valor |
| --- | --- | --- |
| total | Cuantitativo | continuo |
| active\_months | Cuantitativo | discreto |
| name\_services | Cualitativa | nominal |
| domain\_status | Cualitativa | nominal |
| paymentmethod | Cualitativa | nominal |
| billingcycle | Cualitativa | nominal |
| state | Cualitativa | nominal |

##### Tabla 3-3: Descripción de los tipos de variable

**Fuente: Elaboración propia, (2023)**

### Descripción Estadística:

La descripción estadística es una técnica que resume y presenta de manera concisa las características fundamentales de un conjunto de datos.

Recuento: Indica que se tiene 1849 observaciones en ambas columnas. Esto significa que no hay valores faltantes (NaN) en estas columnas.

Media: La media de “meses\_activo” es aproximadamente 20, lo que sugiere que, en promedio, la duración de la cantidad de meses activos es de alrededor de 20 meses. Para la columna total, la media es aproximadamente 318.86 bs.

Mediana: La mediana (o percentil 50) de “meses\_activo” es 12 meses. Esto indica que la mitad de las observaciones tienen 12 meses o menos de actividad. Para la columna total, la mediana es 299 bs.

Valor mínimo: Muestra el valor mínimo en ambas columnas. La duración mínima con el servicio es 1 mes, y el valor mínimo de total es 99 bs.

Valor máximo: Muestra el valor máximo en ambas columnas. La duración máxima con el servicio es 73 meses, y el valor máximo total es 620 bs. Esto indica el rango de valores en tus datos.

Moda: La moda es 12 meses para “meses\_activo”, lo que significa que este valor es el más frecuente. La moda para la columna total es 299 bs.

Rango: El rango, que es la diferencia entre el valor máximo y el mínimo, es de 72 meses para “meses\_activo” y 547 bs para total.

Rango Intercuartil (IQR): El IQR, que representa la dispersión de la mitad central de los datos, es de 23 meses para “meses\_activo” y 134 bs para total.

Primer percentil (25%): El 25% de las observaciones tienen 1 mes o menos de actividad para “meses\_activo” y 225 bs o menos para total.

Tercer percentil (75%): El 75% tienen 24 meses o menos para “meses\_activo” y 399 bs o menos para total.

Varianza: La varianza de “meses\_activo” es 531.09 y para total es 23430.43. Esto mide cuánto varían las observaciones respecto a la media.

Desviación Estándar: La desviación estándar de “meses\_activo” es aproximadamente 23.04 meses, y para total es aproximadamente 153.07 bs, ver la tabla 3-4.

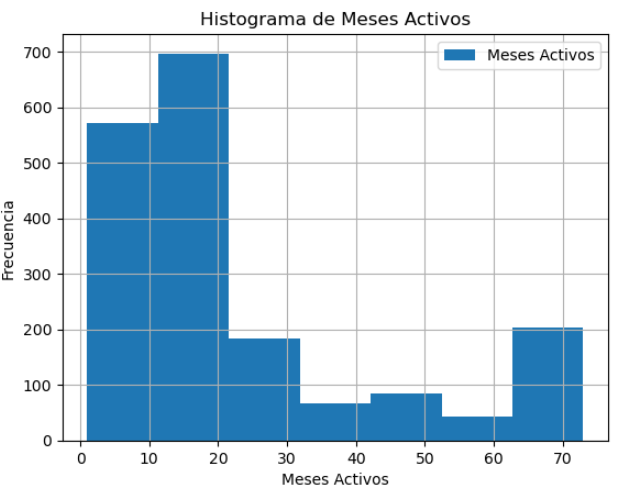
| Nombre de Variable | Meses\_activo | total |
| --- | --- | --- |
| Recuento | 1849 | 1849 |
| Media | 20.46 | 318.86 |
| Mediana | 12.0 | 299.0 |
| Valor mínimo | 1 | 99 |
| Valor máximo | 73 | 620 |
| Moda | 12 | 299 |
| Rango | 72.0 | 547 |
| Rango Intercuartil (IQR) | 23 | 134 |
| Primer percentiles 25% | 1 | 225 |
| Segundo percentiles 50% | 12 | 299 |
| Tercer percentiles 75% | 24 | 399 |
| Varianza | 531.09 | 23430.43 |
| Desviación Estándar | 23.04 | 153.07 |

##### Tabla 3-4: Descripción Estadística

**Fuente: Elaboración propia, (2023)**

#### Análisis Univariable

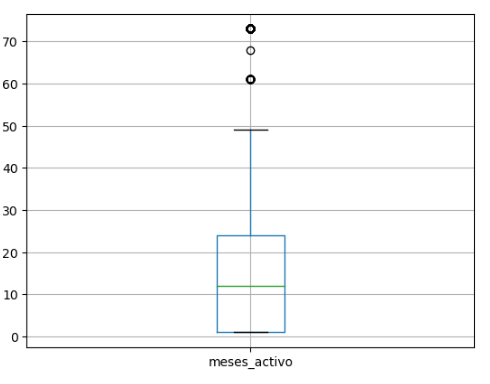
A través del histograma de la variable 'meses\_activos', se observa que la mayoría de los usuarios tienen una antigüedad de 1 a 20 meses. Sin embargo, a partir de los 20 meses, se observa un descenso en la cantidad de usuarios con mayor antigüedad. Este patrón sugiere que hay una concentración significativa de usuarios en los primeros meses, y a medida que aumenta la antigüedad, la cantidad de usuarios tiende a disminuir como se muestra en la figura 3-4.



###### Figura 3-4: Histogramas para la variable ‘meses\_activo’

**Fuente: Elaboración propia, (2023)**

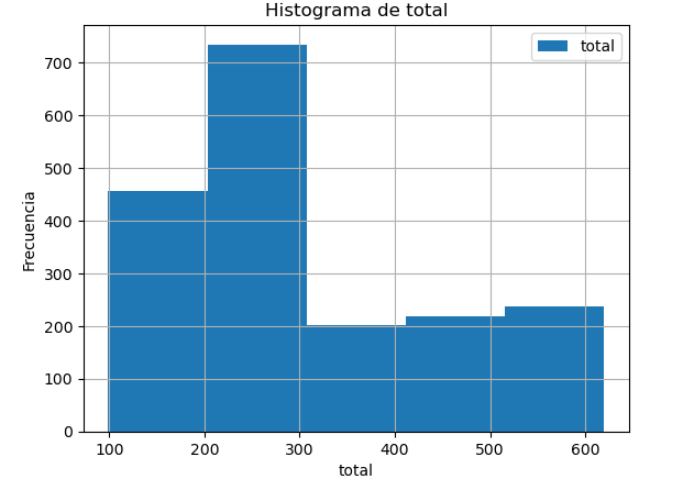
En la gráfica boxplot de la variable 'meses\_activos' se visualiza que la caja está más abajo porque los datos están sesgados hacia la izquierda, es decir, hay más valores pequeños que grandes. Esto significa que la mayoría de los individuos tienen pocos meses activos, pero hay algunos que tienen muchos meses activos y que existen algunos valores atípicos, como se muestra en la figura 3-5.



###### Figura 3-5: Gráfico de boxplot para las variables ‘meses\_activo’

**Fuente: Elaboración propia, (2023)**

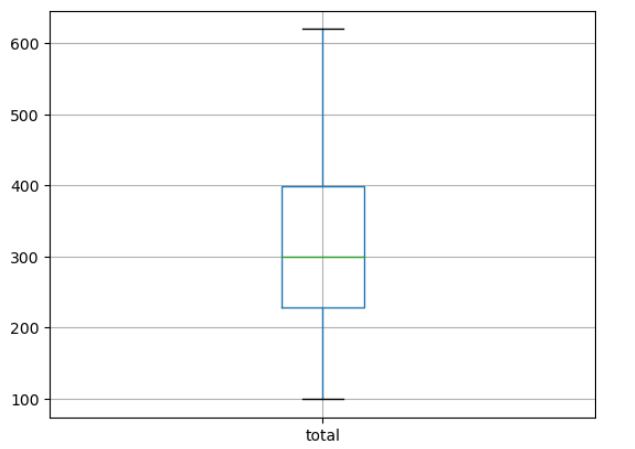
Mediante la visualización del histograma para la variable "Total", se aprecia que la gran mayoría de usuarios muestra una preferencia por servicios cuyos precios oscilan entre 100 y 300 bs. Por el contrario, muy pocos clientes pagan más de 300 bs por el servicio, ya que esta es la categoría con menor frecuencia. Esto sugiere que el servicio tiene una demanda elástica, es decir, que la cantidad demandada varía mucho según el precio como se muestra en la figura 3-6.



###### Figura 3-6: Histograma para las variables “Total”

**Fuente: Elaboración propia, (2023)**

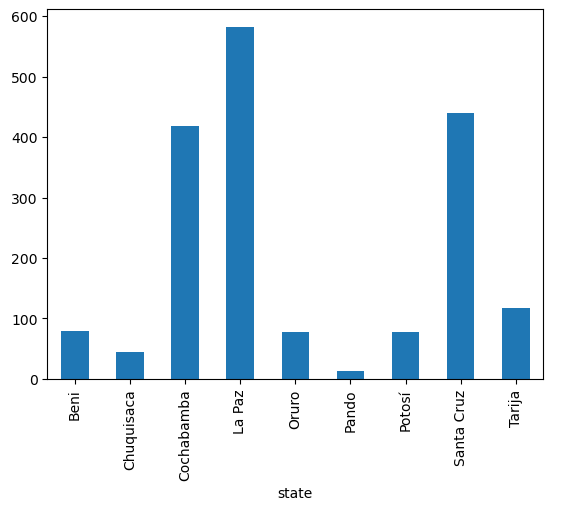
En la gráfica boxplot de la variable “total” se visualiza que el primer cuartil es aproximadamente 200 bs y el tercer cuartil es aproximadamente 400 bs. Esto significa que el 50% de los clientes pagaron entre 200 y 400 bs por el servicio. El valor mínimo es aproximadamente 100 bs y el valor máximo es aproximadamente 600 bs. No hay valores atípicos en este gráfico como se muestra en la figura 3-7.



###### Figura 3-7: Gráfico de boxplot para las variables “Total”

**Fuente: Elaboración propia, (2023)**

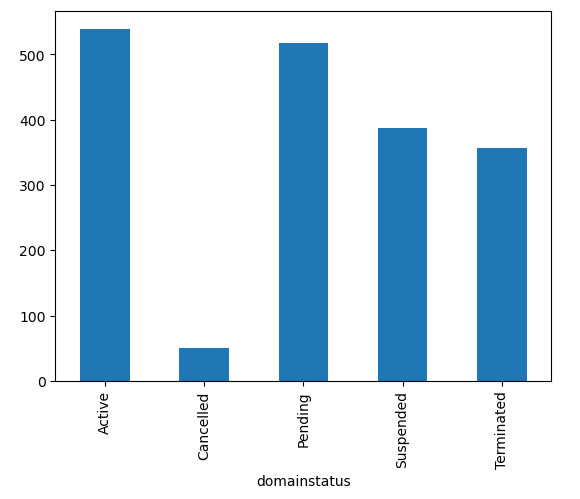
Las ciudades que tienen mayoro cantidad de usuarios son La Paz, Cochabamba y Santa Cruz. Por lo contrario la ciudad que cuenta con menos usuarios es la ciudad de Pando como se muestra en la figura 3-8.

****

###### Figura 3-8: Gráfica de barras para la variable “state”

**Fuente: Elaboración propia,(2023)**

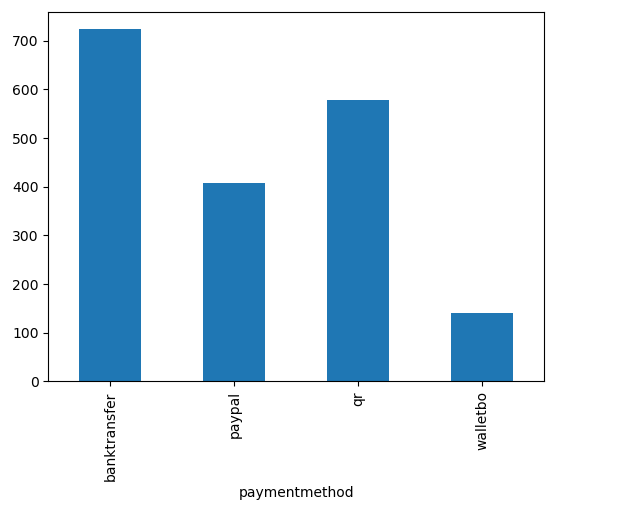
En el estado de las cuentas se puede observar que hay cinco estados posibles: activo, cancelado, pendiente, suspendido y terminado. Cada barra representa el número de cuentas que se encuentran en cada estado. La barra más alta corresponde al estado “active”, las cuentas que se encuentran dadas de bajas son representadas por los estado “cancelled”, “suspended” y “terminated” como se muestra en la figura 3-9.



