Universidad Tecnológica ECOTEC

FACULTAD:
INGENIERÍAS, ARQUITECTURA Y CIENCIAS DE LA NATURALEZA

TÍTULO:
DESARROLLO DE UN MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA ANTICIPAR EL CRECIMIENTO DE LA CARTERA DE CLIENTES EN UNA EMPRESA PROVEEDORA DE SERVICIOS DE FIRMA ELECTRÓNICA.

LÍNEA DE INVESTIGACIÓN:
TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN

MODALIDAD DE TITULACIÓN:
TRABAJO DE INTEGRACIÓN CURRICULAR

CARRERA:
INGENIERÍA EN TECNOLOGÍAS DE LA INFORMACIÓN

TÍTULO A OBTENER:
INGENIERO EN TECNOLOGÍAS DE LA INFORMACIÓN

AUTOR:
NANCY ANGELICA JIMENEZ GALAN

TUTOR:
ING. MARCOS ANTONIO ESPINOZA MINA, PHD.

GUAYAQUIL 2024

ANEXO No. 9

PROCESO DE TITULACIÓN
CERTIFICADO DE APROBACIÓN DEL TUTOR

Samborondón, 19 de diciembre de 2024

Magíster
Erika Ascencio
Facultad de Ingenierías, Arquitectura y Ciencias de la Naturaleza
Universidad Tecnológica ECOTEC

De mis consideraciones:

Por medio de la presente comunico a usted que el trabajo de titulación TITULADO: DESARROLLO DE UN MODELO PREDICTIVO BASADO EN MINERÌA DE DATOS PARA ANTICIPAR EL CRECIMIENTO DE LA CARTERA DE CLIENTES EN UNA EMPRESA PROVEEDORA DE SERVICIOS DE FIRMA ELECTRÓNICA, fue revisado, siendo su contenido original en su totalidad, así como el cumplimiento de los requerimientos establecidos en la guía para su elaboración, por lo que se autoriza al estudiante: **NANCY ANGELICA JIMENEZ GALAN,** para que proceda con la presentación oral del mismo.

ATENTAMENTE,

Firmado electrónicamente por:
MARCOS ANTONIO ESPINOZA MINA

Ing. Marcos Antonio Espinoza Mina, PhD.

Tutor

ANEXO No. 10

PROCESO DE TITULACIÓN
CERTIFICADO DEL PORCENTAJE DE COINCIDENCIAS
DEL TRABAJO DE TITULACIÓN

Habiendo sido revisado el trabajo de titulación TITULADO: DESARROLLO DE UN MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA ANTICIPAR EL CRECIMIENTO DE LA CARTERA DE CLIENTES EN UNA EMPRESA PROVEEDORA DE SERVICIOS DE FIRMA ELECTRÓNICA elaborado por NANCY ANGELICA JIMENEZ GALAN fue remitido al sistema de coincidencias en todo su contenido el mismo que presentó un porcentaje del 9% mismo que cumple con el valor aceptado para su presentación que es inferior o igual al 10% sobre el total de hojas del documento. Adicional se adjunta print de pantalla de dicho resultado.

ATENTAMENTE,

MARCOS ANTONIO ESPINOZA MINA

Ing. Marcos Antonio Espinoza Mina, PhD.
Tutor

Resumen

Este proyecto tiene como propósito desarrollar un modelo predictivo que permita anticipar la renovación o deserción de clientes en la cartera de una empresa proveedora de servicios de firma electrónica. Como caso de estudio, se utilizó Security Data, especializada en la emisión de certificados digitales y soluciones tecnológicas para firmas electrónicas. Esta investigación busca optimizar la retención de clientes en un mercado competitivo y cumplir con la Ley Orgánica de Protección de Datos Personales en Ecuador.

Se adoptó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que estructuró el análisis en fases como la comprensión del negocio, generación de datos, modelado y evaluación. Debido a restricciones legales, se generó un dataset sintético con 22,401 observaciones, diseñado para representar posibles comportamientos reales de los clientes. El dataset incluyó variables como historial de renovaciones, fecha de caducidad del servicio, tipo de servicio, zona geográfica y frecuencia de compras, garantizando representatividad en los análisis.

El modelo predictivo fue construido utilizando el algoritmo Random Forest, conocido por su eficacia en la clasificación. Las métricas de evaluación del modelo, como la precisión, el recall, el F1-score y el AUC, mostraron un desempeño satisfactorio al identificar patrones clave de deserción y renovación. Entre los hallazgos se destaca la relación entre la ubicación geográfica y la probabilidad de deserción, lo que permitió diseñar estrategias personalizadas de retención de clientes, como incentivos específicos para zonas con mayor riesgo de deserción y campañas enfocadas en fidelización.

El impacto del modelo predictivo es significativo: proporciona a Security Data una herramienta basada en datos que permite anticipar comportamientos y tomar decisiones estratégicas en tiempo oportuno. Además, demuestra la utilidad de la minería de datos para resolver problemáticas empresariales reales y ofrece un enfoque replicable para empresas similares en el sector de firmas electrónicas. Este trabajo destaca la relevancia de los modelos predictivos en la mejora de la competitividad empresarial y en la gestión estratégica de la cartera de clientes.

Palabras clave:

Firma electrónica, minería de datos, modelo predictivo, Random Forest, retención de clientes, CRISP-DM, dataset sintético, caso de estudio.

Abstract

This project aims to develop a predictive model to anticipate client renewal or churn within the portfolio of a company providing electronic signature services. Security Data was used as a case study, specializing in issuing digital certificates and technological solutions for electronic signatures. This research seeks to optimize client retention in a competitive market while complying with Ecuador's Organic Law on Personal Data Protection.

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was adopted, structuring the analysis into phases such as business understanding, data generation, modeling, and evaluation. Due to legal restrictions, a synthetic dataset containing 22,401 observations was generated to represent potential real behaviors of clients. This dataset included variables such as renewal history, service expiration dates, type of service, geographic location, and purchase frequency, ensuring representativeness in the analyses.

The predictive model was developed using the Random Forest algorithm, renowned for its effectiveness in classification tasks. Model evaluation metrics, including accuracy, recall, F1-score, and AUC, demonstrated satisfactory performance in identifying key patterns of churn and renewal. Among the findings, the relationship between geographic location and the probability of churn stood out, enabling the design of personalized client retention strategies, such as targeted incentives for high-risk areas and loyalty-focused campaigns.

The impact of the predictive model is significant: it provides Security Data with a data-driven tool to anticipate behaviors and make timely strategic decisions. Furthermore, it demonstrates the utility of data mining to address real-world business challenges and offers a replicable approach for similar companies in the electronic signature sector. This work

highlights the relevance of predictive models in enhancing business competitiveness and strategically managing client portfolios.

Keywords:

Electronic signature, data mining, predictive model, Random Forest, client retention, CRISP-DM, synthetic dataset, case study.

1. Introducción

La toma de decisiones basada en datos se ha consolidado como un factor crucial para el éxito y sostenibilidad de las organizaciones en un mundo globalizado y digitalizado. Las empresas que logran anticiparse a las necesidades y comportamientos de sus clientes pueden diseñar estrategias más efectivas, mejorando su competitividad y alcanzando un crecimiento sostenido. Según Smith (2021), "La minería de datos, junto con el uso de técnicas estadísticas, facilita la conversión de vastos volúmenes de información en insights valiosos" (p. 45).

En este contexto, la minería de datos se presenta como una herramienta poderosa para extraer patrones significativos y conocimientos útiles a partir de grandes volúmenes de información. Esto es especialmente relevante para sectores altamente competitivos y dependientes de la tecnología, como el de los servicios de firma electrónica. Sin embargo, en el entorno ecuatoriano, desafíos como la limitada infraestructura tecnológica y las restricciones legales relacionadas con la protección de datos personales complican la adopción de estas técnicas.

En Ecuador, la firma electrónica está regulada principalmente por la Ley Orgánica de Comercio Electrónico, Firmas Electrónicas y Mensajes de Datos y la Ley Orgánica de Protección de Datos Personales. Estas normativas establecen las bases legales que aseguran la validez y el cumplimiento de las transacciones electrónicas, lo que otorga un marco de confianza tanto a empresas como a usuarios, dado que "las empresas que ofrecen servicios de firmas electrónicas asumirán con los cargos aún siendo leves si un miembro de la organización causa daños hacia un cliente."(Ley de Comercio Electrónico, Firmas y Mensajes de Datos, 2002).

Sin embargo, las estrictas regulaciones en cuanto a la protección de datos personales presentan desafíos para las empresas del sector, que deben cumplir con rigurosos requisitos técnicos para garantizar la seguridad y la privacidad de la información. Según Naranjo Godoy, directora nacional de Registro de Datos Públicos (Dinardap), "la minería de datos no es ilegal, pero debe hacerse con la debida autorización de los propietarios de los datos, respetando la Ley de Protección de Datos Personales" (Oráculo, 2021).

Security Data, es una empresa que emite firma electrónica y otros servicios y es autorizada por la ARCOTEL, es un agente clave en el mercado ecuatoriano de firma electrónica. Su visión corporativa establece que "En el 2027, seremos líderes del mercado ecuatoriano en la creación de una identidad digital para todos los ciudadanos, que les permita interactuar dentro de un entorno tecnológico seguro, a través de servicios digitales y productos revolucionarios" (Security Data, s. f.). Esta visión resalta el compromiso de la empresa con la innovación y el cumplimiento normativo, elementos fundamentales para mantenerse competitiva en un sector cada vez más regulado y dinámico.

Aunque Security Data ha generado una considerable cantidad de datos a través de su portal de ingresos, estos no han sido utilizados de manera estratégica. Como afirman Johnson y Gupta (2020), "Los datos, por sí solo, a menudo no aportan ventajas directas; su verdadero valor reside en la información que podemos obtener de ellos" (p. 23). En este sentido, la minería de datos ofrece una oportunidad única para transformar estos datos en conocimiento útil, lo que podría ayudar a Security Data a predecir comportamientos de renovación y abandono de clientes. Esto es esencial para optimizar sus estrategias de retención y adquisición, asegurando su sostenibilidad frente a la creciente competencia en el mercado ecuatoriano.

Es fundamental que el uso de los datos se alineen con las normativas legales vigentes, especialmente en el contexto ecuatoriano, donde la legislación sobre protección de datos personales establece estrictos lineamientos. Este estudio resalta cómo Security Data puede aprovechar la minería de datos para transformar su vasto repositorio de información en un activo estratégico, al tiempo que reafirma su compromiso con el cumplimiento de la legislación vigente. La adopción de prácticas éticas y responsables no solo fortalecerá su posición competitiva, sino que también contribuirá a la consolidación de un ecosistema digital seguro y sostenible en Ecuador.

El entorno competitivo actual incluye actores como el Banco Central del Ecuador, el Consejo de la Judicatura y FIRNASEGURA, entre otros, quienes también ofrecen servicios de firma electrónica y cumplen con los requisitos normativos establecidos por la ARCOTEL. Este contexto resalta la necesidad de que Security Data implemente estrategias innovadoras basadas en tecnologías avanzadas para consolidar su posición en el mercado.

El presente estudio tiene como objetivo desarrollar un modelo predictivo basado en técnicas de minería de datos para anticipar el crecimiento de la cartera de clientes de Security Data. Este modelo permitirá predecir comportamientos futuros, como probabilidades de renovación y abandono, en una data totalmente sintética y anonimizada, recreando una cartera de clientes con comportamiento similar al servicio ofrecido por una organización proveedora de firma electrónica. Además, proporcionará insights valiosos para mejorar las estrategias comerciales de la empresa por zonas en la ciudad de Guayaquil. Se busca también evaluar la efectividad del modelo a través de métricas de rendimiento, como recomienda Analytics India Magazine (2021): "La evaluación del rendimiento juega un papel dominante en la técnica de modelado predictivo. La elección de las métricas adecuadas es crucial para obtener resultados precisos y útiles."

La importancia de este estudio radica en su potencial para proporcionar a Security Data una ventaja competitiva significativa, al anticipar el crecimiento de su cartera de clientes y optimizar tanto sus estrategias de adquisición como de retención. Estas acciones permitirían a la empresa fortalecer su posición en un mercado dinámico y competitivo, maximizando la eficiencia en el uso de sus recursos. Además, este estudio adquiere mayor relevancia al considerar que, en Ecuador, no se ha identificado un caso de estudio específicamente enfocado en la aplicación de minería de datos o machine learning en el sector de firmas electrónicas. Esto representa una oportunidad única para sentar precedentes en un ámbito aún inexplorado, abriendo la puerta a futuras investigaciones y aplicaciones prácticas.

1.1. Antecedentes

En el contexto de las firmas electrónicas, la literatura aún es emergente sin embargo, en el contexto de las estrategias empresariales, la literatura sugiere que el análisis de datos o la minería de datos juega un papel clave. En este sentido, estudios como el de Huang, J (2019)

indican que "a través de analizar datos de membresía y transacciones, es posible formar con precisión el retrato del cliente y hacer predicciones precisas sobre su comportamiento, lo cual es de gran importancia para la estrategia de marca y la mejora del rendimiento de las empresas". Este hallazgo sugiere que la minería de datos no solo es aplicable a la personalización y retención de clientes, sino también a la anticipación de servicios tecnológicos avanzados, como es el caso de las firmas electrónicas.

Un caso exitoso de la aplicación de estas técnicas es el de Bank of America, que utilizó los historiales de transacciones de sus clientes para desarrollar nuevos modelos predictivos. Estos modelos se enfocan en identificar a los clientes hipotecarios y titulares de tarjetas de crédito en riesgo de deserción. El equipo de científicos de datos del banco colaboró de manera multidisciplinaria con áreas como marketing y atención al cliente, para crear un sistema de recomendaciones que ofreciera ofertas personalizadas a los clientes en riesgo cada vez que contactaran al banco, ya sea en línea, por teléfono o en sucursales. Este sistema permitió identificar de manera proactiva a los clientes en riesgo de deserción, reduciendo el tiempo de cálculo de incumplimiento de pago de préstamos en un 95%. Chitra y Subashini (s.f.), Además, la integración de minería de datos con otros sistemas internos del banco no solo mejoró la eficiencia operativa, sino que también incrementó la retención de clientes. Estos resultados son un testimonio de cómo la minería de datos puede ofrecer soluciones efectivas para la retención proactiva en sectores altamente competitivos.

Si bien el caso de Bank of America demuestra el valor de la minería de datos para la anticipación de la deserción y la mejora de la retención, su aplicación en el sector de las firmas electrónicas aún presenta desafíos. Aunque las soluciones de minería de datos se están implementando con éxito en sectores como el bancario y retail, Security Data y otras empresas

ecuatorianas aún se enfrentan a limitaciones relacionadas con la infraestructura tecnológica y la adopción de herramientas predictivas.

En Ecuador, el uso de técnicas de minería de datos en el ámbito empresarial ha comenzado a aplicarse en sectores como financiero y retail, pero aún está en fase de desarrollo en comparación con otros países con mayores avances tecnológicos. Según un informe de Deloitte (2020) titulado Global Human Capital Trends 2020, "las empresas en Ecuador están adoptando cada vez más tecnologías de análisis de datos para mejorar la eficiencia operativa y la toma de decisiones". Esto refleja una tendencia positiva hacia la digitalización y el uso de análisis de datos en el país, pero también revela que la adopción de técnicas predictivas avanzadas aún es limitada, especialmente en sectores más pequeños o especializados, como el de las firmas electrónicas.

Security Data siendo un caso de estudio muy óptimo tiene la oportunidad de implementar modelos predictivos basados en minería de datos que le permitan anticipar la deserción de clientes y mejorar sus estrategias de retención. Sin embargo, la infraestructura tecnológica de las empresas ecuatorianas, incluidas aquellas que trabajan con firmas electrónicas, debe mejorar para ser compatible con las soluciones de Big Data y análisis avanzado.