###### Figura 3-9: Gráfica de barras para la variable “domainstatus”

**Fuente: Elaboración propia, (2023)**

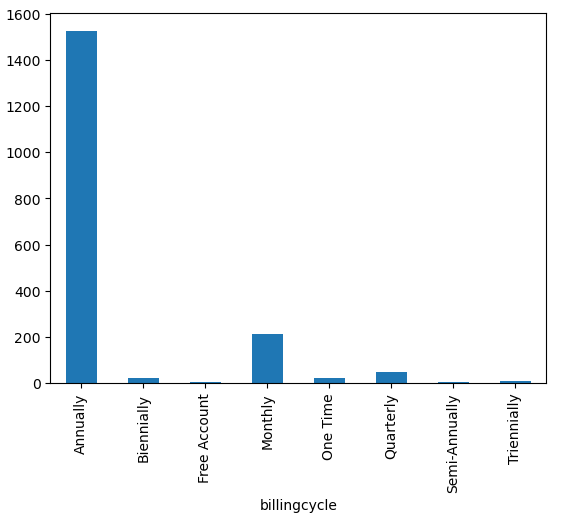
Por parte de las formas de pago se puede observar que los clientes prefieren los métodos de pago “banktransfer”, “paypal” y “qr” y el método de pago que menos usado es el “walletbo” como se muestra en la figura 3-10.



###### Figura 3-10: Gráfica de barras para la variable “domainstatus”

**Fuente: Elaboración propia, (2023)**

La forma de pago más preferida por los usuarios es de manera "anual", y esto se destaca significativamente, como se evidencia en la figura 3-11.

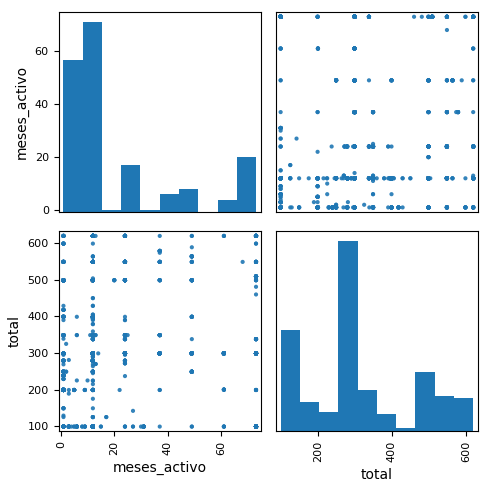


###### Figura 3-11: Gráfica de barras para la variable “billingcycle”

**Fuente: Elaboración propia, (2023)**

#### Análisis Bivariado

La visualización bivariada revela que no hay una correlación lineal entre las variables cuantitativas "Total" y "Meses\_Activo". Asimismo, se observa un sesgo hacia la derecha en la distribución de la variable "Meses\_Activo" como se muestra en la figura 3-12.



###### Figura 3-12: Gráfica bivariado entre variables cuantitativas

**Fuente: Elaboración propia, (2023)**

#### Correlación entre variables cuantitativas

El análisis de correlación “pearson” muestra que hay una relación positiva débil entre las variables "meses\_activo" y "total". El coeficiente de correlación de 0.099 indica una asociación positiva, pero la fuerza de esta asociación es baja. Esto sugiere que, en general, a medida que la variable "meses\_activo" aumenta, la variable "total" tiende a aumentar también, pero la conexión entre ambas no es muy fuerte como se visualiza en la figura 3-13.



###### Figura 3-13: Gráfico de correlación entre las variables ‘total’ y ‘meses\_activo’

**Fuente: Elaboración propia, (2023)**

#### Prueba Chi-cuadrado para Variables Cualitativas

La prueba de Chi-cuadrado se emplea para evaluar la asociación entre dos variables cualitativas (name\_services, paymentmethod, billingcycle, state) con respecto a la variable objetivo "domainstatus".

* Hipótesis Nula (H0): No hay asociación entre las dos variables.
* Hipótesis Alternativa (H1): Existe una asociación entre las dos variables.

| **Nombre de Variable** | **Estadístico Chi-cuadrado** | **P-valor** | **Grados de libertad** |
| --- | --- | --- | --- |
| name\_services | 1570.658873732699 | 1.0053250837863816e-155 | 344 |
| paymentmethod | 67.72080537056128 | 8.528459554454841e-10 | 12 |
| billingcycle | 234.4576660578237 | 1.1642153736640864e-32 | 32 |
| state | 202.9966405545492 | 1.855008243249217e-28 | 28 |

##### Tabla 3-6: Prueba Chi-cuadrado para Variables Cualitativas

**Fuente: Elaboración propia, (2023)**

El análisis de chi-cuadrado revela una asociación altamente significativa entre la variable categórica “name\_services” y la variable de interés. Con un estadístico chi-cuadrado de 1570.66 y un p-valor prácticamente nulo 1.01e-155 para 344 grados de libertad, se confirma que las distribuciones de name\_services difieren sustancialmente entre las categorías. Este hallazgo sugiere que el tipo de servicio tiene un impacto significativo en la variable de interés.

La prueba de chi-cuadrado indica una conexión estadísticamente significativa entre la variable categórica “paymentmethod” y la variable de interés. El estadístico chi-cuadrado de 67.72 con un p-valor de 8.53e-10 para 12 grados de libertad revela diferencias notables en las distribuciones de métodos de pago entre las categorías. Este resultado sugiere que la elección del método de pago podría tener implicaciones importantes en la variable de interés.

El análisis chi-cuadrado demuestra una asociación sustancial entre la variable categórica “billingcycle” y la variable de interés. Con un estadístico chi-cuadrado de 234.46 y un p-valor extremadamente bajo 1.16e-32 para 32 grados de libertad, se confirma que las distribuciones de “billingcycle” difieren significativamente entre las categorías. Este descubrimiento sugiere que el ciclo de facturación juega un papel significativo en la variable de interés.

El examen chi-cuadrado revela una conexión estadísticamente significativa entre la variable categórica state y la variable de interés. Con un estadístico chi-cuadrado de 202.99 y un p-valor de 1.86e-28 para 28 grados de libertad, se evidencia que las distribuciones de state difieren notablemente entre las categorías. Este resultado implica que la ubicación geográfica, representada por el estado, podría tener un impacto importante en la variable de interés ver tabla 3-6.

## Procesamiento de los datos

### Tratamiento de valores nulos

Los valores nulos se detectaron utilizando funciones de Python en las variables tanto cualitativas como cuantitativas, y se procedió a realizar la correspondiente transformación de estos valores nulos en valores adecuados.

### Tratamiento de valores atípicos

Se reemplazan los valores atípicos por valores más cercanos al percentil 95 o 99 puede contribuir a suavizar el impacto de los valores extremos en los datos, proporcionando así una representación más robusta y equilibrada de la distribución de las observaciones.

### Transformación de variables cualitativas

Las variables cualitativas, a excepción de la variable etiqueta, se transformaron utilizando la técnica de codificación one-hot (one hot encoding). Esta técnica consiste en convertir las cualitativas de una variable cualitativa en un conjunto de nuevas variables binarias (0 o 1), donde cada variable binaria representa la presencia o ausencia de una cualitativa específica.

## Desarrollo de Modelo

### Análisis del Problema

Por la naturaleza del problema de abandono de clientes en el servicio de alojamiento web, identificando los requisitos específicos. Se determina que los modelos de clasificación son apropiados para predecir el abandono.

### Exploración de Algoritmos

Se opta por seleccionar una variedad de algoritmos que abarcan modelos lineales, de árboles, probabilísticos y de instancias. Entre las opciones seleccionadas se encuentran Arbol de decisiones, Bosque aleatorio, knn, Regresión Logística y Naive Bayes. Esta diversidad de modelos se elige para abordar distintos aspectos del problema de abandono de clientes en el servicio de alojamiento web (Anexo 4).

### Implementación Inicial

En este código, representa la estructura común para el análisis de modelos de aprendizaje automático del presente proyecto. La estructura se divide en 7 secciones:

**Sección 1**: Importación de Bibliotecas

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

import matplotlib.pyplot as plt

import seaborn as sns

Se importan las bibliotecas necesarias, incluyendo pandas para manipulación de datos, Scikit-Learn para machine learning, y las bibliotecas de visualización matplotlib y seaborn.

**Sección 2:** Carga de datos

data = pd.read\_csv('data\_readyboliviahost.csv')

data = data.drop('client\_id', axis=1)

data = data.drop('regdate', axis=1)

La tabla se carga desde un archivo CSV en un DataFrame de pandas llamado 'data'. Luego, se eliminan las columnas 'client\_id' y 'regdate', que no se utilizarán en el análisis.

**Sección 3:** Preprocesamiento

X = data.drop('domain status', axis=1)

y = data['domain status']

Se separan los datos en dos conjuntos: X, que contiene todas las columnas excepto 'domainstatus' (la variable objetivo), y, que almacena 'domainstatus'. Esta división es esencial para el entrenamiento y evaluación de modelos de clasificación

**Sección 4:** Creación de conjuntos de entrenamientos y prueba

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

Los datos se dividen en conjuntos de entrenamiento (X\_train, y\_train) y prueba (X\_test, y\_test) mediante la función train\_test\_split de Scikit-Learn. El 20% de los datos se reserva como conjunto de prueba, con una semilla aleatoria para reproducibilidad.

**Sección 5:** Entrenamiento

naive\_bayes = GaussianNB()

naive\_bayes.fit(X\_train, y\_train)

Esta sección importa el clasificador de la librería sklearn. Luego, crea un objeto clasificador y lo entrena con los datos de entrenamiento.

**Sección 6:** Predicción

y\_pred = naive\_bayes.predict(X\_test)

Realizar predicciones en el conjunto de prueba (X\_test) utilizando el modelo entrenado y almacena las predicciones en 'y\_pred '.

**Sección 7:** Evaluación

Esta sección calcula las métricas de evaluación del modelo. En este caso, se calculan la precisión, el informe de clasificación y la matriz de confusión.

## Evaluación

Esta sección 7 del código, representa la estructura común para la evaluación de modelos de aprendizaje automático del presente proyecto y se subdivide en 4 secciones.

**Sección 1**: Cálculo de Precisión global

accuracy = accuracy\_score(y\_test, y\_pred)

La precisión del modelo es una medida que evalúa su rendimiento general. Esta métrica representa la fracción de predicciones correctas en comparación con el total de predicciones realizadas. Para calcular la precisión, empleamos la función “accuracy\_score” de Python.

**Sección 2**: Informe de Clasificación

report = classification\_report(y\_test, y\_pred, target\_names=naive\_bayes.classes\_)

El Informe de Clasificación, generado mediante la función “classification\_report” en Python, condensa métricas clave como precisión, recuperación, F1-score y soporte para cada clase del modelo.

**Sección 3**: Matriz de Confusión

confusion\_df = pd.DataFrame(confusion\_matrix(y\_test, y\_pred, labels=naive\_bayes.classes\_),index=naive\_bayes.classes\_,columns=naive\_bayes.classes\_)

El código genera una matriz de confusión para evaluar el rendimiento de un modelo de clasificación. Utiliza las predicciones del modelo (y\_pred) y las etiquetas reales de prueba (y\_test). La matriz de confusión es organizada en un DataFrame, facilitando la interpretación del rendimiento del modelo en términos de clasificaciones correctas e incorrectas para cada clase específica

**Sección 4**: Área bajo la Curva ROC (AUC-ROC)

auc\_roc\_scores = {}

for i in range(len(naive\_bayes.classes\_)):

auc\_roc\_scores[naive\_bayes.classes\_[i]] = roc\_auc\_score(y\_test == naive\_bayes.classes\_[i], y\_pred == naive\_bayes.classes\_[i])

El código calcula el Área bajo la Curva ROC (AUC-ROC) para cada clase en un modelo de clasificación. Utiliza la función “roc\_auc\_score” de Scikit-Learn para evaluar la capacidad del modelo para distinguir entre clases. El resultado es un diccionario “auc\_roc\_scores” que almacena las puntuaciones AUC-ROC para cada clase específica. Este análisis proporciona información sobre la capacidad del modelo para equilibrar la tasa de verdaderos positivos y falsos positivos en cada clase, evaluando así su sensibilidad y especificidad.

## Ajuste de Modelos

Los Procesos de Optimización de Modelos buscan encontrar la mejor configuración de ajustes hiper parámetros para maximizar la precisión de los algoritmos de machine learning. Utilizan estrategias como la búsqueda de cuadrícula y aleatoria, junto con la validación cruzada, para evaluar y ajustar continuamente el rendimiento del modelo. Este enfoque busca lograr un equilibrio óptimo entre ajuste y generalización para obtener modelos más precisos y eficientes.

### Ajuste del modelo Knn

En esta sección, se definen los hiper parámetros que se ajustarán durante la búsqueda. En este caso, para el modelo Knn, los hiper parámetros son:

param\_grid = {

'n\_neighbors': [3, 5, 7, 9],

'weights': ['uniform', 'distance'],

'algorithm':['auto','ball\_tree','kd\_tree','brute'],

'p': [1, 2]

}

Donde:

* **n\_neighbors**: Número de vecinos más cercanos.
* **weights**: Peso dado a los vecinos cercanos, siendo 'uniform' para pesos iguales y 'distance' para pesos inversamente proporcionales a la distancia.
* **algorithm**: Algoritmo utilizado para calcular los vecinos más cercanos, que puede ser 'auto', 'ball\_tree', 'kd\_tree' o 'brute'. Cuando se establece como 'auto', el algoritmo selecciona automáticamente el enfoque más adecuado según los datos de entrada. 'ball\_tree' y 'kd\_tree' utilizan estructuras de árbol para optimizar la búsqueda de vecinos, siendo más eficientes en conjuntos de datos grandes. Por otro lado, 'brute' realiza una búsqueda exhaustiva sin estructuras adicionales.
* **p**: Parámetro de potencia para la métrica de Minkowski. Cuando p = 1, se utiliza la distancia de Manhattan; cuando p = 2, se utiliza la distancia euclidiana.

### Ajuste del modelo Regresión Logística

Los hiper parámetros que se ajustarán para el modelo de Regresión Logística son:

param\_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],

'penalty': ['l1', 'l2']}

Donde:

* **C**: Inverso de la fuerza de regularización; valores más pequeños especifican una regularización más fuerte.
* **penalty**: Tipo de regularización, que puede ser 'l1' (regularización L1) o 'l2' (regularización L2).

### Ajuste del modelo Naive Bayes

Los hiper parámetros que se ajustarán para el modelo de Naive Bayes son:

param\_grid = {

'var\_smoothing': [1e-9, 1e-8, 1e-7, 1e-6, 1e-5]

}

**var\_smoothing**: Parámetro de suavizado, que controla la adición de una pequeña cantidad a las varianzas de las características para mejorar la estabilidad numérica.

### Ajuste del modelo Árbol de decisiones

Los hiper parámetros que se ajustarán para el modelo Árbol de decisiones son:

param\_dist = {

'criterion': ['gini', 'entropy'],

'max\_depth': [None] + list(range(5, 20)),

'min\_samples\_split': list(range(2, 11)),

'min\_samples\_leaf': list(range(1, 11)),

'min\_impurity\_decrease': [0.0] + [i/1000.0 for i in range(1, 11)]

}

Donde:

* **criterion**: especifica la función para medir la calidad de una división, que puede ser 'gini' o 'entropy'.En el contexto de árboles de decisión, 'gini' y 'entropy' son criterios para medir la pureza de una división. 'Gini' se centra en la probabilidad de error, mientras que 'entropy' se basa en la medida de la información. Ambos criterios buscan divisiones que minimicen la mezcla de clases en los nodos del árbol.
* **max\_depth**: representa la profundidad máxima del árbol.
* **min\_samples\_split**: establece el número mínimo de muestras requeridas para dividir un nodo interno.
* **min\_samples\_leaf**: define el número mínimo de muestras requeridas en cada hoja.
* **min\_impurity\_decrease**: Indica la cantidad mínima de disminución requerida en la impureza para realizar una división. La lista incluye valores desde 0.0 hasta 0.01 en incrementos de 0.001, permitiendo una sintonización fina de este hiperparametros

### Ajuste del modelo Bosque aleatorio

Los hiper parámetros que se ajustarán para el modelo Bosque aleatorio son:

param\_dist = {

'n\_estimators': [50, 100, 200, 300],

'max\_features': ['sqrt', 'log2', None],

'bootstrap': [True, False]

}

Donde:

* **n\_estimators**: Este hiperparámetro define el número de árboles en el bosque aleatorio. Cada árbol en el bosque se construye durante una iteración del algoritmo, y la predicción final se realiza agregando las predicciones de todos los árboles.
* **max\_features**: controla cuántas características se consideran al dividir un nodo en Bosques Aleatorios. Puede ser 'sqrt', 'log2', o None, representando la raíz cuadrada, el logaritmo base 2 del número total de características, o todas las características, respectivamente. Esta flexibilidad impacta la diversidad entre los árboles, mejorando la robustez del modelo.
* **bootstrap**: Indica si se debe realizar el muestreo con reemplazo (bootstrap) al construir los árboles. Si es True, cada árbol se construye utilizando una muestra aleatoria con reemplazo de los datos de entrenamiento. Si es False, se utiliza todo el conjunto de datos original para construir cada árbol.

## Técnicas de Optimización

Se emplea GridSearchCV para llevar a cabo una búsqueda exhaustiva de hiper parámetros para para los modelos NaiveBayes, Regresion Logistica y Knn, explorando todas las combinaciones definidas en "param\_grid" durante el ajuste de hiper parámetros. La métrica de rendimiento utilizada, en este caso, es 'accuracy'. Aquí, "model\_classifier" representa el modelo en cuestión, y el proceso evalúa y selecciona la combinación óptima de hiper parámetros para mejorar la precisión del modelo

grid\_search= GridSearchCV(model\_classifier, param\_grid, scoring='accuracy',cv=cv)

grid\_search.fit(X\_train, y\_train)

También se utilizó RandomizedSearchCV para los modelos de árbol de decisión y bosque aleatorio. Se especifica un espacio de búsqueda definido por el diccionario param\_dist, que contiene diferentes valores posibles para los hiperparámetros del árbol de decisión. La métrica de evaluación utilizada es la precisión, y el proceso se lleva a cabo mediante validación cruzada (cv) para evaluar el rendimiento del modelo. La búsqueda aleatoria explora 20 combinaciones diferentes de hiperparámetros, y el resultado es un ajuste del modelo utilizando los datos de entrenamiento (X\_train e y\_train). Este enfoque ayuda a encontrar la combinación óptima de hiperparámetros que maximiza la precisión del modelo en el conjunto de datos de entrenamiento.

random\_search=

RandomizedSearchCV(decision\_tree,

param\_distributions=param\_dist,

scoring='accurecy',cv=cv,

n\_iter=20,random\_state=42)

random\_search.fit(X\_train, y\_train)

## Validación Cruzada

La validación cruzada se realiza utilizando “StratifiedKFold” para todos los modelos de aprendizaje automático con 5 pliegues para garantizar que la distribución de las clases sea similar en cada pliegue.

cv = StratifiedKFold(n\_splits=5, shuffle=True, random\_state=42)

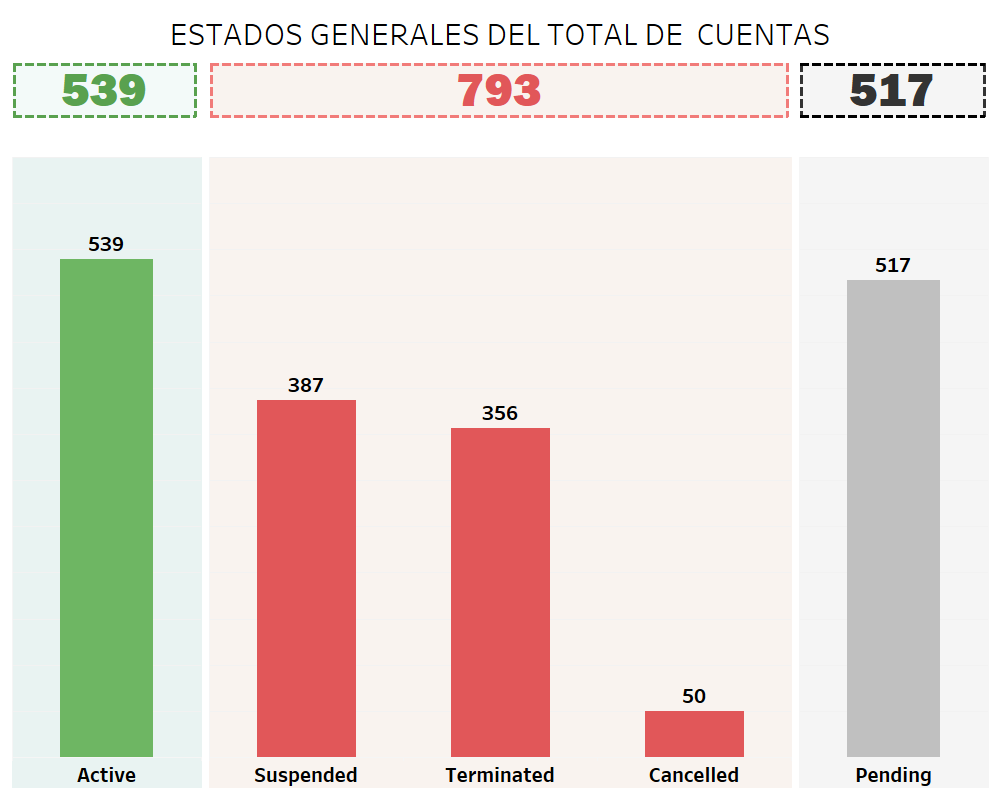
# 

# Resultados y Discusión

A continuación, se presenta el análisis del estudio, que incluye los datos más relevantes, así como los resultados del entrenamiento de los modelos. Se proporciona una breve descripción del trabajo.

## Análisis descriptivo de los datos

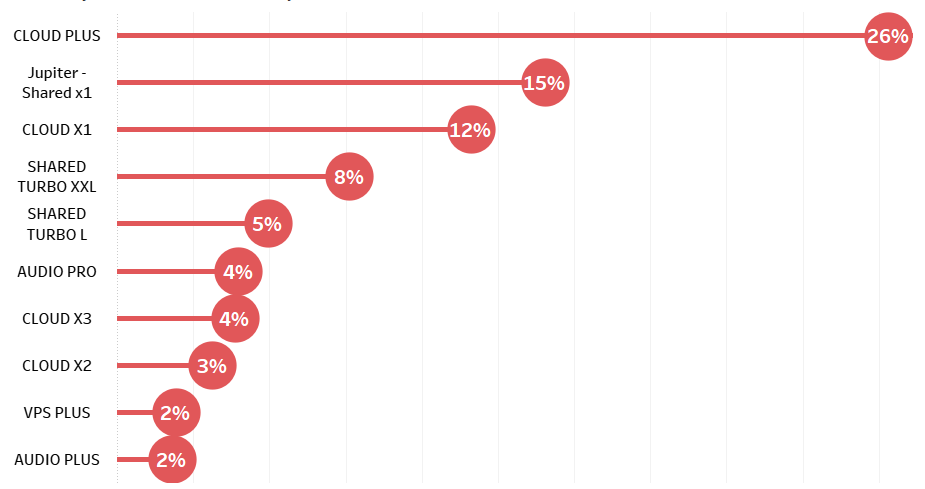
En el intervalo de tiempo que abarca desde 2017 hasta 2023, se ha contabilizado un total de 1,849 cuentas. De este conjunto, 793 cuentas han experimentado suspensiones o bajas, ya sea por motivos de falta de pago o retiro voluntario. Estas suspensiones constituyen aproximadamente el 43% del total de cuentas registradas, lo que se traduce en el porcentaje de churn. Este dato se presenta visualmente en la figura 4-1.



###### Figura 4-1: Estado de las cuentas

**Fuente: Elaboración propia, (2023)**

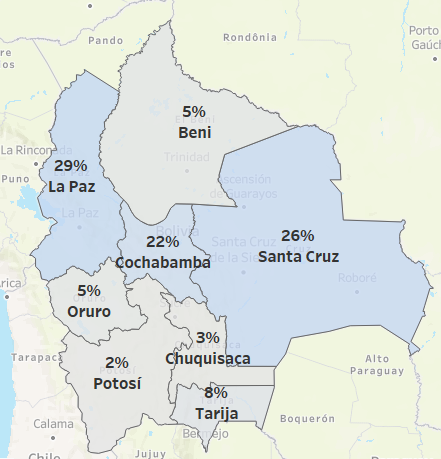
Destacan los servicios de almacenamiento con los nombres "Cloud Plus", "Júpiter-Shared X1", "Cloud X1" y "Shared Turbo XXL" debido a que representan el mayor porcentaje de cuentas suspendidas, con un total del 58% de las 28 cuentas desconectadas, como se observa en la figura 4-2.



###### Figura 4-2: Servicios con el mayor número de cuentas dadas de baja

**Fuente: Elaboración propia, (2023)**

Entre los departamentos del país, se ha identificado que las ciudades de La Paz, Cochabamba y Santa Cruz presentan el 76% de cuentas dadas de baja como se muestra en la figura 4-3.