En Ecuador, la información acerca de las compañías que ofrecen servicios de minería de datos y análisis avanzado continúa siendo escasa, aunque la gestión de datos se ha transformado en un papel fundamental para el triunfo de las organizaciones actuales. En la situación actual, la habilidad para convertir datos en activos estratégicos es crucial para perfeccionar los procesos de decisión y potenciar la eficacia en las operaciones. Como resalta

Actuaria Asesoramiento Estratégico (s.f.), "mediante la analítica y administración de datos, las entidades pueden convertir la información en un recurso estratégico, mejorando la toma de decisiones, la eficiencia en las operaciones y el servicio al cliente, entre otras ventajas." Adicionalmente, compañías como Ingelsi Cia Ltda (s.f.) utilizan herramientas avanzadas como Business Analytics, minería de datos y Big Data para ayudar a sus socios de negocios a interpretar datos de manera más eficaz. Este enfoque permite a las empresas responder a campañas de marketing, anticiparse a la deserción de clientes, analizar productos en redes sociales o realizar una calificación óptima para otorgar créditos, agregando así valor estratégico a sus operaciones. En resumen, estos elementos evidencian la creciente importancia de la analítica de datos en el contexto empresarial de Ecuador y la importancia de disponer de equipos formados para explotar al máximo el valor estratégico que los datos proporcionan.

En respuesta a la falta de empresas locales dedicadas a este ámbito, algunos estudios individuales en Ecuador han logrado avances notables en la resolución de problemas específicos utilizando técnicas de minería.En este contexto, aunque Security Data enfrenta el reto de la adaptación tecnológica, existe una gran oportunidad para implementar técnicas predictivas que optimicen sus operaciones. Un claro ejemplo de la efectividad de la minería de datos en Ecuador es el estudio de Vela López (2022), quien desarrolló un modelo predictivo para identificar clientes con alta probabilidad de deserción en el sector de telecomunicaciones. El modelo mostró una efectividad del 93.20% en la identificación de clientes susceptibles de abandonar la empresa. Este tipo de enfoque podría ser adaptado para el caso de estudio y mejorar la retención de clientes en el sector de firmas electrónicas.

En Guayaquil, la ciudad más grande y principal centro económico de Ecuador, el desarrollo y la adopción de tecnologías de minería de datos aún son relativamente nuevos como lo indica el reporte chequeo digital 2022-2023 donde muestra los hallazgo encontrando

indicando un alto porcentaje de las empresas el cual el (72.08%) se encuentra en un nivel inicial, lo que sugiere que muchas de ellas aún no consideran utilizar tecnologías digitales para obtener información sobre su gestión, en guayaquil, con su dinámica empresarial y comercial, representa un entorno ideal para la implementación de este tipo de tecnologías, siendo el principal centro económico de Ecuador, demás el reporte destaca la importancia de que las empresas ecuatorianas reconozcan el valor de la analítica de datos y la adopción de tecnologías digitales para mejorar su competitividad y eficiencia. Se recomienda que las empresas implementen estrategias que integren el uso de datos y analítica en sus operaciones, lo que les permitiría identificar áreas de mejora y oportunidades de crecimiento, Ministerio de Telecomunicaciones y de la Sociedad de la Información (2024)

Security Data, con sede en Guayaquil, tiene la oportunidad de establecer un precedente en la ciudad y convertirse en un modelo a seguir para otras empresas locales que buscan mejorar su competitividad mediante el uso de técnicas avanzadas. El éxito de este proyecto en Guayaquil podría impulsar la adopción de prácticas similares en otras empresas, contribuyendo al desarrollo tecnológico y económico de la ciudad.

1.2. Planteamiento del problema

Según estudios recientes revisados, uno de los principales impactos negativos que puede tener una empresa comercial, dependiendo de la naturaleza del negocio, es la disminución de utilidades anuales, lo cual está fuertemente vinculado con la cartera de clientes. La disminución de clientes es una alarma, ya que puede ser casi siete veces más costoso atraer a nuevos clientes que conservar a los actuales. Una elevada tasa de deserción de clientes es una

señal de alerta (Ortega, C, 2023). Esto implica que la estabilidad económica de una empresa no solo se basa en sus ingresos y gastos, sino también en su habilidad para conservar y administrar de manera eficaz su clientela.

El caso de estudio de Security Data muestra cómo la alta tasa de deserción de clientes, o "churn rate", afecta la rentabilidad de la empresa. En los últimos dos años, la empresa ha experimentado una caída en sus ingresos netos por ventas (-9,45%) y otros indicadores financieros negativos, como una disminución en activos (-23,06%) y patrimonio (-15,04%), lo que indica que la rentabilidad general sigue siendo baja (EMIS, n.d.). Estos resultados reflejan no sólo un desafío financiero, sino una pérdida de confianza de los clientes, lo que afecta directamente la sostenibilidad de la empresa.

Según Ortega (2023), la deserción de clientes es un reto importante para muchas industrias, ya que no solo afecta los ingresos, sino que también muestra dificultades en la fidelización de clientes. Este fenómeno es complejo y puede ser causado por diversos factores, como la competencia, que ofrece productos de mejor calidad, precios más competitivos y un servicio al cliente superior (Ortega, 2023). En sectores como las telecomunicaciones y el comercio electrónico, este problema se ve agravado por la gran cantidad de opciones disponibles para los consumidores y sus altas expectativas en cuanto a servicio y entrega. En este contexto, Security Data enfrenta dificultades para mantener la lealtad de sus clientes y mejorar las renovaciones de servicios, ya que el mercado más amplio y competitivo impacta su capacidad para retener clientes.

La minería de datos ha sido un factor determinante en el éxito de empresas como Netflix, que utiliza el análisis de los hábitos de visualización de sus usuarios para personalizar la experiencia, recomendar contenido y anticipar preferencias. Esta estrategia ha permitido a

Netflix mejorar la satisfacción de sus suscriptores, adaptar su catálogo a las diferentes regiones y optimizar la calidad de su servicio, superando así a la competencia. Este modelo de análisis de datos se ha convertido en una fuente de inspiración para otras empresas, que buscan incrementar la lealtad de sus clientes y fortalecer su competitividad en el mercado (Datamedia, 2024). Según Olaniyi et al. (2023), Walmart consiguió disminuir la escasez en un 16% e incrementar las ventas en un 4% a través de la utilización de estas tecnologías para anticipar la demanda y modificar los niveles de stock en tiempo real.

En el caso de Security Data, aplicar un enfoque similar de minería de datos podría ser crucial para identificar patrones en el comportamiento de los clientes, mejorar la retención y reducir la deserción en un mercado altamente competitivo.

Un problema transversal que influye en el caso de Security Data en el contexto de la minería de datos son los problemas legales relacionados con el uso de datos personales. En Ecuador, la implementación de la Ley Orgánica de Protección de Datos Personales en 2021 establece regulaciones estrictas sobre cómo las empresas deben recolectar, procesar y almacenar la información personal. Según Naranjo Godoy, directora nacional de Registro de Datos Públicos, el análisis de datos puede ser viable si se respeta el consentimiento informado de los individuos y se cumple con las normativas legales (Oráculo, 2021b). No obstante, estas disposiciones generan barreras importantes para empresas como Security Data, al limitar el acceso y uso de datos sensibles, lo que puede restringir análisis críticos para predecir, como indica la Dirección Nacional de Registros Públicos (2021). El artículo 66, numeral 19, de la Constitución de la República del Ecuador asegura el derecho esencial a la salvaguarda de la información personal. Esto implica que cualquier recolección, procesamiento o difusión de estos datos debe contar con la autorización expresa del titular o estar respaldada por la ley.

En este contexto, sectores como el de la firma electrónica, donde los datos de clientes son cruciales, enfrentan retos similares. Sin embargo, para evitar comprometer la privacidad de sus clientes y al mismo tiempo garantizar la efectividad de sus estrategias analíticas, Security Data debe explorar alternativas como el uso de datos anonimizados, sintéticos, inteligencia artificial explicable y sistemas de cumplimiento automatizado. Adoptar un enfoque ágil, que combine el respeto a la normativa con la innovación tecnológica, permitirá superar las restricciones legales y fomentar la confianza del cliente, garantizando la sostenibilidad de sus operaciones a largo plazo.

Las limitaciones de este análisis incluyen que, debido a que Security Data no puede hacer uso de su data real con fines de análisis debido a restricciones legales, se optó por la elaboración de una data totalmente sintética para el desarrollo que simule la adquisición de los diferentes servicios que ofrece la organización. Aunque la data replicada se generó a partir de comportamientos de clientes reales de Security Data, esta replicación puede no capturar todas las complejidades o variaciones del comportamiento real de los clientes, lo que podría afectar la precisión del modelo. A pesar de estas limitaciones, grandes ejemplos y líderes en innovación resaltan que la solución no está en detenerse frente a los obstáculos, sino en adaptarse y evolucionar. Marc Randolph, cofundador de Netflix, señaló en una entrevista en Ecuador: “Hoy no tenemos el lujo de esperar cientos de años para que las cosas cambien, hay que evolucionar continuamente, y las grandes empresas con miles de empleados deben pensarse como startups altamente tecnológicas. Esa será la única manera de garantizar la supervivencia del más apto” (Ecuavisa, 2019). Este pensamiento refuerza la necesidad de adoptar enfoques creativos y soluciones tecnológicas que permitan a empresas como Security Data operar dentro del marco legal mientras maximizan el potencial de sus herramientas analíticas.

Este proyecto propone abordar y solucionar la problemática planteada en el caso de estudio de Security Data, desarrollando un modelo predictivo que permita identificar y retener a los clientes mediante el desarrollo de un modelo predictivo. El objetivo principal es predecir con precisión la probabilidad de que un cliente renueve o deserte del servicio de firma electrónica, utilizando el lenguaje R y el algoritmo Random Forest, conocido por su alta capacidad de clasificación y manejo de grandes volúmenes de datos. Para garantizar la efectividad del modelo, se adoptó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que permitió estructurar cada fase del análisis, desde la comprensión del negocio hasta la validación de resultados. Los datos utilizados fueron recreados de comportamientos de clientes de Security Data, procesados y anonimizados para cumplir con las normativas legales de protección de datos. Estos datos se integraron en un DataFrame en RStudio, permitiendo realizar la limpieza, transformación y entrenamiento del modelo predictivo. Como resultado, el modelo Random Forest desarrollado ofrece una herramienta precisa y confiable para identificar clientes con alto riesgo de deserción, habilitando a la empresa para implementar estrategias personalizadas de retención, reducir el churn y mejorar la sostenibilidad del negocio en un mercado altamente competitivo.

1.3. objetivos

1.3.1. Objetivo General

Desarrollar un modelo predictivo con una precisión superior al 80%, basado en técnicas de minería de datos y utilizando el lenguaje R con el algoritmo Random Forest, para proyectar la renovación o deserción de clientes en la cartera de Security Data, considerando variables

relevantes como zonas geográficas, historial de renovaciones y tipos de servicio.

1.3.2. Objetivos Específicos

- Generar un dataset sintético compuesto por al menos 22,000 observaciones, que incluya variables clave como fechas de caducidad, historial de renovaciones, zonas geográficas y frecuencia de compras, asegurando que el dataset represente comportamiento de patrones similares a los clientes reales de Security Data.
- Desarrollar y evaluar un modelo predictivo utilizando el algoritmo Random Forest en R, capaz de clasificar clientes con una precisión superior al 80%, empleando métricas como precisión, recall, F1-score y AUC para garantizar la fiabilidad del modelo en la predicción de la renovación o deserción de clientes.
- Analizar y segmentar los resultados de las predicciones por zonas geográficas específicas (Alban Borja y Las Cámaras), identificando patrones diferenciados de deserción o retención con una representación del 95% de los datos totales del dataset sintético.
- Proponer al menos tres estrategias de retención de clientes basadas en los hallazgos del modelo predictivo, con enfoques específicos para las zonas geográficas analizadas, asegurando que al menos dos de las estrategias puedan ser implementadas a corto plazo.

1.4. Justificación

En el campo de las tecnologías de la información, este proyecto se enfoca en la aplicación de técnicas de minería de datos en un contexto empresarial crucial para la competitividad. Desarrollar un modelo predictivo para proyectar el crecimiento de la cartera de clientes en Security Data representa una oportunidad estratégica para la empresa en el sector

de firmas electrónicas. Este enfoque tiene el potencial de transformar la manera en que la empresa gestiona sus relaciones con clientes, optimiza sus recursos y mejora su rentabilidad a través de un uso más eficiente de los datos disponibles.

En el contexto ecuatoriano, la adopción de tecnologías digitales y analíticas es un factor clave para mejorar la competitividad de las empresas. Según el informe del Banco Interamericano de Desarrollo (BID) sobre la transformación digital en América Latina (2021), menos del 30% de las empresas ecuatorianas han implementado tecnologías avanzadas para la toma de decisiones más objetivas. Este escenario genera una gran brecha en la capacidad de adaptación de las empresas ecuatorianas frente a las demandas cambiantes del mercado. Esta deficiencia en la adopción de herramientas digitales limita la capacidad de las empresas para innovar y competir de manera efectiva, lo que destaca la relevancia de este estudio y la urgencia de la transformación digital en Ecuador. Además, según el Instituto Nacional de Estadística y Censos (INEC), en 2022 solo el 27.5% de las empresas grandes y medianas en Ecuador utilizan algún tipo de análisis de datos para optimizar su gestión empresarial, lo que resalta aún más la brecha significativa en la adopción de tecnologías que favorezcan la gestión estratégica de las empresas.

Este proyecto, al centrarse en la minería de datos en el contexto de una empresa de firmas electrónicas, no solo es relevante para Security Data, sino que también se presenta como una oportunidad crucial para las empresas en Guayaquil y en Ecuador en general. En una ciudad como Guayaquil, que es el principal centro económico del país, la adopción de tecnologías analíticas puede ser un motor clave de competitividad. A través de modelos predictivos, las empresas pueden incrementar la confianza de sus clientes, descubrir nuevas fuentes de ingresos y garantizar que los clientes actuales sigan comprometidos, mejorando la retención y aumentando la satisfacción. Las herramientas de análisis, como las que se proponen en este

proyecto, ofrecen la capacidad de anticipar tendencias, predecir comportamientos y mejorar la gestión de las relaciones con los clientes, lo que resulta esencial para fortalecer el posicionamiento en el mercado. Como indica Amazon Web Services (s.f.), las tecnologías analíticas permiten que las empresas "incrementen la confianza en sus clientes, descubran nuevas vías de ingresos y consigan que los clientes vuelvan".

Desde una perspectiva práctica, los resultados obtenidos de este estudio brindarán a Security Data herramientas efectivas para mejorar la gestión de relaciones con sus clientes, optimizar las campañas de marketing y fortalecer su posicionamiento en un mercado de firmas electrónicas altamente competitivo. La implementación de un modelo predictivo en el ámbito local contribuirá significativamente a mejorar la capacidad de las empresas para adaptarse a las tendencias del mercado y al comportamiento del consumidor.Como enfatiza Verhoef et al. (2021), ""Es crucial adaptar los métodos analíticos a algunas exigencias específicas de cada sector para mejorar los resultados comerciales para alcanzar el éxito".

Este estudio tiene el potencial de impactar positivamente el entorno local y empresarial, ya que Security Data no solo logrará mejorar su retención de clientes a través de un enfoque predictivo, sino que también servirá como modelo de referencia para otras empresas que busquen mejorar su competitividad mediante el uso de tecnologías analíticas avanzadas. Los resultados de este proyecto podrían extrapolarse a otros sectores, contribuyendo al fortalecimiento de la competitividad y rentabilidad en un entorno empresarial cada vez más desafiante en Ecuador, especialmente en un contexto donde la adopción de tecnologías digitales aún se encuentra en niveles bajos, como lo demuestran los estudios recientes.

Predecir el crecimiento de la cartera de clientes permitirá a Security Data planificar de manera más eficiente sus recursos, lo que resultará en una mayor satisfacción y lealtad de los

clientes actuales, así como en la atracción de nuevos clientes a la organización. Además, este modelo predictivo puede convertirse en una herramienta replicable para empresas ecuatorianas de diferentes sectores, lo que contribuirá al fortalecimiento de la competitividad y la rentabilidad en un entorno empresarial que necesita adaptarse rápidamente a los cambios tecnológicos globales.

2. Marco Teórico

El presente marco teórico sustenta la investigación sobre el desarrollo de un modelo predictivo basado en técnicas de minería de datos para anticipar el crecimiento de la cartera de clientes en un escenario de estudio para la empresa de Security Data, una empresa proveedora de servicios de firma electrónica en Ecuador. Cada sección del marco teórico aborda conceptos fundamentales y su relación directa con los objetivos específicos del proyecto, permitiendo establecer una base sólida para la implementación y evaluación del modelo

2.1. Introducción a la Firma Electrónica

En este apartado se explorarán los conceptos básicos de la firma electrónica, los tipos más comunes utilizados en Ecuador y las entidades certificadoras que ofrecen estos servicios en el país. Esta información resulta esencial para comprender el marco tecnológico y normativo que sustenta el uso de esta herramienta en el contexto ecuatoriano, así como su impacto en la modernización de las operaciones empresariales.