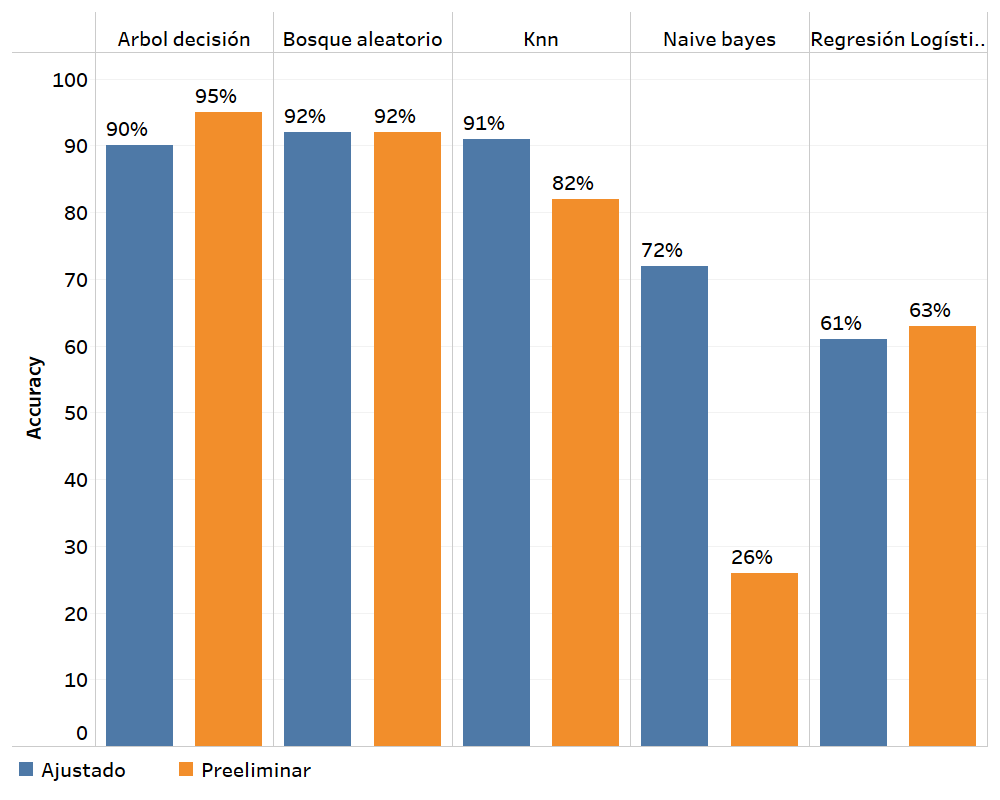
###### Figura 4-3: Porcentaje de cuentas dadas de baja por departamento

**Fuente: Elaboración propia, (2023)**

## Entrenamiento del modelo

Se diseñaron cinco modelos mediante la aplicación de procesos de ajuste de hiper parámetros, técnicas de optimización y validación cruzada. Se llevaron a cabo evaluaciones utilizando los métodos F1-score, recall, matriz de confusión, curva ROC y precisión, en el contexto de una variable de clasificación multiclase. Estos métodos fueron aplicados a cada una de las alternativas o clases dentro de la variable, permitiendo una evaluación detallada de su rendimiento en cada categoría.

El uso de los hiper parámetros en la construcción de modelos de machine learning desempeña un papel crucial en la determinación de su rendimiento predictivo. Los modelos ajustados, que han pasado por un proceso de optimización de hiper parámetros y validación cruzada, muestran un rendimiento sobresaliente en comparación con sus contrapartes preliminares. En particular, al observar el modelo Naive Bayes, la precisión mejora notablemente del 72% en la versión preliminar al 26% después del ajuste. Esta mejora sugiere que la elección cuidadosa de los hiper parámetros ha optimizado la capacidad predictiva del modelo Naive Bayes. Estos resultados subrayan la importancia crucial de considerar y ajustar los hiper parámetros en el proceso de construcción de modelos para garantizar un rendimiento óptimo y confiable ver figura 4-4.



###### Figura 4-4: Comparación entre modelos preliminares y ajustados

**Fuente: Elaboración propia, (2023)**

### Resultados de la precisión de los modelos

Los resultados del Accuracy (precisión global) muestran la capacidad de los diferentes modelos para clasificar correctamente las muestras en el conjunto de prueba. Aquí está la interpretación de los valores de precisión para cada modelo ver figura 4-5.



###### Figura 4-5: Precisión entre modelos

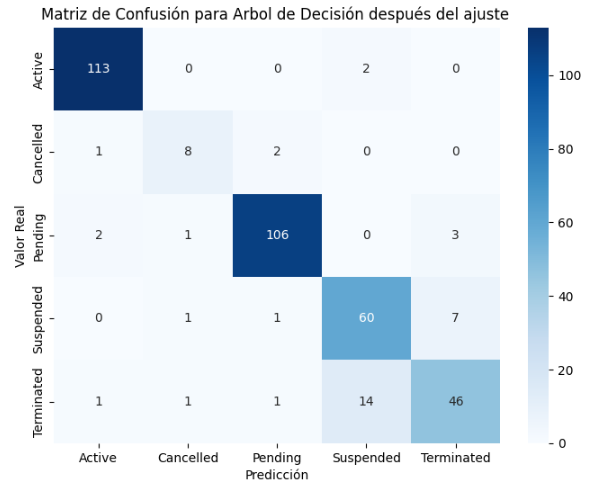
**Fuente: Elaboración propia, (2023)**

* Árbol de Decisión: Destacando un desempeño sólido, este modelo alcanza una precisión del 90%, acertando en la clasificación del 90% de las muestras.
* Bosque Aleatorio: Exhibiendo un rendimiento robusto, en línea con el árbol de decisión, logra una precisión del 92%, clasificando correctamente el 92% de las muestras.
* KNN (Vecinos más Cercanos): Presentando una capacidad aceptable, el modelo KNN obtiene una precisión del 91%, acertando en la clasificación del 91% de las muestras en el conjunto de prueba.
* Naive Bayes: Aunque muestra un rendimiento más limitado con una precisión del 72%, Naive Bayes logra clasificar correctamente el 72% de las muestras, indicando un desempeño intermedio en comparación con otros modelos.
* Regresión Logística: Con una precisión del 61%, la regresión logística sugiere un rendimiento más bajo en comparación con otros modelos, clasificando correctamente el 61% de las muestras.

## Resultados de métricas de evaluación de Árbol de decisiones

La matriz de confusión, cuyos resultados se pueden apreciar en la Figura 4-6, proporciona una visión detallada del rendimiento del modelo.

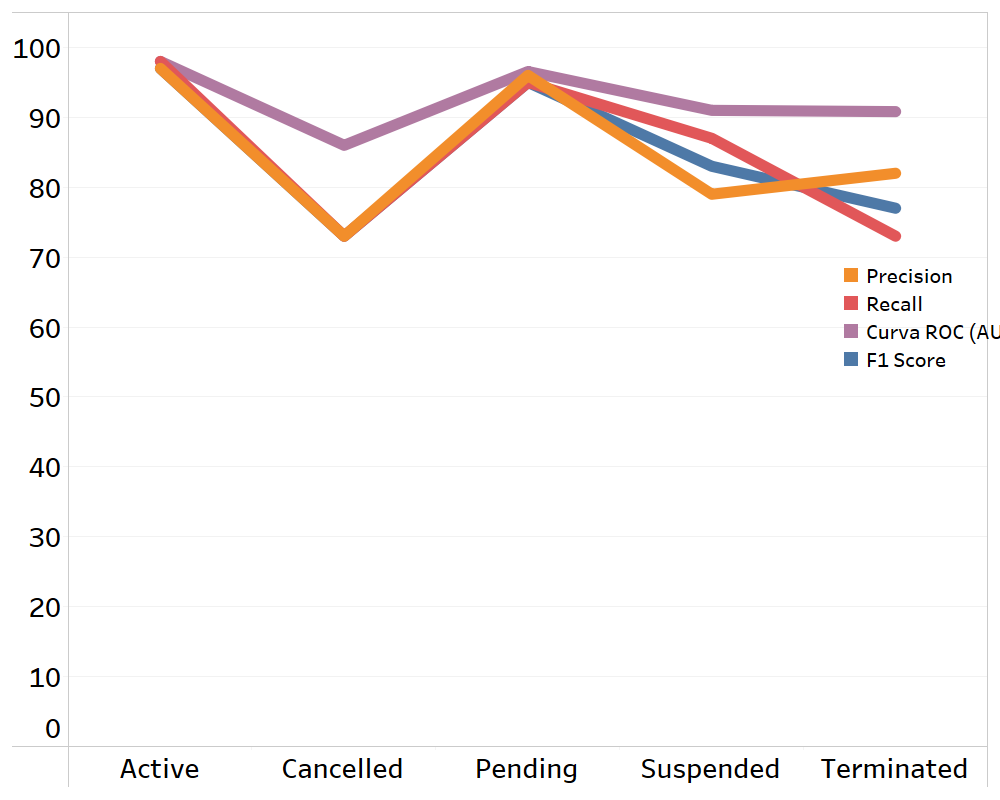
La matriz de confusión ofrece una visión detallada del rendimiento de un modelo de clasificación con cinco categorías: 'Active', 'Cancelled', 'Pending', 'Suspended' y 'Terminated'. En la categoría 'Active', el modelo acertó en 113 instancias y cometió dos errores al clasificar como 'Suspended', sin cometer errores en las demás categorías. En 'Cancelled', el modelo logró 8 clasificaciones correctas y cometió un error al clasificar como 'Active', dos errores al clasificar como 'Pending', y no cometió errores en las demás categorías. Para la categoría 'Pending', el modelo alcanzó 106 clasificaciones correctas, cometiendo dos errores al clasificarse como 'Active', una instancia como 'Cancelled' y tres errores al clasificar como 'Terminated', pero sin errores en las demás categorías. En cuanto a 'Suspended', el modelo acertó en 60 instancias, cometió un error clasificando como 'Cancelled' y otro clasificando como 'Pending', pero tuvo 7 errores al clasificar como 'Terminated'. Finalmente, en la categoría 'Terminated', el modelo acertó en 46 instancias, cometió un error clasificando como 'Active', 'Cancelled' y 'Pending' , 14 errores clasificando como 'Suspended', ver figura 4-6.



###### Figura 4-6: Matriz de confusión árbol de decisiones

**Fuente: Elaboración propia, (2023)**

El rendimiento del modelo destaca en varias categorías. Para 'Active', se logra una precisión del 97%, un recall del 98%, y un f1-score del 97%. En 'Cancelled', se observa un 73% para las métricas de precisión, recall y f1-score. 'Pending' destaca con un 96% de precisión, un 95% de recall, y un f1-score del 95%. 'Suspended' presenta una precisión del 79%, un recall del 87%, y un f1-score del 83%, mientras que 'Terminated' exhibe un equilibrio con un 82% de precisión, un 73% de recall, y un f1-score del 77%. Además, el Área bajo la Curva ROC (AUC-ROC) respalda el rendimiento discriminativo del modelo para cada clase, con valores notables: 'Active' 98.35%, 'Cancelled' 85.95%, 'Pending' 96.55%, 'Suspended' 90.82%, y 'Terminated' 84.88%. Estos resultados ofrecen una visión concisa del desempeño del modelo en la identificación y clasificación de las categorías, expresados en porcentajes, como se muestra en la figura 4-7.

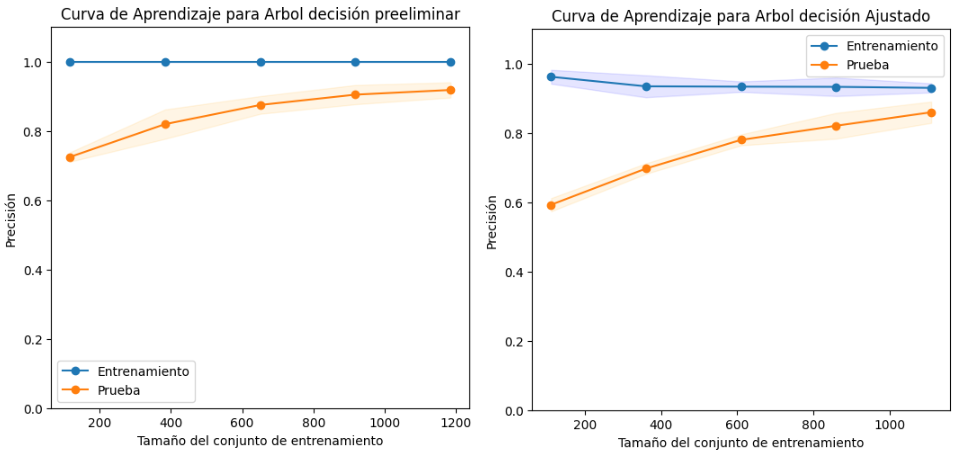


###### Figura 4-7: Métricas de evaluación Arbol de decisiones

**Fuente: Elaboración propia, (2023)**

### Curva de aprendizaje de Árbol de decisiones

Para el modelo preliminar, se observa una precisión perfecta en el conjunto de entrenamiento, lo cual indica un posible sobreajuste, ya que el modelo memoriza los datos de entrenamiento sin generalizar bien a nuevos datos, como se refleja en las puntuaciones de precisión del conjunto de prueba. Por otro lado, en la versión ajustada, se aprecia una disminución en la precisión del conjunto de entrenamiento, lo cual es un indicativo positivo, sugiriendo que el modelo no está memorizando los datos. Además, la precisión en el conjunto de prueba se mantiene alta, demostrando que el modelo ajustado generaliza bien a nuevos datos. La reducción en la brecha entre las puntuaciones de precisión del conjunto de entrenamiento y prueba en el modelo ajustado respalda la afirmación de que este no presenta sobreajuste. Esta mejora sugiere que el modelo ajustado es más robusto y capaz de adaptarse eficientemente a datos no vistos sin comprometer su rendimiento en los datos de entrenamiento. ver figura 4-8

****

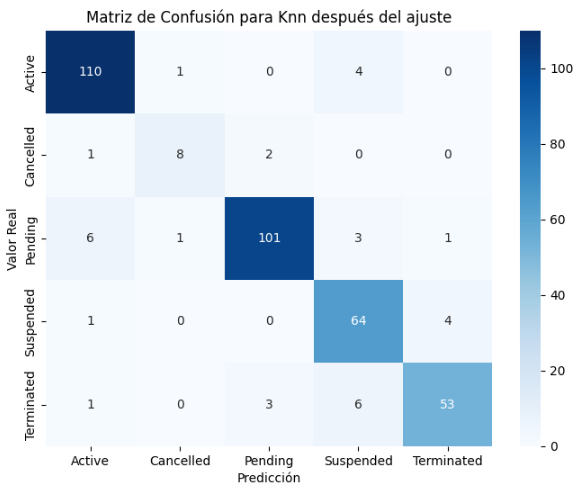
###### Figura 4-8: Curva de aprendizaje Árbol de decisión

**Fuente: Elaboración propia, (2023)**

## Resultados de métricas de evaluación de Knn

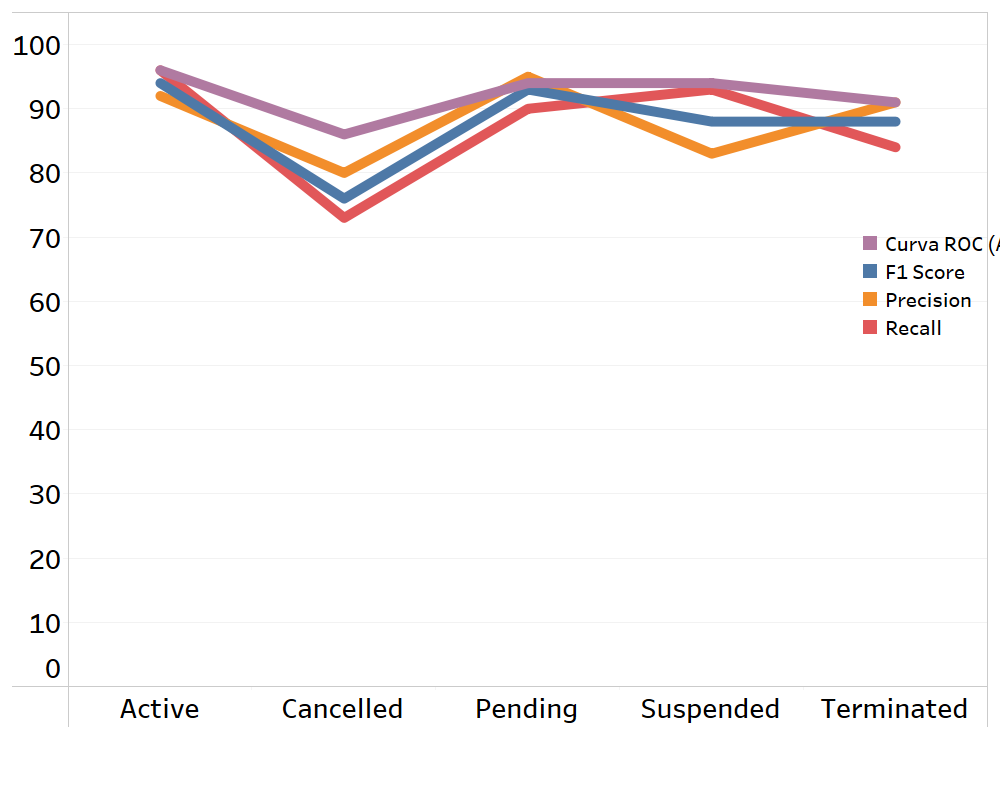
La evaluación detallada de la matriz de confusión del modelo Knn muestra el rendimiento en la clasificación. En la categoría 'Active', el modelo acertó en 110 instancias, pero cometió un error clasificando una instancia como 'Cancelled' y cuatro instancias como 'Suspended', sin errores en las demás categorías. Para 'Cancelled', el modelo tuvo ocho clasificaciones correctas, pero cometió errores al clasificar una instancia como 'Active' y dos instancias como 'Pending'. En la categoría 'Pending', se lograron 101 clasificaciones correctas; sin embargo, se cometieron errores al clasificar seis instancias como 'Active', una instancia como 'Cancelled', tres instancias como 'Suspended', y una instancia como 'Terminated'. En relación con 'Suspended', el modelo acertó en 64 instancias, pero clasificó incorrectamente una instancia como 'Active' y cuatro instancias como 'Terminated', sin errores en las demás categorías. Finalmente, para la categoría 'Terminated', el modelo acertó en 53 instancias, pero cometió errores al clasificar instancias como 'Active', 'Pending' y 'Suspended', sin errores en las demás categorías. Estos detalles brindan una visión completa del rendimiento y los desafíos específicos del modelo Knn en cada categoría, como se muestra en la figura 4-9.

El modelo Knn destaca en la clasificación de diversas categorías con precisión y equilibrio. Para 'Active', logra una precisión del 92%, un recall del 96%, y un f1-score del 94%, respaldado por un AUC-ROC del 96.06%. En 'Cancelled', muestra una precisión del 80%, un recall del 73%, y un f1-score del 76%, con un AUC-ROC del 86.09%. La categoría 'Pending' se distingue con un 95% de precisión, un 90% de recall, y un f1-score del 93%, respaldada por un AUC-ROC del 94.12%. 'Suspended' presenta una precisión del 83%, un recall del 93%, y un f1-score del 88%, con un AUC-ROC del 94.22%. 'Terminated' exhibe un equilibrio con un 91% de precisión, un 84% de recall, y un f1-score del 88%, respaldado por un AUC-ROC del 91.25%, como se muestra en la figura 4-10.



###### Figura 4-9: Matriz de confusión Knn

**Fuente: Elaboración propia, (2023)**



###### Figura 4-10: Métricas de evaluación knn

**Fuente: Elaboración propia, (2023)**

### Curva de aprendizaje de Knn

Los resultados indican que el modelo preliminar experimenta mejoras graduales en la precisión a medida que aumenta el tamaño del conjunto de entrenamiento. Por otro lado, el modelo ajustado muestra una precisión perfecta en el entrenamiento, sugiriendo un ajuste exhaustivo, aunque las puntuaciones en prueba también son altas pero no perfectas, indicando sobreajuste en ambos modelos. ver Figura 4-11

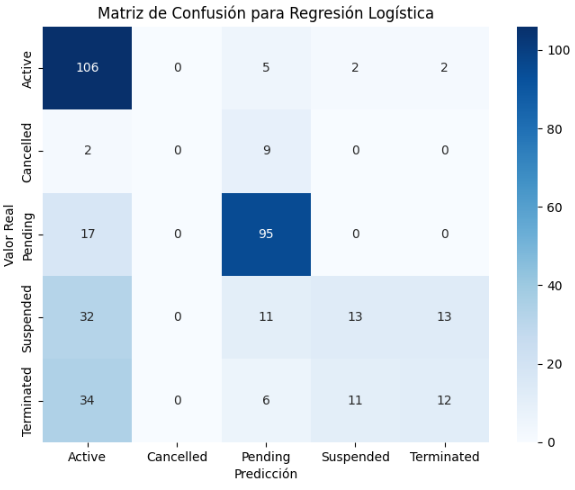
###### Figura 4-11: Curva de aprendizaje para Knn

**Fuente: Elaboración propia, (2023)**

## Resultados de métricas de evaluación de Regresión logística

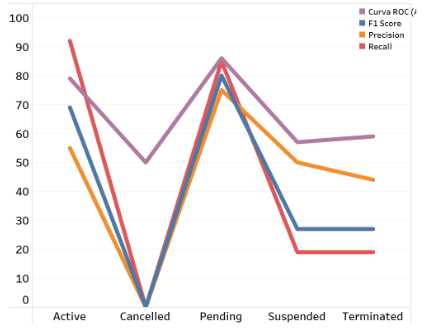
La matriz de confusión revela el rendimiento del modelo de clasificación en distintas categorías. En la categoría 'Active', el modelo clasificó correctamente 106 instancias, cometiendo errores al clasificar 5 instancias como 'Pending', 2 como 'Suspended' y 2 como 'Terminated'. Para 'Cancelled', el modelo no logró clasificar correctamente ninguna instancia, confundiendo 9 instancias como 'Pending'. En la categoría 'Pending', el modelo acertó en 95 instancias, pero cometió errores al clasificar 17 instancias como 'Active'. Para 'Suspended', el modelo acertó en 13 instancias, clasificando erróneamente 32 como 'Active', 11 como 'Pending' y 13 como 'Terminated'. Finalmente, en la categoría 'Terminated', el modelo acertó en 12 instancias, cometiendo errores al clasificar 34 instancias como 'Active', 6 como 'Pending' y 11 como 'Suspended', como se observa en la figura 4-12.

Los resultados de clasificación muestran variados desempeños para cada categoría. En la categoría 'Active', se logra una precisión del 55%, un recall del 92%, un f1-score del 69%, y un AUC-ROC del 79.42%. Para 'Cancelled', la precisión es del 0%, el recall es del 0%, el f1-score es del 0%, y el AUC-ROC es del 50%. 'Pending' destaca con una precisión del 75%, un recall del 85%, un f1-score del 80%, y un AUC-ROC del 86.40%. 'Suspended' presenta una precisión del 50%, un recall del 19%, un f1-score del 27%, y un AUC-ROC del 57.26%. 'Terminated' exhibe una precisión del 44%, un recall del 19%, un f1-score del 27%, y un AUC-ROC del 57.08%, ver figura 4-13.



###### Figura 4-12: Matriz de confusión regresión logística

**Fuente: Elaboración propia, (2023)**

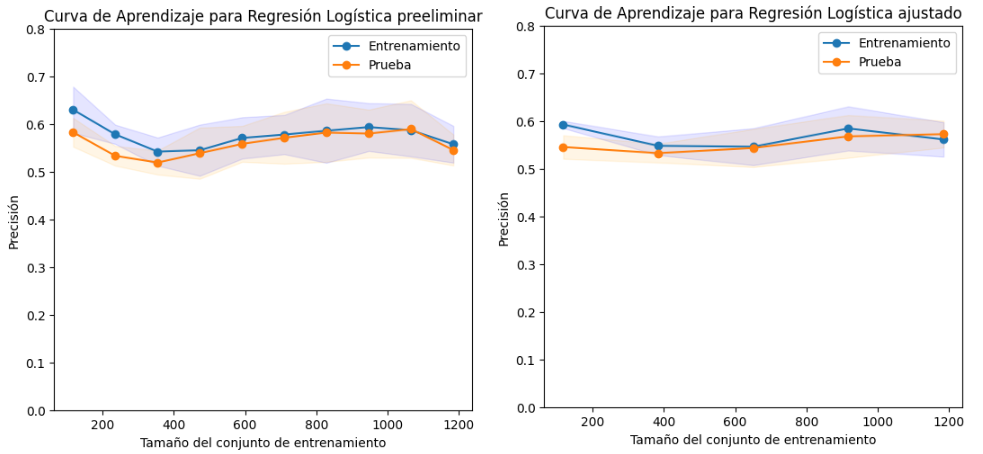


###### Figura 4-13: Métricas de evaluación regresión logística

**Fuente: Elaboración propia, (2023)**

### Curva de aprendizaje de Regresión logística

En la versión preliminar del modelo de regresión logística, se observa un buen ajuste a los datos de entrenamiento, pero una dificultad para generalizar a nuevos datos, como se refleja en la brecha entre la precisión en entrenamiento y prueba. Tras el ajuste, se mejora la capacidad de generalización, aunque persisten algunas discrepancias entre la precisión en ambos conjuntos. En resumen, el ajuste ha mejorado significativamente la capacidad del modelo para adaptarse a nuevos datos, aunque aún existen áreas de posible mejora en términos de consistencia en la generalización ver figura 4-14.

****

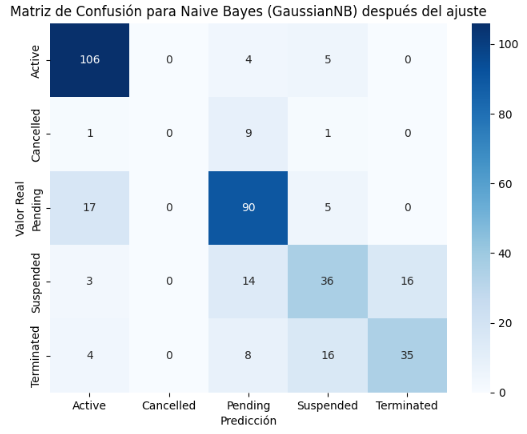
###### Figura 4-14: Curva de aprendizaje regresión logística

**Fuente: Elaboración propia, (2023)**

## Resultados de métricas de evaluación de Naive Bayes

La matriz de confusión ofrece una visión detallada del rendimiento del modelo en la clasificación de diversas categorías. Para la categoría 'Active', el modelo clasificó correctamente 106 instancias, cometiendo errores al clasificar 4 instancias como 'Pending' y 5 como 'Suspended'. En la categoría 'Cancelled', el modelo no logró clasificar correctamente ninguna instancia, confundiendo 9 instancias como 'Pending' y 1 como 'Suspended' y 'Active'. En 'Pending', el modelo acertó en 90 instancias, pero cometió errores al clasificar 17 instancias como 'Active', 5 cómo 'Suspended', y ninguna como 'Terminated'. Para 'Suspended', el modelo acertó en 36 instancias, clasificadas erróneamente 3 cómo 'Active', 14 como 'Pending', y 16 como 'Terminated'. En la categoría 'Terminated', el modelo acertó en 35 instancias, cometiendo errores al clasificar 4 instancias como 'Active', 8 como 'Pending', y 16 como 'Suspended', como se muestra en la figura 4-15.