2.1.1. Definición y Marco legal

La firma electrónica es una herramienta clave en el proceso de transformación digital de las organizaciones esta surgió a partir de una necesidad evidente “la pandemia de COVID-19” cuando las empresas en el Ecuador tuvieron que apegarse más a una era digital pero esta aceleró la transformación digital en el país. Registro Civil, Identificación y Cedulación. (n.d.)

La firma electrónica es un recurso que permite autenticar y vincular la identidad de una persona a un documento o transacción digital de manera segura, promoviendo la confianza en los entornos digitales. Según el Artículo 13 de la Ley de Comercio Electrónico, Firmas y Mensajes de Datos en Ecuador, una firma electrónica consiste en “datos en forma electrónica, que se incluyen en un mensaje de datos o se asocian a él de manera lógica, permitiendo identificar al propietario y garantizar la integridad del mensaje”. Este marco legal asegura la validez y reconocimiento de las firmas electrónicas en el país.

2.1.2. Tipos de firmas electrónicas

En Ecuador, los tipos de firma electrónica más comunes se clasifican en firma electrónica basada en archivo y firma electrónica con token, ambos regulados por la Agencia

de Regulación y Control de las Telecomunicaciones (ARCOTEL). A Continuación, se describe cada uno según, (Primicias, 2024)

2.1.2.1. firma en archivo

La firma electrónica con token utiliza un dispositivo físico, similar a un pendrive, que contiene los certificados digitales. Este token es necesario para realizar el proceso de autenticación.

Características:

- Está diseñada para transacciones que requieren un nivel adicional de seguridad, como trámites de importación, procesos aduaneros y operaciones financieras.

- El dispositivo token genera y almacena la clave privada del usuario, asegurando que no pueda ser exportada o duplicada.

2.1.2.2. Firma Electrónica con Token

La firma electrónica con token utiliza un dispositivo físico, similar a un pendrive, que contiene los certificados digitales. Este token es necesario para realizar el proceso de autenticación.

Características:

- Está diseñada para transacciones que requieren un nivel adicional de seguridad, como trámites de importación, procesos aduaneros y operaciones financieras.

- El dispositivo token genera y almacena la clave privada del usuario, asegurando que no pueda ser exportada o duplicada.

2.1.3. Entidades de certificación de firma electrónica

Dentro del marco normativo de las firmas electrónicas en Ecuador, varias entidades han sido acreditadas para proporcionar servicios de certificación digital, garantizando así la validez y seguridad de las transacciones electrónicas. Estas entidades están autorizadas por la Agencia Nacional de Regulación, Control y Supervisión de las Telecomunicaciones (ARCOTEL) y deben cumplir con los estándares legales y técnicos establecidos en la legislación ecuatoriana. Las principales entidades certificadoras incluyen:

- ALPHA TECHNOLOGIES CIA. LTDA.
- ANFAC AUTORIDAD DE CERTIFICACION ECUADOR C.A
- ARGOSDATA CERTIFICACIÓN DE INFORMACIÓN Y SERVICIOS RELACIONADOS S.A.S
- BANCO CENTRAL DEL ECUADOR
- CONSEJO DE LA JUDICATURA
- CORPNEWBEST CIA. LTDA.
- DATILMEDIA S.A.
- DIRECCIÓN GENERAL DE REGISTRO CIVIL, IDENTIFICACIÓN Y CEDULACIÓN
- ECLIPSOFT S.A.
- LAZZATE CIA. LTDA.
- SECURITY DATA Y FIRMA DIGITAL S.A.
- UANATACA ECUADOR S.A.

Estas entidades no solo deben cumplir con la normativa nacional, sino también con las exigencias internacionales que aseguran que las firmas electrónicas sean válidas y seguras en el país

En el contexto ecuatoriano, Security Data, como entidad certificadora autorizada por la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL), desempeña un papel fundamental en la emisión de certificados digitales que permiten la implementación de firmas electrónicas. Estas firmas son utilizadas en una amplia variedad de procesos, como la validación de documentos contractuales, transacciones financieras y gestiones gubernamentales, consolidándose como un elemento crítico para el comercio electrónico y la modernización empresarial.

2.1.4. Relevancia en el modelo predictivo

En el desarrollo de un modelo predictivo para anticipar el crecimiento de la cartera de clientes, la firma electrónica desempeña un rol fundamental. Los certificados digitales emitidos por Security Data se convierten en variables clave para analizar patrones de renovación y deserción de clientes. Este enfoque permite identificar tendencias específicas relacionadas con el comportamiento de los usuarios que adquieren servicios de firma electrónica. Además, contribuye a optimizar las estrategias comerciales de la empresa.

tabla 1

Entidad de certificación ecuatoriana

Entidades de Certificación										
Código	**Usuario**	**CI/RUC**	**Tipo**	**Servicio**	**Resolución**	**Acta - Página**	**Fecha Res.**	**Registro**	**Vigencia**	**Estado**
1710752	ALPHA TECHNOLOGIES CIA. LTDA.	1791819926001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-CTHB-CTDS-2022-0225	1-14	21/10/2022	22/11/2022	22/11/2032	VIGENTE
1759188	ANFAC AUTORIDAD DE CERTIFICACION ECUADOR C.A	1792601215001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	RESOLUCIÓN ARCOTEL-2016-0737	1-5	28/10/2016	24/11/2016	24/11/2026	VIGENTE
1724833	ARGOSDATA CERTIFICACIÓN DE INFORMACIÓN Y SERVICIOS RELACIONADOS S.A.S	1793130925001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-1060	1-12	28/9/2021	19/10/2021	19/10/2031	VIGENTE
1700203	BANCO CENTRAL DEL ECUADOR	1760002600001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2018-0902	1-6	25/10/2018	5/11/2018	5/11/2028	VIGENTE

1760659	CONSEJO DE LA JUDICATURA	1768097520001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	TEL-556-19-CONATEL-2014	1-4	28/7/2014	5/9/2014	5/9/2024	VIGENTE
0189605	CORPNEWBEST CIA. LTDA.	0190459616001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2023-0111	1-15	13/6/2023	13/7/2023	13/7/2033	VIGENTE
0931019	DATILMEDIA S.A.	0992712554001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-0923	1-11	6/8/2021	30/8/2021	30/8/2031	VIGENTE
1756947	DIRECCIÓN GENERAL DE REGISTRO CIVIL, IDENTIFICACIÓN Y CEDULACIÓN	1768049390001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-0013	1-7	7/1/2021	10/2/2021	10/2/2031	VIGENTE
0949220	ECLIPSOFT S.A.	0992253428001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-0736	1-10	25/6/2021	5/7/2021	5/7/2031	VIGENTE

1731683	LAZZATE CIA. LTDA.	1891756115001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-CTHB-CTDS-2022-0068	1-13	10/5/2022	24/5/2022	24/5/2032	VIGENTE
1748060	SECURITY DATA Y FIRMA DIGITAL S.A.	1792261848001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-0398	1-8	19/2/2021	24/12/2020	24/12/2030	VIGENTE
1783021	UANATACA ECUADOR S.A.	1793081770001	JURÍDICO	ACREDITACIÓN COMO ENTIDAD DE CERTIFICACION	ARCOTEL-2021-0395	1-9	19/2/2021	15/3/2021	15/3/2031	VIGENTE

Nota: datos tomados de la Agencia de Regulación y Control de las Telecomunicacione (ARCOTEL, s.f.)

2.2. Crecimiento de la Cartera de Clientes

En este apartado se analizarán los conceptos clave relacionados con la cartera de clientes, incluyendo su definición, el ciclo de vida del cliente y los factores que afectan su pérdida. Asimismo, se explorará cómo la correcta gestión de esta cartera puede influir en la capacidad de la empresa para fortalecer sus relaciones comerciales y mejorar su competitividad en el mercado. Este análisis sentará las bases para comprender la importancia de las estrategias predictivas aplicadas en el presente proyecto.

2.2.1. Concepto de cartera de clientes

Una cartera de clientes se refiere a un grupo de personas con características similares en términos de ingresos y hábitos de consumo, lo que facilita a las empresas diseñar estrategias y ofrecer nuevas oportunidades comerciales, En el ámbito comercial, la cartera de clientes se considera una herramienta esencial, ya que representa el conjunto de personas con las que un agente ha construido relaciones de confianza a través de un trabajo constante. Este grupo suele incluir a clientes que recurren a servicios ofrecidos por agentes de ventas, corredores de seguros, propiedades u otros servicios personalizados (Orellana Nirian & López, 2020)

2.2.2. Valor del cliente

El valor del clientes es una medida que estima cuánto dinero un cliente invertirá en una empresa a lo largo de su relación comercial. Este valor abarca compras de productos, tarifas por servicios, suscripciones y otros cargos adicionales. En términos sencillos, refleja el monto que se anticipa que un cliente gastará durante el tiempo que permanezca como parte de la empresa (Sprinklr, 2024)

Según el estudio de la Cámara de Comercio de España, el valor del cliente ya no se limita únicamente a los ingresos generados durante un año fiscal, como se hacía

tradicionalmente. Actualmente, muchas empresas consideran el valor del cliente a lo largo de todo el período en que este se mantiene activo.Además, algunas organizaciones emplean herramientas financieras como el Valor Actual Neto (VAN) para estimar este impacto, lo que permite incrementar tanto el valor percibido del cliente como la valoración de la empresa, especialmente en procesos de venta (Cuesta, n.d.)

2.2.3. Ciclo de vida del cliente

El ciclo de vida del cliente se refiere a las interacciones continuas que un cliente tiene con una marca, comenzando desde el primer contacto hasta la compra, el uso del producto, el soporte y finalmente convirtiéndose en un defensor fiel de la marca. Este ciclo influye en todos los aspectos operativos de una empresa, desde las estrategias de marketing y ventas hasta los servicios de atención al cliente, determinando cómo estos departamentos se relacionan con los clientes y evalúan su éxito (Sprinklr, 2024).

2.2.4. Factores clave en la pérdida de clientes

En el caso de Security Data, podemos ver una disminución en sus ingresos anuales asociados directamente con la pérdida de clientes y también mayores costos asociados a la adquisición de nuevos clientes. Este fenómeno afecta no solo la rentabilidad, sino también la sostenibilidad a largo plazo de la empresa. Como indica (Asana , 2024), La pérdida de clientes en una cartera puede derivarse de varios factores clave:

- **Desorganización:** La falta de respuesta a comunicaciones, errores en el manejo de información y descuidos en hitos importantes afectan la relación con los clientes.

- **Falta de claridad en objetivos y servicios:** No comunicar las metas y responsabilidades de manera clara genera confusión y expectativas no cumplidas. Esto puede provocar que los clientes busquen opciones más confiables.

➢ **Desatención a clientes actuales:** Priorizar la adquisición de nuevos clientes sobre el mantenimiento de los actuales es un error frecuente. la poca importancia a los clientes establecidos puede llevar a la pérdida de vínculos con la marca y oportunidades de negocio a largo plazo.

2.2.5. Gestión de carteras de clientes

Este aspecto sugiere cuál debería ser la conducta del encargado de administrar una cartera de clientes (Zendesk, 2023)

➢ **Manejo de cartera activa de clientes**

En la gestión de una cartera activa, el responsable tiene la tarea de generar lazos con los clientes que realizan compras de manera habitual. Por ejemplo, si gestionas una tienda de artículos para animales y un individuo ha adquirido una jaula para pájaros, puede volver a adquirir alimentos, juguetes y otros productos. Este posible acto de adquisición es de interés para la compañía y necesita ser monitoreado de cerca.

➢ **Manejo del portafolio inactivo**

Para la administración de cartera inactiva, el administrador se encarga de las relaciones con los clientes que en algún momento adquirieron a su compañía, pero que rompieron dicha relación comercial debido a razones desconocidas o supuestas. El propósito no solo es reiniciar la relación, sino también entender la causa del conflicto.Por ejemplo, el uso de datos como historial de renovaciones y frecuencia de compras ayudaría a Security Data a diseñar incentivos específicos para fidelizar a los clientes.

2.3. Minería de Datos

En el contexto de Security Data, la minería de datos permite analizar el comportamiento en base de patrones referente al historial de la vasta cartera de clientes, como el historial de renovaciones, zonas geográficas. Estos análisis no solo facilitan la identificación de patrones de deserción, sino que también ayudan a segmentar la cartera de clientes según su probabilidad de renovación.

2.3.1. Concepto de minería de datos

La minería de datos, también denominada como descubrimiento de conocimiento en datos (KDD), es el procedimiento para identificar patrones y otros datos de gran envergadura (IBM, 2024).

La minería de datos es el procedimiento de analizar datos con el objetivo de descubrir patrones y anticiparse a tendencias futuras, cuenta con una amplia historia. A pesar de que el concepto de "data mining" se originó únicamente en los años 90, sus bases se fundamentan en tres campos científicos interconectados: la estadística (que se centra en el estudio numérico de las conexiones entre datos), la inteligencia artificial (que simula la inteligencia humana mediante programas y maquinaria) y el aprendizaje automático (algoritmos que extraen información de los datos para realizar proyecciones, (SAS, 2024) , Se podría deducir que la minería de datos es una práctica moderna que fusiona métodos de estadística, inteligencia artificial y aprendizaje automático para analizar información y prever conductas futuras.

La minería de datos resulta particularmente beneficiosa en contextos donde se gestionan grandes cantidades de información, tales como el ámbito financiero, el marketing, la salud, las telecomunicaciones y el comercio minorista (Rodríguez & Rodríguez , 2024)

2.3.2. Sectores de aplicación de Minería de Datos

El uso de esta técnica está comprobada en su estudio de diversas índoles de aplicación tales como;

- **Entretenimiento:** Para personalizar sugerencias de contenido, comprender la conducta del consumidor y anticipar el triunfo de películas y programas de televisión (Ceupe, 2024)
- **Salud:** Para identificar patologías, anticipar epidemias de enfermedades, adaptar los tratamientos médicos y optimizar la administración de recursos hospitalarios (Ceupe, 2024)
- **Telecomunicaciones:** Para optimizar la administración de redes, evitar la pérdida de clientes y elevar el nivel de servicio (Ceupe, 2024)
- **Mercadotecnia**: Para detectar patrones de adquisición de los consumidores, segmentar el mercado, anticipar tendencias de consumo y personalizar sugerencias de productos (Ceupe, 2024)
- **Finanzas:** Para identificar fraudes, anticipar peligros crediticios, mejorar los portafolios de inversión y evaluar el desempeño del mercado (Ceupe, 2024)
- **Educación:** Para personalizar la enseñanza a través de la modificación del currículo, detectar patrones de rendimiento de los estudiantes y potenciar la permanencia en el aula (Ceupe, 2024)
- **Manufactura**:Para optimizar la incorporación y permanencia de empleados, detectar patrones en el rendimiento laboral y anticipar la falta de asistencia (Ceupe, 2024)

Pero emplearlas en diferentes escenarios trae consigo retos y no es lo mismo emplearlo en la en una entidad bancaria como emplearlo en una entiéndase de firma electrónica también depende del país donde se realice estas prácticas

2.3.2.1. Comparativa del sector bancario extranjero y uno de telecomunicaciones nacional

Tabla #2

Comparativa nacional e internacional de sectores de minería de datos

Aspectos	Bank of America	Empresa Ecuatoriana de Firmas Electrónicas
Marco Regulatorio	Regulaciones bancarias estadounidenses La Ley Gramm-Leach-Bliley (GLBA) exclusivo para las financieras a revelar sus prácticas de intercambio de datos	Ley de Protección de Datos de Ecuador y regulaciones de firma electrónica
Objetivo Principal	Identificar y prevenir la fuga de clientes, mejorando la retención mediante estrategias proactivas.	Con un aspecto similares pero desde Determinar porcentajes de renovación y deserción de clientes en diferentes zonas geográficas de la empresa.
Infraestructura Tecnológica	Sistema bancario robusto con capacidad de procesamiento masivo	Base de datos sintética que replica patrones de comportamiento de clientes, diseñada para cumplir regulaciones de protección de datos.
Procesamiento de Datos	Análisis avanzado de historiales de transacciones reales, incluyendo compras, préstamos y patrones de crédito.	Uso de datos sintéticos centrados en variables como historial de renovaciones, frecuencia de compras y zonas geográficas.
Personal Requerido	Esta empresa cuenta con expertos en el área de minería de datos, expertos en banca y tecnología de la información.	Equipo más reducido con experiencia en criptografía, seguridad digital y desarrollo de páginas web.
Impacto en el Negocio	Reducción del 95% en el tiempo de cálculo para identificar incumplimientos; aumento significativo en retención.	Mejor rentabilidad al optimizar estrategias de retención en un mercado con creciente competencia en firmas electrónicas.
Desafíos Principales	Integración de sistemas complejos, cumplimiento de regulaciones internacionales y procesamiento en tiempo real.	Cumplimiento normativo estricto en protección de datos y diseño de modelos predictivos con datos limitados.

Inversión y Recursos	Alta inversión en infraestructura tecnológica, personal especializado y sistemas integrados de Big Data.	Inversión moderada, enfocada en seguridad y cumplimiento y otros aspectos

Fuente: elaboración propia bajo del análisis de Bank of America y el presente caso de estudio

La comparativa destaca diferencias en la escala y complejidad de las operaciones entre una institución financiera global como Bank of America y una empresa ecuatoriana de firmas electrónicas, pero también resalta similitudes en los objetivos de retención de clientes y el uso de datos predictivos. La empresa ecuatoriana, aunque con menos recursos, enfrenta desafíos similares en términos de seguridad, cumplimiento normativo y optimización de procesos, lo que refuerza la relevancia de la minería de datos en cualquier sector, independientemente del tamaño de la empresa.

2.3.3. Marco Legal en la Práctica de la Minería de Datos en Ecuador

La minería de datos se ha transformado en un recurso imprescindible para las compañías que aspiran a mejorar su proceso de toma de decisiones y potenciar su competitividad en el mercado. Sin embargo, en Ecuador, el progreso de esta práctica se ve afectado por el marco regulatorio que salvaguarda la información personal de los ciudadanos.

2.3.3.1. Legislación Relevante , Ley Orgánica de Protección de Datos Personales

La Ley Orgánica de Protección de Datos Personales, establecida en 2021, proporciona pautas precisas sobre la gestión, almacenamiento y procesamiento de datos, asegurando el derecho a la privacidad y la protección de la información según lo establecido en el (Artículo 82, Ley Orgánica de Protección de Datos Personales, 2021): "Los prestadores de servicios no pueden utilizar los datos personales, la información de uso del servicio, datos de tráfico o patrones de consumo de sus clientes para fines comerciales (como promocionar productos o servicios) sin el consentimiento expreso del usuario. Según la Ley Orgánica de Protección de Datos Personales, el usuario debe dar su autorización explícita para que sus datos sean utilizados con ese propósito"

La Ley Orgánica de Protección de Datos Personales, actualmente en vigor, establece derechos y obligaciones que son pertinentes y de obligado cumplimiento para todas las entidades que gestionan datos personales (PwC, n.d.)

Esta normativa impone varias limitaciones a la implementación de herramientas de minería de datos de inteligencia de negocios en empresas ecuatorianas:

- La ley exige que las empresas obtengan el consentimiento explícito de los usuarios antes de utilizar sus datos personales para fines de análisis. Esto limita la capacidad de las empresas para recolectar y utilizar datos de manera generalizada, lo que puede dificultar la construcción de modelos predictivos efectivos.
- Los datos recolectados deben emplearse sólo con el propósito para el que fueron autorizados. Esto implica que, si los datos se obtuvieron con el objetivo de proporcionar un servicio, no pueden ser reutilizados para fines de marketing

- Los usuarios tienen el derecho de acceder a la información que se posee sobre ellos y solicitar su eliminación.

En este sentido, el papel de la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL) es fundamental para garantizar un crecimiento ordenado y sostenible del sector de telecomunicaciones en Ecuador, fomentando el acceso justo, la innovación y la salvaguarda de los usuarios. En la actualidad, ARCOTEL cuenta con un centro de datos de vanguardia que permite mantener los servicios informáticos disponibles en favor de la ciudadanía (Agencia de Regulación y Control de las Telecomunicaciones, 2024).

2.3.3.2. Constitución de la República del Ecuador

En su artículo 66, numeral 19, la Constitución de la República indica que "el derecho a la protección de datos es de carácter personal comprende el acceso y la decisión sobre información y datos de tal naturaleza, además de su correspondiente salvaguarda". (Ley de Protección de Datos Personales - Dirección Nacional de Registros Públicos, 2024b). Este marco normativo resalta la importancia de manejar adecuadamente la información personal de los clientes, lo cual es crucial para la recolección y análisis de datos.

Desde la perspectiva jurídica, las organizaciones deben ser conscientes de sus responsabilidades hacia los titulares de los derechos, lo cual implica reconocer la importancia de implementar la regulación vigente en materia de protección de datos.

Actualmente, las sanciones por violar esta normativa pueden implicar multas que ascienden entre el 0.7% y el 1% del volumen de negocios anual de la empresa (Alonso, 2023c). Además, el incumplimiento de los derechos de los titulares de los datos puede no solo llevar a procesos judiciales y pérdidas económicas, sino también afectar significativamente la reputación de la organización.

2.3.4. Diferentes Métodos en la minería de Datos

En el caso de estudio de Security Data, se optó por un enfoque de aprendizaje supervisado, por el motivo de la aplicación del modelo Random Forest, que ayude anticipar el comportamiento de la cartera de clientes. La decisión de utilizar este enfoque estuvo motivada por la necesidad de predecir comportamientos específicos de los clientes, tales como la probabilidad de renovación o deserción de servicios, basándose en datos históricos.

2.3.4.1. Aprendizaje supervisados

La finalidad del aprendizaje supervisado es identificar una función que dado un conjunto de datos de entrenamiento y sus correspondientes etiquetas, resulta más beneficioso en la conexión entre las variables de entrada y salida detectadas en la información, Normalmente, se emplea el aprendizaje supervisado para problemas de clasificación y regresión. Algunos de los algoritmos más frecuentemente utilizados incluyen: regresión logística, máquinas de soporte vectorial, bosques aleatorios y redes neuronales artificiales (Hernández Gómez, 2019)

2.3.4.2. Aprendizaje no supervisados

El aprendizaje no supervisado, que no se basa en etiquetas de salida, tiene como objetivo descubrir la estructura "intrínseca" que se encuentra en el conjunto de datos. El aprendizaje no supervisado generalmente se emplea para agrupar, segmentar y disminuir la dimensionalidad, algunos de los algoritmos más comunes son: k-medias, análisis de componentes principales y factorización no negativa de matrices (Hernández Gómez, 2019)

2.3.5. Tipos de Datos en la minería de Datos

Para llevar a cabo este análisis, se utilizaron datos sintéticos creados con el propósito de replicar patrones reales de comportamiento de clientes. Estos datos fueron organizados en

diferentes tipos, que jugaron un papel crucial en la construcción y mejora del modelo predictivo, permitiendo evaluar con mayor precisión la probabilidad de renovación de servicios primeramente se muestran estos tipo de datos frecuentemente utilizados y como contribuyeron en el caso de estudio.

2.3.5.1. Datos estructurados

Los datos estructurados son datos que han sido organizados y transformados en un modelo de datos con una delimitación precisa. Los datos sin procesar se organizan en campos prediseñados que posteriormente pueden ser extraídos y leídos con facilidad a través de SQL. Las bases de datos SQL, consisten en tablas con filas y columnas, son un claro ejemplo de datos estructurados (Naeem, 2024)

Un ejemplo en el caso de estudio fue la base del modelo predictivo desarrollado utilizando Random Forest. Los datos sintéticos fueron organizados en un formato tabular, siguiendo un esquema definido con filas y columnas que representan variables clave como el historial de renovaciones, frecuencia de uso del servicio, y características demográficas de los clientes.

2.3.5.2. Datos semiestructurados

Las cifras semiestructuradas, también denominadas datos parcialmente estructurados, constituyen otra categoría entre los datos estructurados y los no estructurados. Los datos semiestructurados son una clase de información que poseen ciertas propiedades consistentes y claras,No se restringe a un diseño estricto como el requerido para las bases de datos relacionales. Las compañías combinan características organizacionales como metadatos o etiquetas semánticas con datos semiestructurados para simplificar su manejo (Naeem, 2024)

2.3.5.3. Datos no estructurados

Los datos no estructurados se caracterizan como información que se encuentra en estado absoluto sin ser procesada. Esta información resulta complicada de manejar debido a su intrincada organización y formato. Los datos no estructurados comprenden anuncios en redes sociales, chats, fotografías satelitales, información de sensores de IoT, emails y exposiciones (Naeem, 2024)

2.3.6. herramientas en la minería de datos

Las aplicaciones de minería de datos son programas informáticos que asisten a los usuarios en la identificación de patrones, tendencias y conexiones en grandes volúmenes de datos. Surgen en varias variantes, desde sencillas hasta complejas, y cubren diversas necesidades.

tabla 2
Herramientas utilizadas en minería de datos

Tipo de herramienta de minería de datos	Ventajas	Inconvenientes	Mejor para
Herramientas simples (p. ej., Excel, Tableau)	– Fácil de usar, especialmente para quienes están comenzando. – Permiten visualizar datos y detectar patrones básicos.	– Limitadas a tareas sencillas. – No incluye funciones o algoritmos avanzados.	Visualización básica de datos y análisis sencillo
Herramientas avanzadas (por ejemplo, bibliotecas de Python, R)	– Ofrecen algoritmos complejos para análisis avanzados. – Soporte para aprendizaje automático robusto y personalización.	– Requieren conocimientos de programación. – Pueden ser difíciles de dominar para principiantes.	Análisis avanzado y desarrollo de modelos personalizados.
Herramientas específicas de dominio	– Adaptadas a industrias específicas con funciones especializadas.	– Poco flexibles fuera de su ámbito especializado. – No suelen cubrir	Procesamiento de datos específico de una industria o sector.

	– Eficientes para resolver problemas concretos.	todas las necesidades generales.	
Herramientas de Big Data (por ejemplo, Apache Spark, Hadoop)	– Escalables para manejar grandes volúmenes de datos. – Capacidad para procesar datos de forma distribuida y eficiente.	– Configuración e instalación complejas. – Se necesita experiencia en informática distribuida.	Procesamiento de grandes conjuntos de datos de manera distribuida.
Herramientas de minería de texto (p. ej., NLTK, spaCy)	– Permiten extraer información de textos no estructurados. – Útiles para análisis como el de sentimientos o temas.	– Limitadas a datos textuales. – Dificultades con textos mal estructurados o ruidosos.	Procesamiento de texto y análisis de contenido basado en lenguaje.
Herramientas de minería web (p. ej., Scrapy, Beautiful Soup)	– Automatizan la extracción de contenido de sitios web. – Útiles para análisis competitivo o recopilación de datos en línea..	– Se necesita conocimiento de técnicas de web scraping. – Pueden tener problemas legales o éticos.	Extracción de datos de sitios web y análisis de contenido en línea.

Khan, 2024

2.3.7. Tipos de Metodologías en la minería de datos

Para entender mejor las diversas metodologías de minería de datos y su aplicabilidad en la minería de datos, es fundamental considerar los enfoques establecidos en la literatura donde los autores Chapman et al. (2000) .Nisbet et al. (2019), Xia y Li (2022) .Duan et al. (2020). SAS Institute (2021).Hawkins (2019).Güçlü y Yildirim (2022), nos hablan de la

comparación de cada unos de ellos y se puedo construir una comparativa respecto a lo que indican estos autores

tabla 3
Metodologías en la minería de datos

Ventajas	Descripción	Ventajas	Desventajas	Lenguajes Empleados
CRISP-DM (Cross-Industry Standard Process for Data Mining	Metodología de minería de datos que abarca seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.	- Flexible y adaptable a diferentes industrias. - Fomenta la colaboración entre equipos. - Estructura clara.	- Puede ser visto como un enfoque lineal, no siempre refleja la naturaleza iterativa del análisis.	Python, R, SQL, SAS, Java
KDD (Knowledge Discovery in Databases)	Proceso que implica la identificación de patrones significativos en grandes conjuntos de datos, incluyendo etapas de preprocesamiento y postprocesamiento.	- Énfasis en el descubrimiento de conocimiento significativo. - Abarca un amplio rango de técnicas.	- Puede ser complejo y requiere un conocimiento técnico avanzado.	Python, R, SQL, Java, Weka
SEMMA (Sample, Explore, Modify, Model, Assess)	Metodología desarrollada por SAS que se centra en la preparación de datos y la construcción de modelos.	- Enfoque práctico y orientado a la acción. - Buena para la exploración y visualización de datos.	- Menos estructurada que CRISP-DM, lo que puede dificultar su aplicación en proyectos más grandes.	SAS, R, Python
DAS (Data Analysis Solutions)	Metodología enfocada en la solución de problemas específicos mediante el análisis de datos, incluyendo la formulación de hipótesis y el análisis de resultados.	- Enfocada en la solución de problemas concretos. - Adaptable a diversas situaciones y contextos.	- Puede ser demasiado flexible, lo que dificulta la estandarización de procesos.	Python, R, SQL
FODA (Fortalezas, Oportunidades, Debilidades, Amenazas)	Análisis que ayuda a identificar los factores internos y externos que afectan un proyecto o negocio.	- Simple y fácil de entender. - Ayuda a la toma de decisiones estratégicas.	- No es específica de minería de datos; puede no abordar adecuadamente el análisis técnico.	Generalmente no requiere un lenguaje específico; puede utilizarse con herramientas de oficina.
AGILE	Metodología que enfatiza la adaptabilidad y la colaboración, permitiendo	- Fomenta la flexibilidad y la rápida adaptación a	- Puede carecer de una estructura clara	Python, R, Java, SQL, herramientas de software

	cambios rápidos en los requisitos durante el proceso de desarrollo.	cambios. - Mejora la colaboración en equipo.	y rigurosa, lo que puede dificultar el seguimiento de proyectos.	específicas según el contexto del proyecto.

Fuente: elaboración propia

2.4. De Datos a Información: Fundamentos y Estructuración para la Toma de Decisiones

En este apartado se profundizará en cómo los datos, inicialmente dispersos y sin contexto, se estructuran en conjuntos organizados llamados datasets, que permiten su análisis y extracción de patrones. Además, se abordará el concepto de almacén de datos, una herramienta esencial que centraliza y organiza esta información para facilitar su acceso y procesamiento en modelos predictivos. Esta estructura de datos es clave para el éxito de proyectos como el de Security Data, donde el análisis de datos y su conversión en información estratégica es fundamental para anticipar el crecimiento de la cartera de clientes

2.4.1. Dato

El dato simplemente detalla una porción de lo que ocurre en la realidad y no proporciona valoraciones de juicio o interpretaciones, por lo que, por consiguiente, no ofrece pautas para el comportamiento. (Carrión, s.f.)

2.4.1.1. Dato sintéticos

Los datos sintéticos son información producida de manera artificial que replica las características y distribuciones de los datos auténticos, sin incluir datos personales o delicados. Estos datos se generan mediante algoritmos y procedimientos que preservan la estructura y las propiedades estadísticas de los datos preliminares. (datos.gob.es, 2024)

2.4.2. Información

la información posee un propósito y relevancia. No solo tiene la capacidad de formar potencialmente al que la recibe, sino que está estructurada con un objetivo específico. Cuando su autor les confiere un sentido, los datos se convierten en información. Convertimos datos en información otorgándoles valor en múltiples aspectos (Carrión, s.f.)

2.5. Dataset

Como su nombre sugiere, un Dataset es un conjunto de datos estructurado de acuerdo a un sistema de almacenamiento que ofrece las directrices fundamentales para la búsqueda o el rastreo (Cáceres, 2023).

Fundamentalmente, se refiere al contenido de una tabla en una base de datos, que cuenta con diversas columnas donde se anotan datos en cada una de ellas. Estas filas pueden tomarse en cuenta.como las clasificaciones de los datos, en tanto que las columnas simbolizan las variables posibles que los conforman. La mezcla de columnas y filas es lo que se conoce como combinación Set de datos.

2.6. Almacén de datos

Conocido como almacén de datos empresarial, es un sistema que recopila información de varias fuentes en un único repositorio central y unificado para respaldar el análisis de datos, la minería , la inteligencia artificial(IBM, s.f.).