Los resultados de clasificación proporcionan una evaluación detallada del rendimiento del modelo para cada categoría. En la categoría "Active", se obtiene un recall del 92%, una precisión del 81%, un f1-score del 86%, y un AUC-ROC del 91%. Para "Cancelled", el recall es del 0%, la precisión es del 0%, el f1-score es del 0%, y el AUC-ROC es del 50%. "Pending" destaca con un recall del 80%, una precisión del 72%, un f1-score del 76%, y un AUC-ROC del 83.40%. "Suspended" presenta un recall del 52%, una precisión del 57%, un f1-score del 55%, y un AUC-ROC del 71.60%. "Terminated" exhibe un recall del 56%, una precisión del 69%, un f1-score del 61%, y un AUC-ROC del 75.17%, como se muestra en la figura 4-16.



###### Figura 4-15: Matriz de confusión Naive Bayes

**Fuente: Elaboración propia, (2023)**

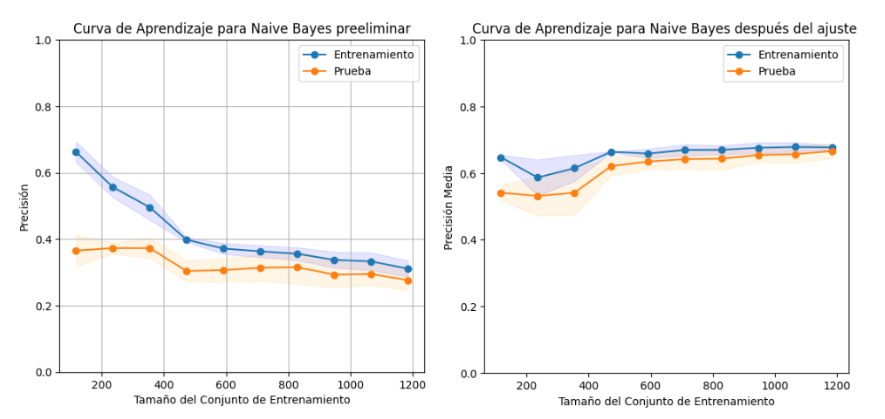


###### Figura 4-16: Métricas de Naive Bayes

**Fuente: Elaboración propia, (2023)**

### Curva de aprendizaje de Naive Bayes

Inicialmente, el modelo Naive Bayes mostró dificultades para aprender patrones en los datos, ya que tanto la precisión en el entrenamiento como en la prueba disminuyeron con conjuntos de datos más grandes. Sin embargo, después de ajustar el modelo, observamos una mejora significativa. En la versión mejorada, la precisión en ambos conjuntos aumentó a medida que se incrementó el tamaño del conjunto de entrenamiento, indicando que el modelo ahora aprende mejor de los datos y generaliza de manera más efectiva a nuevas situaciones. En resumen, el ajuste mejoró considerablemente el rendimiento del modelo, haciéndolo más capaz de comprender y aplicar patrones en los datos ver figura 4-17.

****

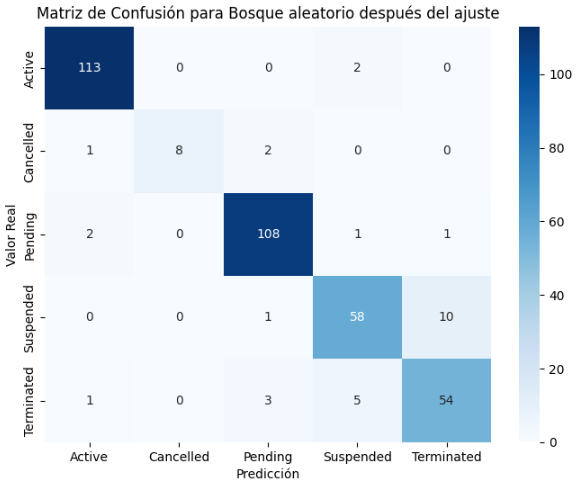
###### Figura 4-17: Curva de aprendizaje para Naive Bayes

**Fuente: Elaboración propia, (2023)**

## Resultados de métricas de evaluación Bosque aleatorio

La matriz de confusión refleja el rendimiento detallado del modelo en la clasificación de cinco categorías. En la categoría 'Active', el modelo acertó en 113 instancias, aunque cometió errores al clasificar 2 instancias como 'Suspended'. Para 'Cancelled', clasificó correctamente 8 instancias, pero erró al etiquetar 2 instancias como 'Pending' y uno como 'Active'. En la categoría 'Pending', el modelo acertó en 108 instancias, pero cometió errores al clasificar 2 instancias como 'Active', 1 cómo 'Suspended' y 1 como 'Terminated'. Respecto a 'Suspended', acertó en 58 instancias, clasificando incorrectamente 1 cómo 'Suspended' y 10 como 'Terminated'. Finalmente, para la categoría 'Terminated', acertó en 54 instancias, pero cometió errores al clasificar 1 instancias como 'Active', 3 cómo 'Pending' y 5 como 'Suspended', como se muestra en la figura 4-18.

Los resultados de clasificación destacan el rendimiento sólido del modelo en diversas categorías. Para "Active", se logra un recall del 98%, una precisión del 97%, un f1-score del 97%, y un AUC-ROC del 98%. En "Cancelled", se obtiene un recall del 73%, una precisión del 100%, un f1-score del 84%, y un AUC-ROC del 86.22%. La categoría "Pending" sobresale con un recall del 96%, una precisión del 95%, un f1-score del 96%, y un AUC-ROC del 97.05%. "Suspended" presenta un recall del 84%, una precisión del 88%, un f1-score del 86%, y un AUC-ROC del 91%. "Terminated" exhibe un recall del 86%, una precisión del 83%, un f1-score del 84%, y un AUC-ROC del 91%. Estos valores detallados ofrecen una evaluación específica del rendimiento del modelo en términos de recall, precisión, f1-score y AUC-ROC para cada categoría, ver figura 4-19.



###### Figura 4-18: Matriz de confusión Bosque aleatorio

**Fuente: Elaboración propia, (2023)**

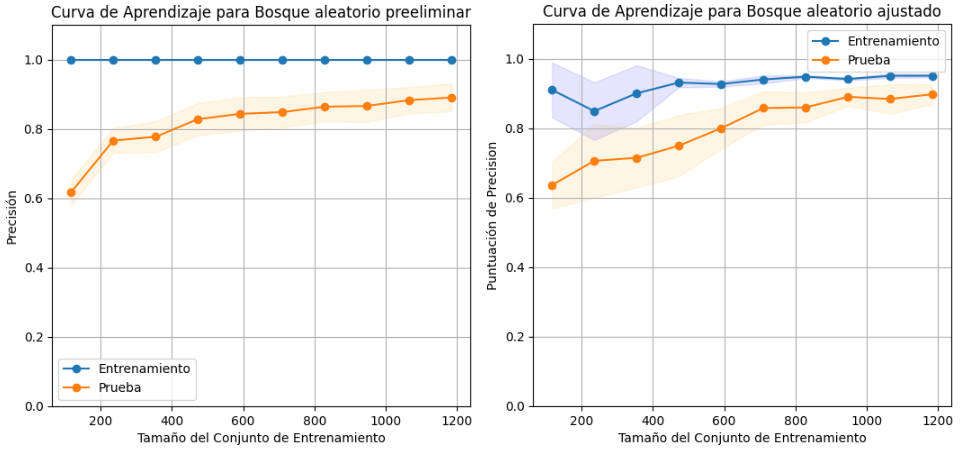


###### Figura 4-19: Métricas de evaluación Bosque aleatorio

**Fuente: Elaboración propia, (2023)**

### Curva de aprendizaje de Bosque aleatorio

En la comparación entre el modelo antes y después del ajuste, se observa que los modelos muestran una mejora significativa después del ajuste, evidenciada por una mayor consistencia y mejora en las puntuaciones de precisión en el conjunto de prueba. El modelo preliminar exhibía sobreajuste, al lograr una puntuación perfecta en el conjunto de entrenamiento pero presentando dificultades en la generalización al conjunto de prueba, especialmente con tamaños de conjunto más pequeños. El ajuste del modelo, al reducir la puntuación de precisión en el conjunto de entrenamiento y mejorar las métricas en el conjunto de prueba, indica una mitigación del sobreajuste y una mejor capacidad para enfrentar datos no vistos, resaltando la importancia del ajuste del modelo en la construcción de modelos más robustos y generalizables ver figura 4-20.

****

###### Figura 4-20: Curva de aprendizaje Bosque aleatorio

**Fuente: Elaboración propia, (2023)**

Los resultados de las métricas de todos los modelos se resumen en la tabla 4-1.

| Modelo | Roc Auc | Precisión | Recall | F1 scores |
| --- | --- | --- | --- | --- |
| Árbol de Decisión | 0,91 | 0.90 | 0.90 | 0.90 |
| **Bosque aleatorio** | **0,92** | **0.92** | **0.87** | **0.90** |
| Knn | 0,92 | 0.88 | 0.87 | 0.88 |
| Regresión logística | 0.62 | 0.45 | 0.43 | 0.41 |
| Naive Bayes | 0.74 | 0.56 | 0.56 | 0.56 |

##### Tabla 4-1: Evaluación de modelos de aprendizaje automático

**Fuente: Elaboración propia, (2023)**

En la evaluación comparativa de los modelos, el Bosque Aleatorio destaca como líder con una precisión de 0.92 en el ROC AUC, un recall de 0.87 y un F1 score de 0.90, según el promedio macro. A pesar de su rendimiento destacado, la diferencia con otros modelos no es sustancial. Tanto el Árbol de Decisión como el KNN muestran un desempeño competitivo en las métricas de evaluación En contraste, la Regresión Logística presenta un rendimiento inferior, con una precisión de 0.45, recall de 0.43 y F1 score de 0.41, todos basados en el promedio macro. Aunque Naive Bayes tiene un rendimiento intermedio, las diferencias entre algunos modelos no son significativas en términos de estas métricas de evaluacion. No obstante, es importante tener en cuenta que el modelo KNN, a pesar de su altas métricas de evaluación, exhibe signos de sobreajuste, lo que sugiere que podría estar demasiado adaptado a los datos de entrenamiento y no generalizar de manera efectiva a nuevos conjuntos de datos. Por lo tanto, se procederá con cautela al interpretar y utilizar las conclusiones basadas en el modelo KNN.

## Discusión de resultados

En esta investigación, se abordó la construcción de diversos modelos de machine learning con el objetivo de predecir el abandono de clientes. Estas técnicas se inscriben en el ámbito de los algoritmos de aprendizaje automático supervisado de clasificación, mostrando su utilidad significativa en el pronóstico de fenómenos como el churn.

En investigaciones sobre la aplicación de aprendizaje automático supervisado para prever el abandono de clientes, se evidencian los resultados del modelo propuesto por (Falla Arango, 2021). En dicho estudio, se emplean diversos algoritmos y se observa una notable coincidencia entre los modelos aplicados, como se refleja en la tabla 4-2, donde "A" hace referencia al estudio de (Falla Arango, 2021) y "B" al presente proyecto, destacándose especialmente XGBoost por su sobresaliente desempeño en métricas de evaluación “Roc Auc”. Este algoritmo se revela como una herramienta eficaz para anticipar el abandono de clientes en empresas del ámbito de las telecomunicaciones, como se ilustra en la tabla 4-3.

| Modelo | Roc Auc | | Precisión | | Recall | | F1 scores | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | B | A | B | A | B | A | B |
| Bosque aleatorio | 0,7849 | 0,92 | 0,7334 | 0,92 | 0,6793 | 0,87 | 0,7053 | 0,90 |
| Knn | 0,7304 | 0,92 | 0,6843 | 0,88 | 0,6709 | 0,87 | 0,92 | 0,88 |
| Regresión logística | 0,7662 | 0,62 | 0,7490 | 045 | 0,6076 | 0,43 | 0,6709 | 0,41 |
| Naive Bayes | 0,7488 | 0,74 | 0,7526 | 0,56 | 0,5783 | 0,56 | 0,654 | 0,56 |

##### Tabla 4-2: Métricas de evaluación entre modelos comunes

**Fuente: Elaboración propia, (2023)**

| **Modelo** | **Roc auc** | **Precisión** | **Recall** | **F1 scores** |
| --- | --- | --- | --- | --- |
| GradienBoosting Classifier | 0,7851 | 0,7328 | 0,6812 | 0,7061 |
| MLPClassifier | 0,7848 | 0,7492 | 0,6519 | 0,6972 |
| Random ForestClassifier | 0,7849 | 0,7334 | 0,6793 | 0,7053 |
| **XGBClassifier** | **0,7854** | **0,7413** | **0,6675** | **0,7025** |
| Logistic Regression | 0,7662 | 0,7490 | 0,6076 | 0,6709 |
| GaussianNB | 0,7488 | 0,7526 | 0,5783 | 0,654 |
| KNeighborsClassifier | 0,7304 | 0,6843 | 0,6709 | 0,6775 |
| LinearDiscriminatAnalysis | 0,7660 | 0,7548 | 0,5945 | 0,6652 |

##### Tabla 4-3: Evaluación de modelos de aprendizaje automático

**Fuente: (Falla Arango, 2021)**

Aunque los resultados obtenidos en el estudio de (Falla Arango, 2021) presentan métricas de evaluación de roc auc, precisión, recall y f1-score inferiores en comparación con el presente proyecto, como se detalla en la tabla 4-4 y se visualiza en la figura 4-21, es crucial tener en cuenta las posibles diferencias entre ambos proyectos que podrían explicar estas disparidades.

| Modelo | Roc auc | Precisión | Recall | F1 scores |
| --- | --- | --- | --- | --- |
| Xgb classifier | 0,7854 | 0,7413 | 0,6675 | 0,7025 |
| **Bosque aleatorio** | **0,92** | **0,92** | **0,87** | **0,90** |

##### Tabla 4-4: Comparación entre modelos

**Fuente: Elaboración propia, (2023)**

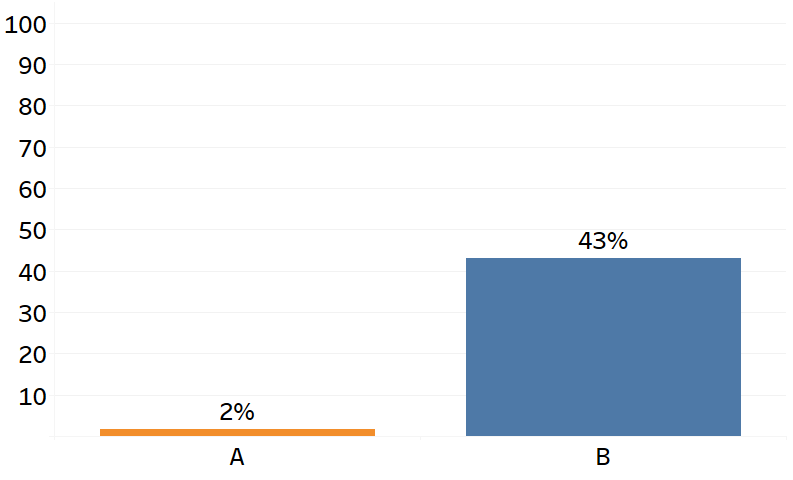
##### 

###### Figura 4-21: Comparación entre métricas de evaluación

**Fuente: Elaboración propia, (2023)**

Una de las discrepancias más destacadas se encuentra en la cantidad y diversidad de datos utilizados para cada estudio. Mientras que (Falla Arango, 2021) trabajó con un porcentaje de churn del 1.8%, nuestro proyecto manejó un considerable 43% de churn, como se muestra en la figura 4-22, donde "A" hace referencia al estudio de (Falla Arango, 2021) y "B" al presente proyecto. Además, se observan variaciones en los periodos de tiempo considerados y en la cantidad de variables manejadas en ambos proyectos.

A pesar de estas diferencias, es importante destacar las similitudes, ya que ambos proyectos lograron resultados notables con modelos de árbol. Esta consistencia refuerza la idea de que los modelos de árbol son altamente efectivos para abordar proyectos de esta índole.

****

###### Figura 4-22: Comparación entre porcentajes de churn

**Fuente: Elaboración propia, (2023)**

Basándonos en los resultados de la investigación de este estudio, concluimos que el objetivo establecido se logró exitosamente. El modelo de Bosque aleatorio ha demostrado su eficacia al predecir la rotación de clientes con alta precisión alcanzando un valor de 92% de acierto en la variable objetivo multivariable. Además, este modelo ha identificado de manera precisa las características que contribuyen a la rotación de clientes

# Conclusiones

En el transcurso de esta investigación sobre la tasa de abandono de clientes, hemos identificado patrones significativos que proporcionan información valiosa para las estrategias futuras de la empresa. Estos hallazgos no solo ofrecen una instantánea de la situación actual, sino que también iluminan áreas potenciales de mejora.

Durante el periodo comprendido entre 2017 y 2023, analizamos 1,849 cuentas registradas en el sistema. Se reveló que un 43% de las cuentas experimentaron suspensiones o bajas, ya sea por falta de pago o por el retiro voluntario de los clientes, marcando el porcentaje de churn.

Según los datos obtenidos por el análisis univariado, se evidencia un periodo crítico durante los primeros 20 meses, caracterizado por la mayor concentración de desconexiones de clientes. Este patrón podría estar asociado a servicios específicos como "Cloud Plus," "Júpiter-Shared X1," "Cloud X1," y "Shared Turbo XXL," los cuales emergen como los principales impulsores de la desconexión, abarcando un considerable 58% de las desconexiones registradas. La interpretación de estos hallazgos sugiere que estos servicios particulares podrían no estar satisfaciendo las expectativas de los clientes o podrían presentar inconvenientes que resultan insatisfactorios para la clientela.

Según el análisis de la deserción de clientes por departamento, se evidencia que las ciudades de La Paz, Cochabamba y Santa Cruz constituyen los principales contribuyentes al abandono de clientes en la empresa, representando un notable 76% de las desconexiones de cuentas. Esta información destaca la necesidad de una estrategia enfocada geográficamente en estas tres ciudades para la retención efectiva de clientes. En base a estos hallazgos, la compañía puede direccionar de manera prioritaria sus recursos para contrarrestar la deserción de clientes en estas áreas geográficas clave.

Se desarrollaron cinco modelos de aprendizaje automático, revelando que tanto el modelo de Bosque Aleatorio como el modelo Árbol de Decisión exhibieron las precisiones más elevadas. Para mejorar aún más el rendimiento y la generalización de estos modelos, se llevó a cabo un ajuste de hiperparámetros, un proceso esencial para optimizar la configuración interna de los algoritmos. Este ajuste permitió refinar la capacidad predictiva de los modelos, asegurando que estuvieran sintonizados de manera óptima para abordar la complejidad de los datos relacionados con el abandono de clientes.

Se utilizaron diversas métricas para medir la eficacia de los modelos en la predicción del abandono de clientes. En términos de Roc Auc, el Bosque Aleatorio lideró con un sólido 92%, seguido de cerca por el Árbol de Decisión y KNN, ambos con un destacado 92%. La interpretación de este indicador revela la capacidad de los modelos para distinguir suficientemente entre clientes que abandonan y aquellos que no lo hacen.

En cuanto a la precisión, el Bosque Aleatorio se destacó con un impresionante 92%, seguido por el Árbol de Decisión con un 90%, KNN con un 88%, Regresión Logística con un 45%, y Naive Bayes con un 56%. La precisión indica qué proporción de las desconexiones identificadas son realmente casos de abandono, y estos valores altos sugieren una identificación precisa de los clientes propensos a abandonar.

En relación con el recall, el Bosque Aleatorio lideró nuevamente con un 87%, seguido por el Árbol de Decisión, KNN, y Naive Bayes, todos con un 90%, y Regresión Logística con un 43%. Este indicador mide la eficacia del modelo para identificar la mayoría de los casos reales de abandono.

En la métrica F1 Score, el Bosque Aleatorio mantuvo su liderazgo con un 90%, seguido por el Árbol de Decisión con el mismo puntaje, KNN con un 88%, Regresión Logística con un 41%, y Naive Bayes con un 56%. El F1 Score ofrece una medida equilibrada entre precisión y recall, resaltando el equilibrio efectivo de estos modelos en la identificación precisa de casos de abandono y la cobertura total.

Durante el proceso de ajuste de hiperparámetros y evaluación de los modelos, se observó que KNN mostraba indicadores inicialmente favorables en términos de métricas como Roc Auc, precisión, recall y F1 Score. Sin embargo, al realizar un análisis más profundo en la curva de aprendizaje, se evidenciaron claros indicios de sobreajuste.

Debido a estos indicios de sobreajuste, se tomó la decisión de excluir el modelo Knn. Este enfoque se basa en la importancia de garantizar que el modelo seleccionado no solo se desempeñe bien en los datos utilizados para su entrenamiento, sino que también mantenga su capacidad predictiva en nuevas instancias, lo cual es esencial para una aplicación exitosa en el entorno empresarial. En contraste, los modelos de Bosque Aleatorio y Árbol de Decisión, que no exhibieron tales problemas de sobreajuste, se consideraron opciones más robustas y confiables para la predicción del abandono de clientes.

Los resultados destacan la efectividad de los modelos de árbol, como el bosque aleatorio y el árbol de decisiones, al abordar de manera eficaz la investigación sobre la pérdida de clientes. Estos modelos se perfilan como preferidos en comparación con las alternativas evaluadas, subrayando su capacidad para ofrecer descubrimientos precisos y valiosos que contribuyen a la comprensión de los clientes que abandonan los servicios. Este respaldo a los modelos de árbol sugiere que su implementación puede ser una estrategia estratégica y beneficiosa para anticipar y gestionar proactivamente la retención de clientes.

Es importante destacar que este respaldo se fundamenta en la curva de aprendizaje de los modelos, que demuestra su capacidad para generalizar patrones y evitar el sobreajuste. La curva de aprendizaje evidencia la consistencia del rendimiento del modelo en diferentes conjuntos de datos.

Para reducir la tasa de abandono de clientes, resulta imperativo mejorar la satisfacción de los usuarios respecto a los servicios ofrecidos. Además, es crucial dirigir estratégicamente las iniciativas de retención de clientes hacia las ciudades de La Paz, Cochabamba y Santa Cruz, donde se concentra un notable porcentaje de desconexiones. La aplicación de herramientas de aprendizaje automático se muestra como un recurso valioso para prever el abandono de clientes, permitiendo a la empresa anticiparse y abordar proactivamente los factores que contribuyen a la pérdida de usuarios.

# Recomendaciones

A partir de los resultados arrojados por la investigación, se derivan sugerencias clave para mejorar la aplicación y efectividad de los modelos de aprendizaje automático en la predicción del churn de clientes:

El análisis de predicción del churn de clientes para el presente proyecto se llevó a cabo empleando una variable objetivo multiclase. No obstante, se sugiere considerar la opción de utilizar una variable uniclase en futuras investigaciones, dependiendo de las circunstancias específicas del proyecto y los datos disponibles. Una variable uniclase podría facilitar la interpretación de los resultados y la implementación de estrategias de retención más directas. Esto podría ser especialmente útil en entornos donde la simplicidad y la claridad son elementos clave en la toma de decisiones estratégicas.

Se recomienda realizar una exploración exhaustiva y la incorporación de variables adicionales relacionadas con incidencias técnicas, el tiempo promedio de respuesta del servicio de soporte, problemas de facturación, entre otros aspectos relevantes. La inclusión de estas variables permitirá capturar aspectos más específicos y detallados del comportamiento de los clientes, brindando una visión más completa de los factores que podrían influir en su decisión de abandonar la empresa. Además, se aconseja incrementar la cantidad de datos utilizados para el entrenamiento del modelo.

Se aconseja una definición clara de las fuentes de datos y el periodo de observación para obtener una visión completa del historial de interacción con el cliente. La calidad de datos desempeña un papel crucial en la mejora de la precisión de los modelos predictivos. Utilizar herramientas de código abierto, como Python y sus bibliotecas de análisis de datos, puede facilitar este proceso de manera accesible y económica.

Los resultados resaltan la importancia de realizar ajustes en los hiper parámetros y realizar técnicas de optimización de modelos. Estas acciones son esenciales para potenciar la efectividad de los modelos, garantizando un rendimiento más robusto e identificar sobreajuste.

La selección cuidadosa de características cruciales influye directamente en el rendimiento predictivo. Se sugiere priorizar características cualitativas para mejorar la precisión de las predicciones. Herramientas de código abierto, como scikit-learn, pueden ser aprovechadas para implementar algoritmos de selección de características de manera efectiva

Dada la destacada precisión de los modelos de árbol de decisión y bosque aleatorio, se sugiere un análisis más profundo de los patrones identificados por estos modelos. La comprensión detallada de las características que más influyen en las predicciones puede orientar estrategias específicas de retención.

Para abordar eficazmente la problemática del abandono de clientes, se sugiere la implementación de técnicas avanzadas, como el Deep Learning, con un enfoque en una variable objetivo única. La aplicación de modelos de Deep Learning permitirá una exploración más profunda de los patrones y relaciones en los datos, brindando una comprensión más precisa y detallada de los factores que contribuyen al abandono de clientes. La elección de una variable objetivo única garantizará una claridad enfocada en los resultados, facilitando la interpretación de los modelos y mejorando su capacidad predictiva.