Según el autor (Biscobing, J. 2021), "Se describe como el núcleo central que recopila y asegura la información producida y recopilada por los sistemas operativos de una empresa y es un elemento esencial”

El concepto de almacén de datos o almacén de datos fue ideado por Bill Inmon y R. D. Hackathorn como "El almacén de datos se refiere a un conjunto de datos enfocados en el tema, integrados, estables e historiados, diseñados para proporcionar respaldo a procesos de asistencia en la toma de decisiones". Esta definición indica que hace referencia a una clase de bases de datos, cuyo valor reside en el respaldo que pueden proporcionar a las organizaciones Sin embargo, el mayor reto al establecer un almacén de datos consiste en identificar, de manera anticipada, qué información es necesaria y cómo deben ser estructurados (Casado, S., & Giménez, J ,2021)

2.6.1. Componentes de la arquitectura del almacén de datos

En el contexto del proyecto de Security Data, los componentes de la arquitectura del almacén de datos podrían ser esenciales para organizar, proteger y procesar los datos reales utilizados para desarrollar un modelo predictivo. Utilizando ETL, los datos podrían ser extraídos, transformados y cargados de manera eficiente en un almacén de datos.

- **ETL:** Cuando los analistas de bases de datos desean mover información desde un origen de datos hacia su depósito de datos, este es el procedimiento que emplean.
- **Los metadatos:** son información basada en datos. En esencia, detallan todos los datos guardados en metadatos facilitan la organización de sus datos para su uso, permitiéndole examinarlos para elaborar un sistema para su búsqueda. Algunos tipos de metadatos comprenden autores, fechas o lugares de un artículo, fecha de generación de un archivo, tamaño de dicho archivo, entre otros. Considere estos como los títulos de una columna en un tablero de cálculos. Los Paneles e informes.
- **Procesamiento de consultas SQL:** Este es el idioma que los analistas emplean para obtener datos de sus datos guardados en el depósito de datos. Usualmente, los almacenes de datos cuentan con tecnologías patentadas para el procesamiento de consultas

SQL, estrechamente vinculadas con la informática. Esto posibilita un desempeño muy elevado en relación a sus análisis. No obstante, es importante considerar que el precio de un almacén de datos puede empezar a incrementarse a medida que aumente la cantidad de datos y recursos computacionales SQL.

- **Capa de datos**

La capa de datos es el nivel de acceso que posibilita que los usuarios obtengan información. Es en este lugar donde usualmente se ubica un depósito de datos. Esta capa segmenta los segmentos de la información según a quién busca conceder acceso, lo que permite obtener un alto nivel de detalle en toda la organización. Por ejemplo, podría no querer brindarle a su equipo de ventas acceso a la información de su equipo de Recursos Humanos, y a la inversa.

- **Gobernanza y seguridad**

Esto tiene un vínculo con la capa de datos, dado que debe tener la capacidad de establecer políticas de seguridad y acceso exhaustivas para todos los datos de su entidad. Usualmente, los almacenes de datos poseen excelentes capacidades de seguridad y gobernanza de datos incorporadas, por lo que no se requiere una gran labor de ingeniería de datos a medida para incorporarlas. Es crucial organizar la gobernanza y la seguridad conforme se incorporan más datos en su almacén y conforme su empresa se expande. (IBM, s.f.)

2.7. Analítica de Datos: Técnicas, Desafíos y Aplicaciones en la Toma de Decisiones Predictivas

Este apartado se adentra en los distintos tipos de análisis que se utilizan en la analítica de datos, tales como la analítica descriptiva, que ayuda a entender los sucesos previos; la analítica diagnóstica, que busca comprender las causas de esos sucesos; y la analítica predictiva, que anticipa lo que podría ocurrir, apoyándose en los patrones extraídos de los datos.

Además, exploramos los desafíos que enfrenta la analítica predictiva, como la calidad de los datos, la integración de sistemas y la seguridad, aspectos fundamentales en un entorno empresarial cada vez más dependiente de la tecnología.

A través de la aplicación de estas técnicas, la analítica predictiva se convierte en una herramienta poderosa, especialmente en el contexto de Security Data, al proporcionar insights cruciales para anticipar comportamientos futuros respecto a renovaciones y deserciones de la firma electrónica. Este análisis tiene el potencial de optimizar las estrategias comerciales y mejorar la eficiencia operativa, permitiendo a las empresas mantenerse competitivas en un mercado dinámico.

2.7.1.1. Analítica de Datos

la analitia de datos implica el estudio de un grupo de datos con el objetivo de obtener una perspectiva global y las tendencias de los datos estudiados. Se examinan los datos en bruto para poder deducir una conclusión de que no están "contaminados" (Ekcit, 2023)

“La analítica de datos se presenta como una disciplina tecnológica que emplea varias herramientas y técnicas para adquirir información relevante para la toma de decisiones a partir de los datos sin procesar” (Paz, 2022)

2.7.2. Tipos de Analitica de Datos

2.7.2.1. Analitica Descriptiva

"Esta analítica implica llevar a cabo operaciones en datos históricos y exhibirlos de tal forma que se pueda entender el estado actual y futuro del negocio." (Instituto de Ingeniería del Conocimiento [IIC], s. f.)

“Se compone de un grupo de arquitecturas y tecnologías que pueden recolectar, depurar y mostrar datos con el objetivo de lograr obtener información inmediata o en tiempo real.Por lo general, es la primera interacción con los datos, intentando responder a preguntas simples.Muchas de las herramientas de este nivel están vinculadas con aplicaciones de Inteligencia Empresarial, que mediante software facilitan la creación de tableros y cuadros de indicadores en tiempo real.La analítica descriptiva tiene como objetivo examinar el pasado y dar respuesta a la interrogante ¿Qué sucedió?” (Paz, 2022)

Los entornos de inteligencia de negocios (BI) tradicionales permiten implementar alertas basadas en reglas para informar a quienes toman las decisiones sobre eventos o cambios importantes. Sin embargo, las acciones en BI siempre se definen a través de la interacción humana y son realizadas por humanos (Gopal, 2019)

2.7.2.2. Analitica Diagnóstico

Este tipo de estudio aborda la interrogante: "¿Por qué sucedió?". Este estudio descubre patrones significativos.(Ekcit, 2023)

Los análisis de diagnóstico ofrecen más valor que los análisis descriptivos y requieren un conjunto de habilidades más especializadas.(Arcitura Education Inc., n.d.)

2.7.2.3. Analitica predictiva

La analítica predictiva aplica técnicas de análisis más avanzadas como la minería de datos, machine learning , análisis estadístico , análisis predictivos con el objetivo claro de identificar patrones en los datos históricos y actuales y prever resultados futuros (IBM, s. f.).

La analítica predictiva, se basa en métodos avanzados, utiliza datos recientes e históricos para prever actividades, comportamientos y tendencias futuras (Calisaya Choque, 2021, p. 37).

2.7.2.4. Analitica Prescriptiva

El análisis prescriptivo sugiere diversas acciones potenciales orientadas a la resolución de problemas, proporcionando orientación sobre los resultados posibles antes de la toma de decisiones (Arroyo Ávila et al., 2023, p. 11).

2.7.3. Desafíos de la analítica predictiva

Esta analítica aplicada en Security Data demostró ser una solución integral para anticipar comportamientos futuros en renovación y deserción y obtener una solución informada de cómo fortalecer la posición competitiva de la empresa en el sector de servicios de firma electrónica. Para la organización se presenta estos desafíos a tener en cuenta

2.7.3.1. Datos complejos

La complejidad en la recolección, purificación y evaluación de datos representa un reto habitual en la aplicación de la analítica predictiva. Las compañías se encuentran con el desafío complicado de administrar grandes cantidades de información dispersa y variada, lo que puede complicar el proceso de análisis y perjudicar la calidad de los resultados predictivos. Es necesario encontrar soluciones que faciliten la administración y preparación de datos, tales como plataformas de unificación de datos y herramientas de análisis unificado (Alarcón, 2021)

2.7.3.2. Interpretación de los resultados

La necesidad de profesionales con competencias en ciencia de datos y análisis predictivo se incrementa, sin embargo, numerosas compañías batallan para hallar y mantener talento competente en este ámbito de alta competencia. Para enfrentar esta disparidad de competencias, las entidades están destinando recursos en programas de formación interna y alianzas. con entidades educativas y la elaboración de tácticas de reclutamiento revolucionario. Igualmente, se nota una inclinación hacia la automatización de labores reiterativas en el análisis

de datos, lo que facilita la realización de tareas repetitivas. a los expertos enfocarse en labores de mayor valor agregado a los profesionales enfocarse en labores de mayor valor añadido (Centeno, 2020)

2.7.3.3. Calidad de los datos

Esta es fundamental para la exactitud y fiabilidad de los modelos de predicción. Las compañías se topan con retos para asegurar la calidad de los datos, los cuales pueden fluctuar significativamente en cuanto a precisión, integridad y coherencia. Para optimizar la calidad de los datos, se están poniendo en marcha procedimientos de purificación y normalización de la información, además de tecnologías para la identificación y rectificación de fallos. Además, se nota un incremento en el manejo y la gobernanza de datos para asegurar su calidad y confiabilidad a través del tiempo (Sas, 2021)

2.7.3.4. Integración de los sistemas

La integración de sistemas representa un reto significativo en la aplicación de la analítica predictiva, particularmente en contextos corporativos complejos y diversos. Las compañías generalmente funcionan con una diversidad de sistemas y plataformas de datos que pueden complicar la integración y el proceso de integración. la capacidad de interoperar. Para enfrentar este reto, se están implementando plataformas de analítica integrada que facilitan la conexión y el intercambio de información entre sistemas. Adicionalmente, se nota una inclinación hacia la normalización de procesos y tecnologías con el fin de promover la integración y cooperación entre equipos y departamentos (Vera, 2023)

2.7.3.5. Escalabilidad de los modelos

La capacidad de escalar de los modelos predictivos es esencial para su eficacia.aplicación en ambientes corporativos en expansión y cambio. Los modelos deben ser capaces de gestionar grandes cantidades de datos y ajustarse conforme se incrementa la

complejidad y la magnitud de los problemas. Para enfrentar este reto, se están empleando tecnologías de computación en la nube que proporcionan recursos escalables y adaptables para el tratamiento de este desafío (Vera, 2023)

2.7.3.6. seguridad de datos

La seguridad de datos es una inquietud creciente en la era digital, particularmente en el ámbito de la analítica predictiva, donde se gestionan grandes cantidades de información delicada. Las compañías deben asegurar la privacidad, integridad y accesibilidad de los datos para resguardarse de amenazas tanto internas como externas. Para tratar este reto, se están poniendo en marcha soluciones avanzadas de ciberseguridad, tales como encriptación de datos, autenticación de múltiples factores y seguimiento constante de la seguridad (Alarcón, 2021)

2.7.3.7. Cumplimiento normativo

El acatamiento de las regulaciones es un elemento crucial en la aplicación de la analítica predictiva, particularmente en áreas altamente reguladas como la salud y las finanzas. Las compañías tienen la obligación de acatar una serie de normativas y estándares vinculados con la privacidad, la seguridad y la ética de los datos. Para enfrentar este reto, se están estableciendo políticas y procesos internos para asegurar la observancia de las regulaciones en todas las fases del ciclo de vida de los datos. Adicionalmente, se están empleando instrumentos de supervisión y auditoría para asegurar el cumplimiento de los requisitos regulatorios y minimizar el riesgo (Centeno, 2020)

2.7.3.8. Costos de implementación

Los costos de implementación pueden representar un impedimento considerable para numerosas compañías que aspiran a implementar la analítica predictiva. La inversión en tecnología, infraestructura y personal puede resultar costosa, en particular para entidades de menor tamaño y en fases tempranas.Para enfrentar este reto, se están creando soluciones de

analítica predictiva más asequibles y lucrativas, tales como servicios en la nube y plataformas de análisis basadas en abonos. Adicionalmente, se están poniendo en marcha modelos empresariales innovadores, tales como el pago por uso y el financiamiento colaborativo, que posibilitan a las compañías disminuir los gastos iniciales y compartir recursos con otros usuarios (Centeno, 2020)

2.7.4. Técnicas aplicables al análisis predictivo

En el desarrollo de un modelo predictivo para anticipar el comportamiento de los clientes de Security Data, las diversas técnicas de análisis y modelado predictivo ofrecen herramientas robustas que pueden ajustarse a diferentes necesidades y tipos de datos. A continuación, se describen algunas de estas técnicas y su relevancia potencial en el caso de estudio:

2.7.4.1. Técnicas de regresión

Uno de los modelos más comunes en el análisis predictivo es el modelo de regresión lineal. Este modelo se centra en estudiar la correlación entre una variable dependiente (o de respuesta) y un conjunto de variables independientes (o predictivas).

Este modelo es útil para anticipar números y puede aplicarse en una variedad extensa de industrias. Por ejemplo, en la industria minorista, puede emplearse para anticipar las ventas futuras basándose en factores como el precio, la promoción y la estación del año. En el ámbito financiero, se podría utilizar para anticipar el desempeño futuro de un portafolio de inversiones en base a factores financieros y económicos" (Sas, 2021)

2.7.4.2. Redes neuronales

Estos modelos poseen la habilidad de ajustarse a una variedad de problemas y resultan especialmente ventajosos en campos que administran grandes volúmenes de datos. En el ámbito tecnológico, se podría implementar para optimizar el reconocimiento de voz en

aplicaciones de asistentes digitales. En el ámbito laboral de transporte, podría emplearse para estimar la demanda de viajes basándose en datos de uso históricos" (Centeno, 2020).

Adicionalmente, una red neuronal es un modelo o programa de aprendizaje automático que toma decisiones de forma parecida al cerebro humano, empleando procesos que imitan el modo en que las neuronas se comportan biológicas para detectar fenómenos, evaluar alternativas y derivar conclusiones" (IBM, s.f.)

2.7.4.3. Árboles de decisión

Este modelo es adaptable y puede emplearse en diversas industrias para solucionar problemas de clasificación y regresión. Por ejemplo, en el sector de las telecomunicaciones, podría utilizarse para estimar la posibilidad de que un cliente rescinda su contrato basándose en su historial de uso y satisfacción. En la industria manufacturera, se podría emplear para estimar la calidad del producto basándose en variables de proceso (Centeno, 2020)

2.7.4.4. Random forest

Random forest es un algoritmo supervisado de aprendizaje automático que se emplea para resolver problemas de clasificación y regresión. Elabora árboles de decisión basados en diversas muestras y utiliza su voto predominante para determinar la clasificación y el promedio en situaciones de regresión.

Una de las particularidades más relevantes del algoritmo de bosque aleatorio es su habilidad para gestionar un conjunto de datos que incluya variables continuas, tal como sucede en la regresión, y variables categóricas, tal como sucede en la clasificación. Por lo tanto, proporciona resultados más favorables para problemas de clasificación, (Inesdi , s.f.)

2.7.4.5. Máquinas de soporte vectorial (SVM)

Este modelo resulta eficaz para problemas de clasificación tanto lineal como no lineal, lo que lo hace pertinente para diversas industrias. En la industria bancaria, podría ser utilizado

para identificar operaciones fraudulentas. En el ámbito del marketing, se podría emplear para anticipar la segmentación de los clientes basándose en sus conductas y gustos”(Centeno, 2020)

2.7.4.6. Análisis de series temporales

Según (IBM, 2024), Una serie de tiempo es un conjunto de observaciones obtenidas midiendo una sola variable regularmente durante un período de tiempo. Además, una de las principales razones para realizar análisis de series temporales es intentar predecir los valores futuros de la serie. Un modelo de serie que ilustra los valores anteriores también puede anticipar si los valores venideros se incrementarán o reducirán y en qué medida lo harán. La habilidad para efectuar estas predicciones de manera adecuada es crucial para cualquier negocio o campo de la ciencia”.

2.8. Lenguaje R en minería de datos en proyectos

En esta sección se explorará el papel fundamental que juega R en la minería de datos dentro del contexto de Security Data, con el objetivo de desarrollar modelos predictivos que anticipen el comportamiento de los clientes y fortalezcan las estrategias comerciales. Además, se destacarán ejemplos prácticos de cómo R ha sido utilizado en otros estudios para resolver problemas complejos mediante la minería de datos, mostrando sus aplicaciones en diferentes sectores, como gestionar inventarios y siniestro de tránsito.

A lo largo de esta sección, se demostrará por qué R es una herramienta clave para cualquier investigación que requiera procesamiento, análisis y visualización de datos, y cómo su versatilidad y eficiencia contribuyen al éxito de proyectos de análisis predictivo en diversos campos.

2.8.1. Lenguaje R

R es un lenguaje de programación que se utiliza principalmente para estudiar grandes volúmenes de datos y realizar cálculos estadísticos. Es un software libre y de código abierto que contiene numerosas herramientas y bibliotecas valiosas para gestionar información y elaborar gráficos (Worsley, 2024).