# Referencias bibliográficas

Amazon web service. (n.d.). ¿En qué consiste el alojamiento web? AWS. Retrieved December 8, 2023, from https://aws.amazon.com/es/what-is/web-hosting/

Barrientos, F., & Ríos, S. A. (2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. Ingeniería Industrial. Retrieved August 23, 2023, from https://www.dii.uchile.cl/~ris/RIS2013/rios.pdf

BET-22. (2023, April 25). ATT. Retrieved August 19, 2023, from https://www.att.gob.bo/sites/default/files/archivos\_listados\_pdf/2023-04-25/Boleti%CC%81n%20Telecomunicaciones%20ANUAL%202022.pdf

Contreras Morales, E. F., Ferreira Correa, F. M., & Valle, M. A. (n.d.). Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión. Dialnet. Retrieved August 23, 2023, from https://dialnet.unirioja.es/servlet/articulo?codigo=6219421

De Leon, A. (2017, Enero 10). Hosting: Definición, Historia y Tipos de Alojamiento Web Hosting: Definición, Historia y Tipos de Alojamiento Web. Infranetworking. Retrieved January 9, 2024, from https://blog.infranetworking.com/web-hosting-definicion-historia-tipos-de-hosting/#Historia\_de\_los\_Servidores\_en\_Internet

Falla Arango, J. D. (2021). PREDICCION DE ABANDONO DE CLIENTES ´ EN TELECOMUNICACIONES MEDIANTE EL APRENDIZAJE AUTOMATICO. Retrieved August 23, 2023, from https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/22247/Tesis%20Churn-Jesus%20David%20Falla.pdf?sequence=1&isAllowed=y

Fiuza Pérez, M. D., & Rodríguez Pérez, J. C. (2000, 12). La regresión logística: una herramienta versátil. 20(6), 477-565. https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-articulo-X0211699500035664

García Herrero, J., Molina López, J. M., Berlanga de Jesús, A., Patricio Guisado, M. Á., Bustamante, Á. L., & Padilla R., W. (2018). Ciencia de datos: técnicas analíticas y aprendizaje estadístico. Un enfoque práctico. Alfaomega.

Gutierrez, M. O. (2015). Advocatus. Establecimiento de comercio en relación con los hosting web, 12(24), : 67 - 77.

i Barrientos Martínez, R. E., Cruz Ramírez, N., Acosta Mesa, H. G., Suárez2, I. R., Gogeascoechea Trejo, M. d. C., Pavón León, P., & Blázquez Morales, S. L. (2009). Árboles de decisión como herramienta en el diagnóstico médico. 9(22), 19-24.

Ine. (2020, Abril). Aspectos Geográficos - INE. Retrieved January 8, 2024, from https://www.ine.gob.bo/index.php/bolivia/aspectos-geograficos/

Kramer, O. (2013). Dimensionality Reduction with Unsupervised Nearest Neighbors. Springer Berlin Heidelberg.

Lazarov, V., & Capota, M. (2007). Churn Prediction.

López, J. F. (n.d.). Teorema de Bayes - Definición, qué es y concepto. Economipedia. Retrieved August 29, 2023, from https://economipedia.com/definiciones/teorema-de-bayes.html

López Navas, J., Villadangos Alonso, J., & Pérez Rández, D. (2020, octubre 2). Business Intelligence y el análisis predictivo: COVID 19. Academica-e. Retrieved December 10, 2023, from https://academica-e.unavarra.es/bitstream/handle/2454/39107/TFG%20-%20Javier%20L%c3%b3pez%20Navas.pdf?sequence=1&isAllowed=y

Luján Mora, S. (2002). Programación de aplicaciones web: Historia, principios básicos y clientes web. Editorial Club Universitario.

McKinney, W. (2013). Python for Data Analysis. O'Reilly Media, Incorporated.

Naive Bayesian Classifier. (2007, November 28). Computer Science and Engineering. Retrieved August 29, 2023, from https://cse.engineering.nyu.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf

Periódico Opinión. (2017, March 4). Minería de datos y fuga de clientes. Retrieved December 8, 2023, from https://www.opinion.com.bo/articulo/opini-oacute-n/miner-iacute-datos-fuga-clientes/20170304200100572703.html

Pineda Pertuz, C. (2022). Aprendizaje automático y profundo en python: Una mirada hacia la inteligencia artificial. Ediciones de la U.

Pineda Pertuz, C. M. (2022). Aprendizaje automático y profundo en Python. Ra-Ma S.A. Editorial y Publicaciones.

Raschka, S., & Mirjalili, V. (2019). Python machine learning: aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow. Marcombo.

Rigatti, S. J. (2017, enero 1). Random Forest. Journal of Insurance Medicine, 47(1), 31–39.

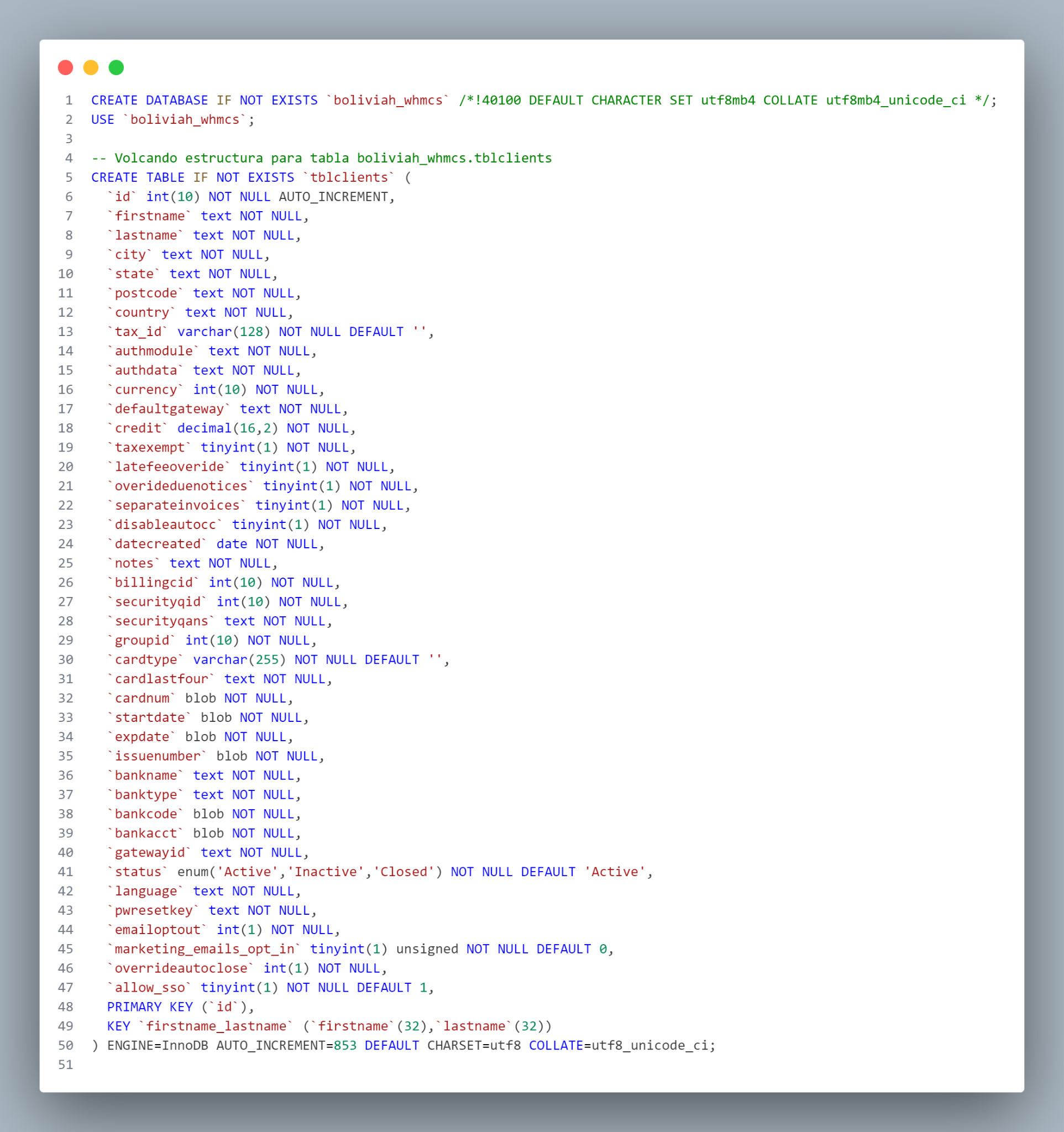
Saldaña, R. (2011). La prueba chi-cuadrado. Enfermería del Trabajo, 1, 31 - 38.

Sandoval, L. J. (2018, 07 19). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. (11). http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6\_RT2018.pdf

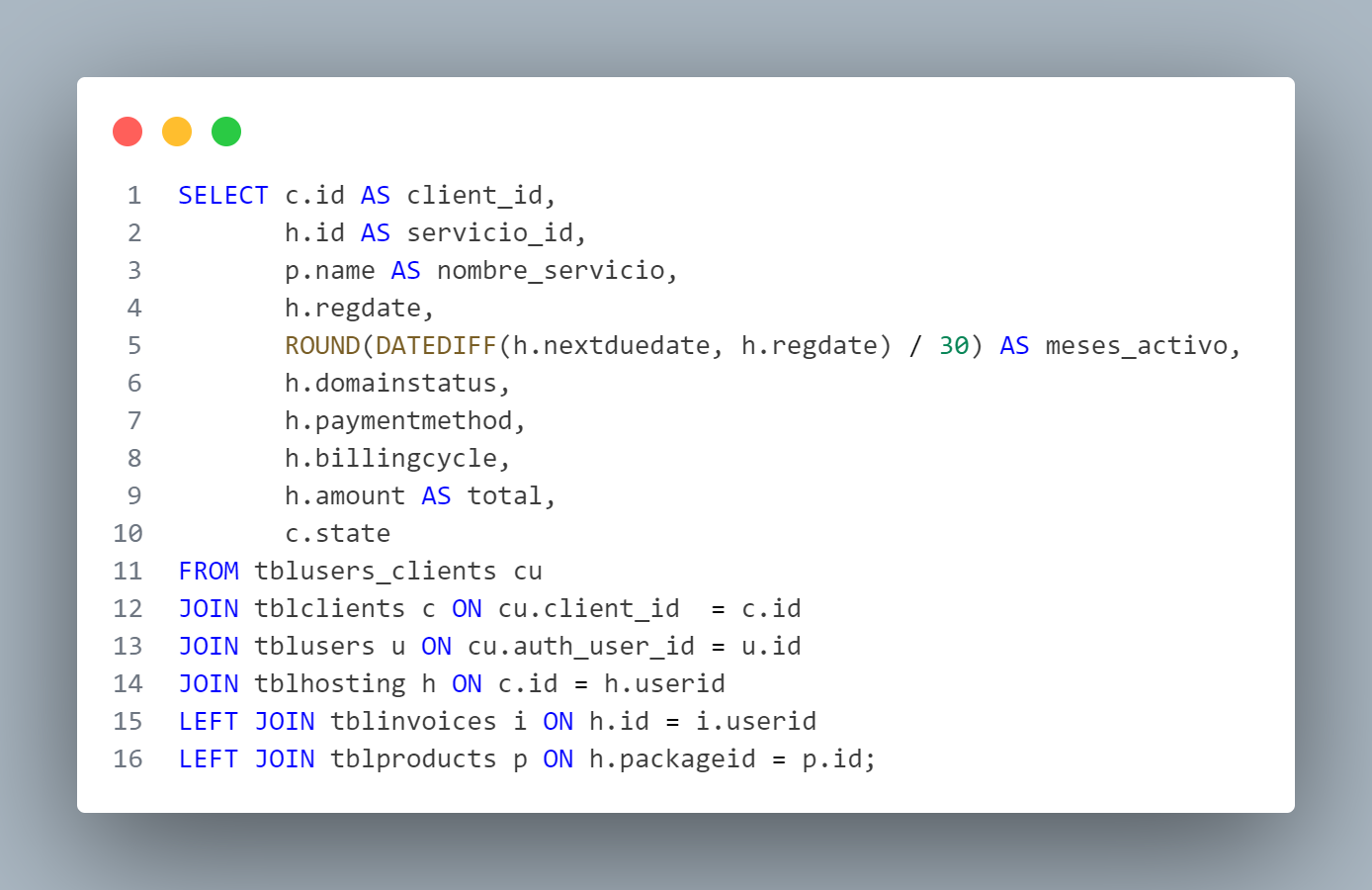
Valle Díaz, F. R., & Huamán Romaní, Y. L. (2022, Abril-Junio). Revista Venezolana de Gerencia. Análisis predictivo de los indicadores macroeconómicos del sector turístico postCovid-19, Perú 2019-2023, 27(98), 580-596. https://dialnet.unirioja.es/servlet/articulo?codigo=8890646

# Anexos

## Base de datos



## Consulta Sql



<https://drive.google.com/file/d/13U5Lc6iqpJDtn6TTfVnK9LG3DQ1ocyRy/view?usp=drive_link>

## Análisis y limpieza de Datos



<https://drive.google.com/file/d/1HnOX_91mTF_cWpe2kI7vQV7rOMgpXf3c/view?usp=drive_link>

## Modelado



<https://drive.google.com/file/d/138216RttSb2OG5kp8k9Ij_8OYRpzbYV7/view?usp=drive_link>

## Cd

## 