De software libre y de código abierto que proporciona una extensa variedad de herramientas y bibliotecas especializadas para el procesamiento, la representación y modelado de datos(Gonzalo, 2023) , Además, se destaca por su capacidad de ejecutarse en diversas plataformas como UNIX, Windows y MacOS, lo que lo convierte en un entorno versátil para la computación estadística y la visualización de datos (R Core Team, n.d.). Estas características hacen de R una opción ideal para la minería de datos, permitiendo a Security Data desarrollar modelos predictivos precisos y efectivos que contribuyan al crecimiento de su cartera de clientes.

2.8.2. lenguaje R en casos de estudios predictivos

El lenguaje R se ha consolidado como una herramienta valiosa en diversos estudios, como el de Yulissa Stefania Torres, que determinó los factores influyentes en los siniestros de tránsito en Ecuador mediante minería de datos utilizando R permitió un análisis estadístico profundo y manipulación eficiente de datos, facilitando las fases de limpieza, extracción, transformación y carga (ETL) y la aplicación de algoritmos predictivos. Los resultados de este estudio, con porcentajes de precisión global variaron entre 20,48% y 73,42% (Torres Quezada, 2020), destacan la capacidad de R para identificar patrones significativos. En el estudio realizado por Cedeño Troya y Carpio Torres utilizaron el lenguaje R, para evaluar un modelamiento predictivo en la gestión de abastecimiento en el sector ferretero de Rocafuerte, Guayaquil. Esta investigación no solo mejoró la comprensión de los factores que afectan la gestión de

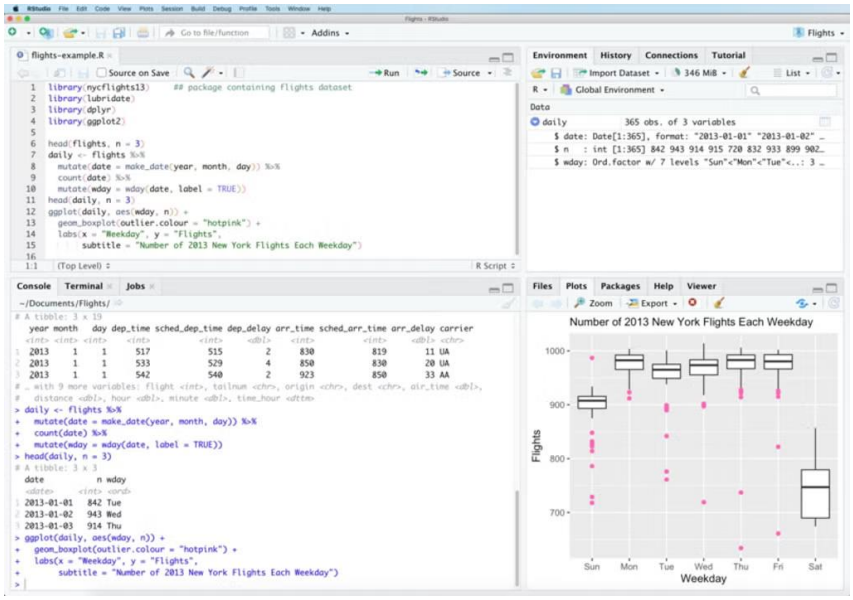
inventarios, sino que también estableció relaciones de cooperación entre participantes, potenciando la competitividad en el mercado(Cedeño Troya & Carpio Torres, 2022). Estas aplicaciones subrayan las ventajas de R en la capacidad de procesamiento, análisis y visualización de datos complejos en diversos contextos.

2.8.3. RStudio

RStudio Desktop es una aplicación desarrollada por RStudio Inc. Esta aplicación opera como un ambiente de desarrollo integrado (IDE), ofreciendo a los usuarios herramientas optimizadas para el lenguaje de programación R. Facilita la creación de scripts, la compilación de códigos, la creación de gráficos e incluso el manejo de múltiples conjuntos de datos en un ambiente altamente preparado.(RStudio, n.d.,)

Imagen 3

Interfaz de RStudio



Nota: imagen recuperada de https://rstudio-desktop.softonic.com/

3. Metodología del proceso de investigación

En este capítulo se detallan los métodos, técnicas y procedimientos empleados para el desarrollo del modelo predictivo. Se describen las fases de la metodología CRISP-DM, desde la creación de datos sintéticos, hasta el balanceo de clases y el entrenamiento del modelo con Random Forest. Además, se incluyen los pasos para la preparación, análisis y evaluación de los datos, con el objetivo de proporcionar una solución efectiva al problema planteado en la investigación.

3.1. Enfoque de la investigación

En esta investigación se adoptó un enfoque cuantitativo, utilizando técnicas de minería de datos y modelos predictivos para proyectar el crecimiento de la cartera de clientes de la empresa se servicios de firma electrónica. Este enfoque fue elegido debido a su capacidad para proporcionar resultados medibles y precisos, aplicados a una cartera de servicios basada en datos sintéticos. El objetivo principal del estudio es desarrollar un modelo predictivo que utiliza datos históricos para anticipar el comportamiento de los clientes al final del año. El desarrollo de modelos predictivos, como Random Forest, requiere métricas objetivas para evaluar su desempeño. Estos modelos, fundamentados en cálculos matemáticos y algoritmos, transforman los datos en resultados específicos y de fácil interpretación. Según Delgado, Morales y Fernández (2023), “el enfoque cuantitativo permite la recolección de datos de manera sistemática, facilitando análisis estadísticos que conducen a conclusiones replicables y generalizables” (p. 12).

El proyecto se centra en detectar patrones y tendencias a partir de datos históricos de una data que simule el comportamiento de una cartera de clientes, con el objetivo de predecir la renovación o deserción de clientes, un ámbito donde el análisis cuantitativo demuestra ser especialmente eficaz. Para ello, se empleó el lenguaje R, lo que permitió realizar un análisis estadístico y predictivo riguroso. Además, este enfoque garantizo un manejo controlado y sistemático de datos personales, cumpliendo con las normativas nacionales de protección de datos.

Si bien la incorporación del componente cualitativo pudo haber sido más valiosa para explorar percepciones o experiencias de los clientes por ejemplo, a través de entrevistas o encuestas abiertas, en este caso, la prioridad radica en generar predicciones confiables que respalden decisiones estratégicas basadas en datos que la empresa ya tiene y cumpliendo con el objetivos abordado el presente proyecto.

3.2. Alcance de Investigación

Para el presente estudio se eligió un alcance descriptivo , predictivo y explicativo con el propósito de analizar exhaustivamente el comportamiento de la cartera simulada para Security Data que cumplan como datos históricos y, posteriormente, anticipar tendencias futuras que orienten la toma de decisiones estratégicas.

En primer lugar, el alcance descriptivo permitió una visión completa y detallada de los datos actuales, facilitando el reconocimiento de patrones de retención y abandono de los clientes para obtener en función de características específicas. Como señala (Azevedo y Santos, 2022), “el análisis descriptivo permite a las organizaciones examinar y resumir datos históricos para identificar patrones y tendencias que son

fundamentales para entender la situación actual y tomar decisiones basadas en evidencia" (p. 56).

El análisis explicativo permitio interpretar y justificar las razones detrás de los hallazgos obtenidos en las fases descriptiva y predictiva. En el caso de Security Data, este enfoque facilitao

- La identificación de las variables más relevantes en el modelo predictivo, como la zona comercial y la zona empresarial.
- La explicación de por qué ciertos patrones de deserción son más frecuentes en determinadas zonas o segmentos de clientes.
- La construcción de recomendación estratégicas basadas en evidencia

Por su parte, el enfoque predictivo fue esencial para alcanzar el objetivo de esta investigación, que es desarrollar un modelo capaz de proyectar el crecimiento de la cartera de clientes y que su rendimiento precisión fue superior al 80%. En este sentido, Coussement y Van den Poel (2022) destacan que "la combinación de enfoques descriptivo y predictivo en el análisis de datos es clave para obtener una perspectiva completa de los fenómenos actuales y sus posibles proyecciones, permitiendo a las empresas formular mejores estrategias". informadas" (p. 65).

La relevancia de estos alcances para el proyecto fue fase crucial en su capacidad para convertir datos en información estratégica. Mientras que el análisis descriptivo facilitó la comprensión del contexto actual y la situación específica de los clientes, el análisis predictivo permitió proyectar posibles comportamientos futuros, dotando al proyecto de un enfoque integral que abarca tanto el presente como el futuro. Esta

combinación favorecio la implementación de estrategias fundamentadas en información cuantificable, aumentando la precisión en la toma de decisiones y brindando a Security Data una ventaja competitiva al anticipar necesidades y mejorar la fidelización en un mercado dinámico y altamente competitivo.

3.3. Delimitación de la investigación

Esta investigación se desarrolló en el contexto de la empresa Security Data, ubicada en la ciudad de Guayaquil, Ecuador, con un enfoque en dos áreas clave: la zona comercial y la zona empresarial. Security Data se especializa en servicios de firma electrónica, certificación de información y gestión de identidad digital, atendiendo a una clientela variada que incluye tanto individuos como empresas que demandan soluciones seguras y confiables en el ámbito de la autenticación de datos.

El análisis se llevó a cabo utilizando un dataset diseñado específicamente para los objetivos del estudio, compuesto por 17 variables y 22,401 observaciones. Este conjunto de datos simula una población que abarca un período de dos años previos al análisis. En términos de distribución temporal, el año 2022 representa el 42,19% de los datos, equivalente a 9,451 clientes; el año 2023 constituye el 31,44%, con 7,043 clientes; mientras que el año 2024 corresponde al 25,66%, representando 5,749 clientes.Este conjunto de datos contiene información estructurada sobre simulación del comportamiento histórico de los clientes, incluidas variables como frecuencia de uso, características demográficas, zonas geográficas y patrones de consumo. Este enfoque permitió segmentar a los clientes y modelar sus comportamientos, lo que facilitó la implementación de un modelo predictivo

El enfoque principal de esta investigación fue proyectar un comportamiento real de una cartera de clientes para Security Data, donde la base sería datos históricos para predecir

patrones de renovación y deserción. Este análisis, apoyado en el dataset, no solo permitió obtener insights estratégicos para la toma de decisiones, sino que también destacó la importancia de un enfoque cuantitativo para identificar tendencias clave en una región donde la demanda de servicios digitales continúa en aumento. Al delimitarse en un ámbito geográfico y poblacional bien definido, esta investigación contribuye al diseño de estrategias comerciales adaptadas a las dinámicas específicas del mercado local.

3.3.1. Periodo Y Lugar De La Investigación

El período de análisis se delimitó en la creación de los datos históricos desde el año 2022, 2023 y 2024, durante los cuales se recrearon datos históricos relacionados con transacciones que tendrían los clientes de Security Data. Estos datos, correspondientes a servicios caducados y por caducar, los cuales fueron anonimizados previamente para garantizar el cumplimiento de las normativas nacionales de protección de datos y recrear un comportamiento utilizado por empresas que hacen uso de datos sensibles. La selección de este período permite obtener una representación actualizada de los patrones de comportamiento de los clientes, proporcionando una base sólida para realizar predicciones precisas sobre tendencias de retención y abandono.

La recreación de los datos se realizó utilizando herramientas de generación aleatoria en el lenguaje R, garantizando que los datos sintéticos reflejaran patrones estadísticos similares a comportamientos reales, pero sin comprometer la información real de la empresa.

La delimitación en términos de ubicación, período, población y la incorporación de datos sintéticos fue crucial para asegurar que los resultados del análisis fueran representativos del entorno real en el que opera Security Data, a la vez que se evitó la utilización de cartera real que pudiera perjudicar la investigación actual. La selección de los años 2022, 2023 y 2024

como base del análisis asegura que las conclusiones y predicciones sean relevantes y aplicables a las dinámicas actuales del mercado.

3.4. Método empleado

El método empleado en esta investigación se fundamenta en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptada específicamente para el desarrollo de un modelo predictivo de renovación y deserción de clientes en Security Data. Esta metodología se complementó con herramientas de procesamiento y análisis de datos implementadas en el lenguaje R. CRISP-DM es un enfoque estándar en la minería de datos que organiza el procedimiento en seis etapas fundamentales: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelación de los datos, evaluación y despliegue. Esta metodología fue seleccionada por su flexibilidad y adaptabilidad, permitiendo que cada fase se ajuste a las necesidades específicas del proyecto y facilite la obtención de resultados precisos.

En esta investigación, se emplearon métodos empíricos y estadísticos con el propósito de analizar comportamientos que tendría la cartera real de clientes de Security Data y desarrollar un modelo predictivo para proyectar su crecimiento. Estos métodos fueron integrados dentro del marco de la metodología CRISP-DM, lo que permitió estructurar la investigación de manera organizada y orientada a la toma de decisiones basadas en datos. A continuación, se detallan los métodos empíricos y estadísticos utilizados en cada fase de CRISP-DM.

3.4.1. Métodos empíricos: observación estructurada

Durante la fase de comprensión de los datos , se empleó la observación estructurada del negocio de ventas de firma electrónica en Security Data que ayudarían a la recreación de la data correspondiente a los años 2022, 2023 y 2024. Esta observación incluyó la identificación

de precios, tipos de servicios, la duración que suelen tener las firmas electrónicas, la variedad de productos adquiridos y la duración de la permanencia de los clientes en la compañía. La observación empírica permitió extraer datos objetivos que son fundamentales para las fases posteriores del modelado, sin la necesidad de utilizar encuestas.

3.4.2. Métodos estadísticos: análisis descriptivo

En la fase de preparación de los datos , se emplearon técnicas estadísticas descriptivas para explorar y entender el comportamiento de los clientes. A través de estadísticas como frecuencias, promedios y distribuciones, se obtuvo una visión detallada de los datos históricos creados, identificando patrones y tendencias en el uso de los servicios. Este análisis descriptivo proporcionó una base sólida para el modelado, permitiendo categorizar a los clientes de acuerdo con su comportamiento y preparar los datos para su análisis predictivo. El análisis descriptivo permitió examinar los datos históricos y entender las características y tendencias en el comportamiento de los clientes. Estas métricas ofrecieron un punto de partida crucial para los modelos predictivos al revelar los factores comunes entre los clientes retenidos y los que abandonan el servicio y las variables del dataframe. ver anexo

3.4.3. Modelo predictivo con Random Forest

En esta investigación, se utilizó el algoritmo Random Forest para desarrollar un modelo predictivo con el fin de anticipar la renovación o deserción de los clientes en Security Data. Este algoritmo fue elegido por su capacidad de manejar grandes volúmenes de datos y variables múltiples, permitiendo estimar la probabilidad de retención o abandono de cada cliente, adicional este modelo estadístico ayudo a transformar los datos históricos en proyecciones sobre el comportamiento futuro de los clientes, apoyando así la toma de decisiones estratégicas.

3.4.4. Método Estadístico: Análisis Descriptivo

El método estadístico se implementó a través de análisis descriptivos para obtener una visión detallada del comportamiento de los clientes y las tendencias de renovación de los servicios. Estos análisis permitieron comprender mejor la dinámica de los datos y proporcionaron información valiosa sobre la relación entre las variables del conjunto de datos.

3.4.4.1. Análisis de la Variable "Fecha de Caducidad" y Renovación

Se realizó un análisis preliminar de la relación entre la fecha de caducidad de los servicios y la decisión de renovación de los clientes. Este análisis ayudó a identificar la influencia de la duración del servicio en la probabilidad de renovación. Se observó que la variable "**año de caducidad**" resultaba ser clave para evaluar su impacto en la renovación, ya que los servicios con una duración promedio entre uno y dos años mostraban tendencias de renovación que variaba según el tiempo.

3.4.4.2. Visualización de las Tendencias de Renovación a lo Largo del Tiempo

Se generó un gráfico para visualizar cómo la fecha de caducidad influye en la renovación de los servicios. Este gráfico mostró la tendencia general de las renovaciones a lo largo del tiempo, destacando tanto las renovaciones reales como las tendencias generales que pueden guiar las decisiones estratégicas.

3.4.4.3. Análisis de Compras por Día de la Semana en las Zonas de Seguridad Data

El análisis de compras por día de la semana se centró en comparar la actividad en las dos zonas de Security Data: la zona comercial (Albán Borja) y la zona empresarial (Las Cámaras). A través de gráficos de barras, se visualizó el número de compras realizadas en cada día de la semana, permitiendo identificar los días con mayor actividad en cada zona durante los

años 2022 a 2024. Esta visualización también ayudó a detectar patrones estacionales en las compras de servicios, permitiendo realizar comparaciones entre años.

3.5. Procesamiento y análisis de la información

3.5.1. Fuente de los datos

En este proyecto, el procesamiento y análisis de la información fueron etapas fundamentales para extraer valor de los datos históricos recreados para Security Data y obtener predicciones precisas sobre su comportamiento, el procesamiento y análisis de la información se realizaron utilizando herramientas y librerías del entorno R, como dplyr y stringi, para garantizar una preparación eficiente de los datos.

Además se generaron datos adicionales sintéticos con el fin de complementar y expandir el conjunto de datos existente. La creación de los datos sintéticos fue un paso importante para asegurar que los patrones del comportamiento de los clientes fueran representativos y, al mismo tiempo, recrear un escenario de anonimización.

3.5.2. Estandarización y control de calidad

Una vez exportada la data histórica creada con datos sintéticos en estudio, se procedió, se generó la expansión de las variables mediante la creación de nuevos atributos derivados de las variables creadas, como el año, mes y día de compra, el día de la semana de la transacción, y la duración del servicio en días aplicando las librerías dplyr, lubridate, y readxl

Estas variables proporcionaron un contexto adicional y facilitaron el análisis predictivo, permitiendo una mejor comprensión de los patrones de comportamiento de los clientes. Paralelamente, se ejecutaron controles de calidad rigurosos para identificar y gestionar valores nulos o duplicados, asegurando que los datos fueran precisos y representativos, y estableciendo una base sólida para el desarrollo de modelos predictivos confiables.

3.5.3. Generación de Datos Sintéticos

La base de datos sintética fue creada utilizando la librería stringi en R, que permitió generar cadenas de texto aleatorias para las variables como nombres, apellidos, correos electrónicos y números de teléfono. Estos datos fueron diseñados para replicar las características estadísticas y estructurales de comportamientos reales que tendría una cartera de cliente, como la variabilidad en los nombres de los clientes, los tipos de servicios contratados y las fechas de compra y caducidad.Se generaron 22,401 registros de clientes registros fueron utilizados para identificar patrones de comportamiento de los clientes, particularmente en relación con su decisión de renovar o abandonar el servicio. ver anexo

3.5.4. Datos anonimizados

En el contexto de esta investigación, la anonimización de los datos fue un paso esencial durante la fase de Preparación de los Datos del proceso CRISP-DM, garantizando el cumplimiento con las normativas de protección de datos. Dado que los datos utilizados en este estudio contienen información sensible, como nombres, correos electrónicos y números de teléfono, se implementó un proceso riguroso para reemplazar cualquier dato identificable para hacer más real el estudio.

se aplicaron técnicas de anonimización basadas en pseudonimización generando identificadores anónimos únicos para cada cliente, utilizando valores alfanuméricos como "CLIENTE_XXXX" y "Usuario_XXXX", con el fin de proteger la identidad de los usuarios ya que los identificadores generados no contienen información derivada de los datos originales, lo que asegura que no se pueda revertir el proceso de anonimización para identificar al cliente.

Además, las variables categóricas, como el tipo de servicio y el ambiente, fueron transformadas a valores binarios o numéricos, permitiendo el análisis sin comprometer la privacidad. Las variables como los correos electrónicos y números de teléfono no fueron

tomadas en cuenta, mientras que las fechas de compra y caducidad fueron transformadas a componentes temporales, como mes, día de la semana, y año, con el objetivo de mantener la utilidad de los datos sin revelar detalles específicos. Este proceso de anonimización permitió que los datos fueran utilizados en el análisis y modelado predictivo de manera ética y segura, asegurando así la integridad del estudio y el cumplimiento con las normativas de protección de datos.

3.5.5. Almacenamiento y exportación

En este proyecto, los datos utilizados para el modelo predictivo no se integraron directamente de fuentes históricas, sino que se recrearon a partir de comportamientos que tendría una cartera real estructurada para la creación de los datos históricos para Security Data. Este enfoque permitió generar un conjunto de datos sintéticos que replicó las características estadísticas y patrones, respetando siempre las normativas de protección de datos personales.

Una vez que los datos sintéticos fueron generados y procesados, se consolidaron en un formato listo para su análisis. Este conjunto incluyó variables derivadas, como componentes de fechas (año, mes, día), indicadores de tipo de servicio, zonas comerciales o empresariales, y estados del entorno. Finalmente, los datos fueron almacenados en un archivo CSV utilizando la función write.csv() en RStudio, lo que permitió exportarlos de forma estructurada y organizada. Este archivo, denominado "clientes_preparados_modelo.csv", se guardó en el sistema local y está diseñado específicamente para su uso en la creación y validación del modelo predictivo, asegurando la accesibilidad y fiabilidad necesarias para las etapas posteriores del proyecto.

3.5.6. Visualización de las primeras filas del dataset

Se visualizan las filas del dataset creado con función head(). Este paso permitió obtener una comprensión preliminar de la estructura y el contenido de las variables incluidas en el dataset. Entre las variables destacadas se encontraban atributos como las fechas de compra y caducidad, el tipo de servicio, indicadores geográficos, y otras características derivadas como los días desde la compra y el tiempo total de servicio en días.

Esta inspección inicial fue esencial para validar que los datos se cargaron correctamente desde el archivo CSV, confirmar que las transformaciones realizadas durante el procesamiento se aplicaron de manera efectiva. Además, permitió identificar posibles inconsistencias o problemas en los registros antes de proceder a fases más avanzadas de análisis y modelado predictivo. Este enfoque asegura un manejo sistemático y organizado del conjunto de datos, estableciendo una base sólida para el desarrollo del modelo.

3.5.7. Verificación de tipos de datos

Una etapa clave en el procesamiento de los datos fue la verificación de los tipos de datos para garantizar que cada variable estuviera correctamente definida según su propósito en el análisis y modelado predictivo. Para este proceso, se utilizó la función str() en RStudio, que permite examinar la estructura de un dataframe y confirmar los tipos de datos asignados a cada columna.

En este análisis, se verificó que las variables categóricas, como "TipoServi_Renovacion", "estado_ambiente", y "zona_comercial", estuvieran correctamente configuradas como factores, lo que es esencial para el modelado y la interpretación de datos dicotómicos. Asimismo, se confirmó que las variables numéricas, como "dias_desde_compra"

y "Dias_servicio", estuvieran definidas como enteros o numéricos, permitiendo realizar cálculos estadísticos y matemáticos sin inconvenientes.

Este proceso de validación no solo garantizó la integridad del dataset, sino que también permitió identificar inconsistencias, como la necesidad de convertir ciertos valores de texto o cadenas en factores. Además, aseguró que las herramientas estadísticas y de aprendizaje automático utilizadas en el proyecto funcionarán de manera óptima con los datos preparados, eliminando posibles errores asociados a discrepancias en los tipos de datos.

3.5.8. conteo de valores faltantes

En este estudio, se realizó un conteo exhaustivo de los valores faltantes en el conjunto de datos utilizando la función colSums(is.na(datos_procesados)). Esta operación permitió identificar las columnas que contenían datos ausentes y cuantificar la cantidad de valores faltantes en cada una de ellas. El conteo de valores faltantes es un paso esencial en el proceso de preparación de los datos, ya que la presencia de datos ausentes puede afectar la precisión y la validez de los análisis subsiguientes. Para asegurar la calidad de los datos, se optó por eliminar las filas con valores faltantes mediante la función na.omit(), lo que permitió mantener la consistencia del conjunto de datos sin comprometer la integridad del análisis. Este proceso es crucial para preparar los datos antes de la aplicación de técnicas de modelado y análisis predictivo, garantizando resultados más fiables y representativos.

3.5.9. Detección de valores nulos y duplicados

Una parte esencial del proceso de limpieza de datos fue la detección y el tratamiento de valores nulos y duplicados. Este paso resultó crítico para garantizar la integridad y calidad del dataset antes de proceder con el modelado predictivo.Para identificar valores nulos, se utilizó la función is.na() en combinación con colSums() en **RStudio**, que permitió calcular el número

total de valores faltantes por columna. Los registros que contenían valores nulos fueron evaluados para decidir su manejo, utilizando estrategias como la eliminación de las filas afectadas mediante la función na.omit() o la imputación de valores faltantes según las características de las variables. Este tratamiento evitó que las ausencias de datos influyeran negativamente en los resultados del análisis y en el desempeño del modelo.

En cuanto a los valores duplicados, se emplearon funciones como duplicated() para identificar registros redundantes. Estos duplicados, si existían, se eliminaron para evitar sesgos en el análisis y asegurar que cada observación fuera única y representativa. La eliminación de duplicados permitió preservar la diversidad del dataset y mejorar la precisión de las conclusiones derivadas del mismo.Este proceso de detección y corrección fue fundamental para mantener la consistencia del conjunto de datos y asegurar que los resultados obtenidos del modelo predictivo fueran confiables y relevantes para los objetivos del proyecto.

3.5.10. Exploración de variables temporales

Se exploraron varias variables temporales clave, como "Fecha de Compra" y "Fecha de Validez", las cuales proporcionan información crucial sobre la duración de los servicios contratados y la frecuencia de las compras. Se utilizaron funciones como year(), month(), day(), y wday() del paquete lubridate en RStudio para extraer componentes específicos de las fechas, como el año, mes, día de la compra, la variable de día de la semana en que se realizó la transacción, como "dias_desde_compra" (que calcula la diferencia entre la fecha actual y la fecha de compra) y "Dias_servicio" (que mide el tiempo de validez del servicio) fue crucial para identificar patrones de comportamiento, como la tendencia a renovar un servicio cerca de su fecha de caducidad. Este tipo de análisis ayuda a detectar estacionalidades o variaciones en las decisiones de los clientes relacionadas con ciertos períodos del año, lo que puede influir en la predicción del comportamiento de renovación.

Además, la visualización de las distribuciones de estas variables temporales, por ejemplo, mediante histogramas o gráficos de barras, permitió una mejor comprensión de cómo los tiempos de compra y caducidad se distribuyen a lo largo del tiempo, lo que facilitó la identificación de puntos críticos o periodos de alta o baja actividad en las renovaciones de servicios.

3.5.11. Codificación de variables categóricas

En este proyecto, se empleó una codificación adecuada de las variables categóricas para garantizar que pudieran ser utilizadas eficientemente en el desarrollo del modelo predictivo.Las variables categóricas, como "Tipo de Servicio", "Ambiente", "zona_comercial", y "zona_empresarial", contienen valores no numéricos que representan distintas categorías, por ejemplo, el tipo de servicio contratado (como "NUEVO" o "RENOVACION") o la ubicación de los clientes (como "Albán Borja" o "Las Cámaras"). Para convertir estas variables en un formato numérico, se utilizó la técnica de codificación binaria o one-hot encoding.

En la codificación binaria, las variables categóricas fueron transformadas en variables de tipo binario (0 o 1), donde cada categoría en la variable original se representó por una columna separada con valores 0 o 1. Por ejemplo, la variable "Tipo de Servicio" fue dividida en varias columnas de indicadores binarios: "TipoServi_Renovacion", "TipoServi_nuvfirma", y "TipoServi_otros", donde cada columna toma el valor 1 si la categoría es verdadera para ese registro y 0 si no lo es.La misma técnica se aplicó para la variable "Ambiente", donde se codificaron las categorías "PRODUCCION" y "PRUEBA" como columnas separadas con valores 0 o 1, y lo mismo ocurrió con las variables de "zona_comercial" y "zona_empresarial" para representar las ubicaciones de los clientes en formato binario.

Este proceso de codificación es crucial, ya que permite que las variables categóricas puedan ser interpretadas correctamente por los algoritmos de aprendizaje automático, como Random Forest, que fueron utilizados en este proyecto para hacer predicciones sobre la renovación de servicios. Además, la codificación binaria facilita el manejo de grandes volúmenes de datos y mejora la precisión del modelo al proporcionar información más clara y estructurada sobre las categorías de cada variable.

3.5.12. Escalado de variables numéricas

En este proyecto, se aplicó el escalado de variables numéricas como parte del proceso de preprocesamiento de datos, un paso crucial para asegurar que las variables con diferentes escalas y unidades de medida contribuyan de manera equitativa al modelo predictivo. Las variables como "dias_desde_compra", "Dias_servicio", y "año_caducidad" presentaban diferentes rangos, lo que podría haber generado un sesgo en los resultados, especialmente en algoritmos sensibles a la magnitud de los valores. Para abordar esto, se utilizó la estandarización de las variables numéricas, transformándolas a una distribución con media 0 y desviación estándar 1. De esta manera, el escalado permitió que el modelo predictivo procesara los datos de manera eficiente, garantizando que las diferentes magnitudes no afectarán el desempeño del algoritmo.

3.5.13. Creación de nuevas variables derivadas

En este proyecto, se llevaron a cabo diversas transformaciones de las variables originales para crear nuevas características que podrían mejorar el rendimiento del modelo y ofrecer una visión más profunda del comportamiento de los clientes.

Por ejemplo, a partir de las variables relacionadas con las fechas de compra y caducidad, se derivaron variables como "dias_desde_compra" y "Dias_servicio", las cuales

representan, respectivamente, el número de días transcurridos desde la última compra y la duración del servicio desde la fecha de compra hasta su fecha de caducidad. Estas nuevas variables proporcionan información clave sobre el comportamiento temporal de los clientes y pueden tener un impacto significativo en las decisiones de renovación de servicio.

Otra variable derivada fue "diasemana_compra", que se extrajo de la fecha de compra y representa el día de la semana en que se realizó la transacción. Esta variable es relevante para identificar patrones en las compras y podría ser útil para entender la estacionalidad o los picos de demanda en días específicos. La creación de estas variables adicionales enriqueció el conjunto de datos y facilitó un análisis más completo de los patrones de comportamiento de los clientes, mejorando así la capacidad predictiva del modelo. se puede observar en el anexo

3.5.13.1. Frecuencias de Variables Categóricas

En este estudio, se realizó un análisis detallado de las frecuencias de las variables categóricas, tales como TipoServi_Renovacion, TipoServi_nuvfirma, y las variables relacionadas con la ubicación (zona_comercial, zona_empresarial). Estas variables categóricas fueron analizadas para identificar la distribución de las clases dentro del conjunto de datos. Por ejemplo, la variable TipoServi_Renovacion, que distingue entre clientes que renovaron o no su servicio, mostró una distribución desbalanceada, con una mayor proporción de renovaciones. Este desbalance es crítico para la precisión del modelo predictivo, ya que podría sesgar los resultados. El análisis de las frecuencias permitió identificar estos desequilibrios, lo que llevó a la aplicación de técnicas de balanceo de clases, como ROSE, para asegurar que el modelo pudiera aprender de manera equitativa de ambas clases, mejorando así la capacidad predictiva y generalización del modelo.

3.5.13.2. Frecuencias de Variables Numéricas

El análisis de las frecuencias de las variables numéricas, como dias_desde_compra, Dias_servicio, y año_caducidad, permitió entender mejor la distribución y dispersión de los datos continuos. Estas variables fueron sometidas a un análisis estadístico descriptivo, donde se calcularon medidas como la media, la mediana, la desviación estándar y los cuartiles. Por ejemplo, la variable dias_desde_compra mostró una distribución sesgada, donde una gran mayoría de los clientes realizaba sus compras dentro de un rango corto de días desde la última compra. Este análisis fue clave para la preparación de los datos, ya que permitió identificar posibles outliers y ajustar los rangos de las variables para su uso en el modelado predictivo, garantizando que las variables numéricas fueran adecuadamente representadas y que no distorsionaran el rendimiento del modelo.

3.5.13.3. Frecuencias de Variables Derivadas

El análisis de las frecuencias de las variables derivadas fue crucial para comprender cómo factores temporales, como año_compra, mes_compra, y diasemana_compra, influyen en el comportamiento de los clientes. Por ejemplo, la variable año_compra mostró un patrón de mayor actividad en ciertos años, lo que sugiere una tendencia de compra estacional. Además, la variable diasemana_compra permitió identificar los días de la semana en que los clientes realizaban más compras, lo que es relevante para comprender las fluctuaciones en la demanda. Estas variables derivadas, que fueron creadas a partir de las fechas originales de compra y caducidad, proporcionaron una visión temporal del comportamiento de los clientes, permitiendo ajustar los modelos para reflejar con precisión los efectos estacionales y los comportamientos de compra en función del tiempo.

3.5.14. Balanceo de clases

Debido a que la variable objetivo, TipoServi_Renovacion, tiene un desbalance significativo entre las clases de renovación y no renovación, se aplicó un balanceo de clases utilizando el paquete **ROSE**. El balanceo de clases es una técnica importante para evitar que el modelo esté sesgado hacia la clase mayoritaria. En el código proporcionado, se utilizó la función ROSE() para generar un conjunto de datos balanceado, lo cual mejoró la capacidad predictiva del modelo al entrenarlo con datos más equilibrados.

3.5.15. Herramienta utilizadas para la analítica de datos

En este proyecto, se utilizó RStudio como el entorno de desarrollo principal, con R como el lenguaje de programación para realizar todo el análisis de datos, desde la preparación hasta la construcción de modelos predictivos. Se utilizaron librerías como dplyr y tidyr para transformar y organizar los datos, asegurando que estuvieran listos para el análisis. Además, se emplearon herramientas para la anonimización de datos, como la creación de identificadores anónimos y la codificación de variables categóricas, garantizando la protección de la privacidad en el escenario real de una organización. En la fase de modelado, R permitió la implementación de técnicas avanzadas de minería de datos, utilizando librerías como randomForest y caret para construir y entrenar modelos predictivos, específicamente un modelo de Random Forest que proyecta la renovación y deserción de clientes. ggplot2 fue utilizado para la creación de gráficos visuales que ayudaron a interpretar las tendencias y patrones en los datos, como la distribución de las renovaciones y las compras por día de la semana. También se utilizó el paquete ROSE para balancear las clases en los datos, lo que mejoró la precisión de las predicciones. En conjunto, RStudio y R proporcionaron un entorno robusto y flexible para realizar un análisis profundo, eficiente y visualmente accesible, lo que facilitó la toma de decisiones estratégicas basadas en datos reales y procesados de manera ética. ver anexo

3.5.16. Lenguaje de programación

Para el análisis de datos y desarrollo de modelos predictivos en esta investigación, se utilizó el lenguaje R con librerías especializadas para cubrir distintas fases del proceso a continuación se detallan las librerías empleadas en el proyecto:

- **R:**

Fue elegido por su potencia en análisis estadístico y modelado predictivo, así como por sus capacidades para la visualización de datos. Las principales librerías utilizadas en R fueron:

- **dplyr**: se utilizó para la manipulación y transformación de datos.
- **stringi**: Utilizada para la generación de cadenas de texto aleatorias (nombres, apellidos, etc.).
- **lubridate**: utilizado para trabajar con las fehcas.
- **readxl**: Utilizada para leer archivos de Excel (.xlsx).
- **caret**: Usada para la creación y evaluación de modelos predictivos, facilita la construcción y validación de modelos.
- **randomForest**: Utilizada para la implementación del algoritmo de **Random Forest**, empleado en el modelado predictivo.
- **pROC**: Utilizada para el análisis de curvas ROC y cálculo de **AUC**.
- **ggplot2**: se utilizó para la creación de gráficos dinámicos y visualizaciones de datos.

➢ **ROSE**: Utilizada para el balanceo de clases, especialmente en el contexto de datos desbalanceados.

➢ **tidyr**: Usada para transformar y organizar los datos en un formato adecuado para el análisis.

➢ **viridis**: Utilizada para crear paletas de colores para las visualizaciones.

➢ **reshape2**: Utilizada para reestructurar y modificar los datos (especialmente en el formato de matrices).

3.5.17. Eliminación de variables no utilizadas

En el proceso de preparación de los datos para el modelado predictivo, se procedió a la eliminación de las variables no utilizadas, un paso crucial para optimizar el conjunto de datos y mejorar la precisión del modelo. Para ello, se utilizó la función select() de la librería dplyr, la cual permitió seleccionar únicamente las columnas relevantes para el análisis, descartando aquellas que no aportaban valor al objetivo de la investigación. Además, se empleó la función setdiff() para identificar las variables no esenciales y eliminarlas del conjunto de datos.

En este caso, las variables que se eliminaron fueron aquellas que no aportaban información relevante para el análisis del comportamiento de los clientes y su renovación de servicio. Se eliminaron las columnas relacionadas con identificadores como id_anonimo y cliente_anonimo, Telefono1 y Correo, ya que no son necesarias para el modelado predictivo y su inclusión podría generar ruido en el análisis. Además, las variables de identificación como id_anonimo y cliente_anonimo, aunque útiles para la anonimización de los datos, no aportan valor al análisis predictivo. Este proceso contribuyó a reducir la complejidad del modelo y a evitar el uso de información irrelevante, garantizando que solo las variables significativas fueran utilizadas en el análisis. La eliminación de variables no relevantes es una técnica

estándar en el preprocesamiento de datos, que mejora la eficiencia y la interpretabilidad de los modelos predictivos.

3.5.18. Entrenamiento y prueba del modelado

En este proyecto, se utilizó el algoritmo Random Forest para la construcción de un modelo predictivo que permitiera anticipar la renovación de servicios entre los clientes de Security Data. Este proceso de modelado se ubicó dentro de la fase de Modelado de la metodología CRISP-DM. El conjunto de datos previamente procesado y preparado fue dividido en dos subconjuntos: uno de entrenamiento y otro de prueba. El modelo se entrenó utilizando el 70% de los datos para ajustar los parámetros del modelo y aprender patrones significativos relacionados con la renovación de los servicios, mientras que el 30% se empleo para evaluar su precisión

3.5.18.1. Evaluación del Modelado en RStudio

La evaluación del modelado en RStudio se lleva a cabo en la fase de Evaluación dentro de la metodología CRISP-DM. En esta fase, se analiza la efectividad del modelo desarrollado durante la fase de Modelado para garantizar que cumpla con los objetivos del proyecto. Es un paso crítico para determinar si el modelo tiene un buen rendimiento y si es adecuado para ser desplegado en la toma de decisiones. Se utilizan diversas métricas de rendimiento, como la matriz de confusión, el recall, la precisión, el F1-score, la curva ROC y el AUC,Además, durante esta fase, se realiza la validación cruzada y se ajustan los parámetros del modelo si es necesario. La evaluación asegura que el modelo final sea robusto, confiable y útil para la toma de decisiones estratégicas, alineándose con los objetivos del negocio establecidos en la fase de Comprensión del Negocio de la metodología CRISP-DM.

3.5.18.2. Visualización de la Importancia de las Variables

Una parte importante del proceso fue identificar las variables más importantes para el modelo. Random Forest proporciona la capacidad de calcular la importancia de las variables a través de métricas como la reducción de impureza de Gini. En el código, se utilizó la función varImpPlot() para visualizar qué variables contribuyeron más a la predicción del modelo, ayudando a comprender los factores clave que afectan la decisión de renovación de los clientes.

3.6. Elementos metodológicos específicos para TI

- **Diseño del proyecto**

Tipo de proyecto: Desarrollo de un modelo predictivo de retención y abandono de clientes.

Enfoque metodológico: CRISP-DM

Modelo de desarrollo:.Random forest

- **Recopilación de información**

Fuentes de datos: Datos sintéticos creados para el caso de estudio Security Data (2022 -2023-2024), replicación de datos sintéticos

Herramientas de compilación: RStudio

Procedimientos de recopilación: Importación y consolidación de archivos mensuales de cada año.

- **Desarrollo y diseño**

Planificación: Definición de fases según CRISP-DM.

Requisitos del sistema: Sistema operativo Windows 10 o superior, 8 GB de RAM, procesador 2,5 GHz.

Diseño del sistema/producto: Construcción de modelos predictivos usando algoritmos como Random Forest.

Tecnologías y herramientas: RStudio, librerías como pandas, dplyr, randomForest, caret, pROC , dplyr , smotefamily , ggplot2 , ROSE , tidyr, viridis, reshape2

- **Análisis de datos**

Métodos de análisis: Exploratorio y predictivo.

Herramientas de análisis: RStudio

Interpretación de resultados: Visualización de patrones renovación y abandonó del servicio de la firma electrónica de duración de uno y dos años presentación en gráficos y tablas.

- **Evaluación del proyecto**

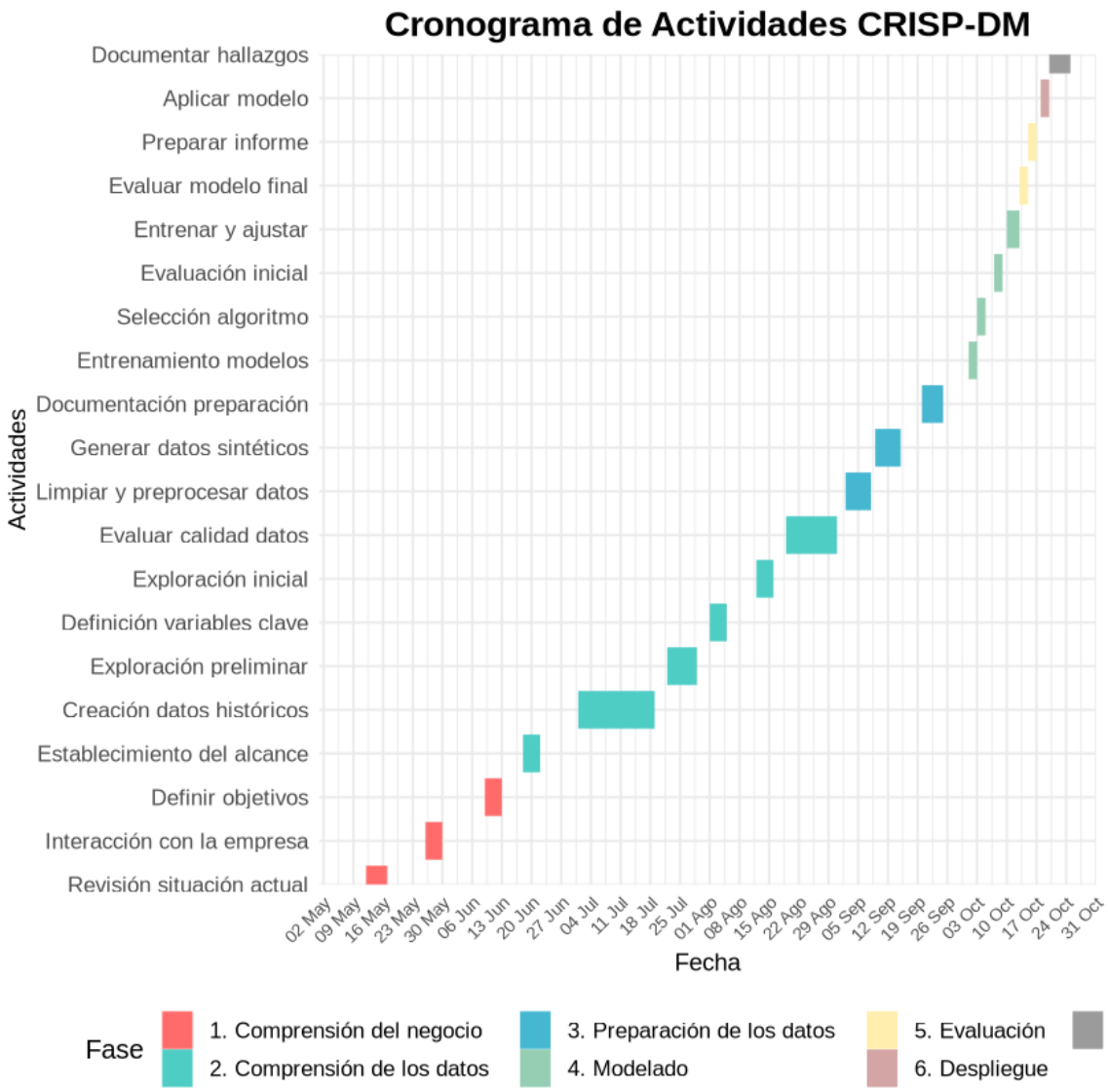
Criterios de evaluación: Precisión y recuerdo del modelo predictivo.

Instrumentos de evaluación: RStudio .

- **Cronograma de actividades**

A continuación, se presenta el cronograma detallado para las actividades planificadas en el desarrollo del modelo desarrollado en el caso de estudio

Imagen# 4
Cronograma de actividades



Fuente:elaboración propia en Rstudio

4.**Resultados**

La implementación de los pasos detallados en el capítulo de Metodología de este proyecto posibilitó la consecución de los siguientes resultados, logrando las metas concretas establecidas en el capítulo pertinente. Estos hallazgos se muestran en las secciones de este capítulo.

4.1. Creación de los datos

Para la creación del conjunto de datos utilizado en este análisis, se implementó un enfoque sistemático de generación de registros sintéticos a través de herramientas avanzadas

de manipulación de datos en R. Inicialmente, se utilizó la función data.frame() para estructurar un conjunto de datos con 22,401 registros y múltiples atributos, tales como RUC, nombre completo, fechas de compra y validez, entre otros. Las funciones de la librería stringi, como stri_rand_strings(), permitieron generar valores aleatorios para campos clave, asegurando una estructura coherente y representativa.Además, mediante la función sample(), se asignaron valores categóricos como el tipo de servicio, el ambiente y la zona ya sea comercial o empresarial,. Para las fechas de compra y validez, se definieron intervalos lógicos y se emplearon operaciones aritméticas sobre objetos tipo fecha, simulando periodos de vigencia de uno o dos años.

Finalmente, utilizando la librería dplyr, se realizaron transformaciones adicionales, como el formateo de las fechas al estándar "YYYY-MM-DD" y la generación de nombres completos combinando listas predefinidas de nombres y apellidos. Este proceso culminó en un dataset bien estructurado, diseñado a partir de patrones observados en los datos reales de la empresa, que sirvió como base para el desarrollo y validación de modelos analíticos.

Imagen 5
Creación de los datos

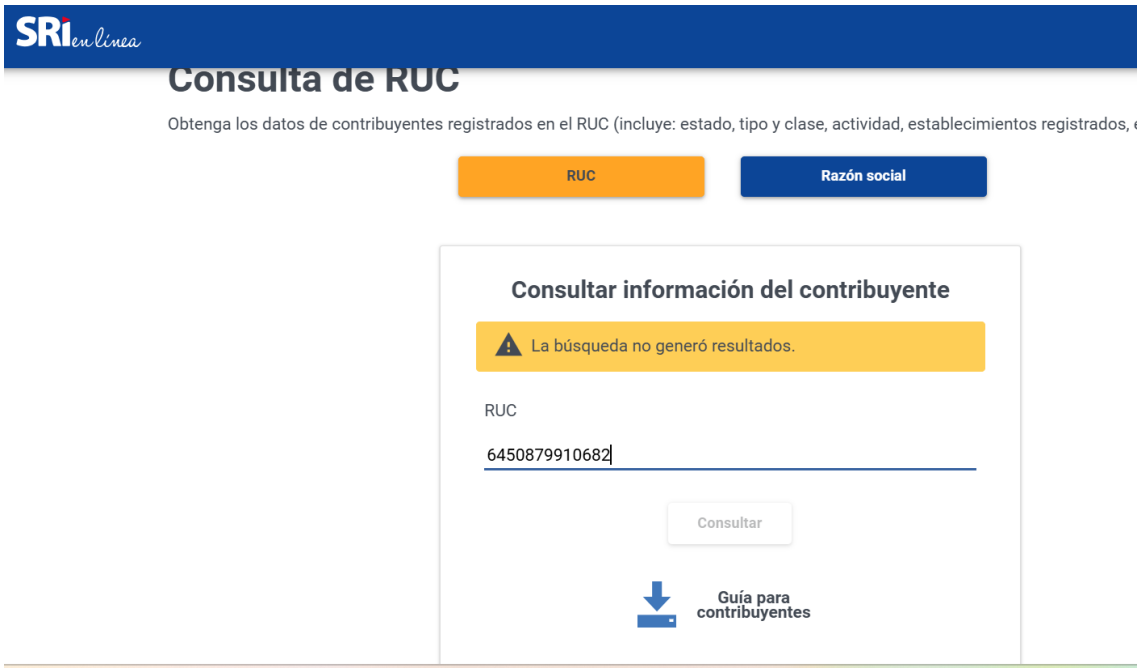
Filter

	RUC	Nombre	FechaCompra	FechaValidez	Telefono1	Correo
1	6450879910682	CASTRO ALVAREZ JORGE	2024-10-12	2025-10-12	9221519319	ipafuun@gmail.com
2	9857219331958	FERNANDEZ CASTILLO JORGE	2024-05-24	2026-05-24	8934960695	ussvvcg@gmail.com
3	5041357680787	GIL SERRANO DIANA	2022-11-08	2024-11-07	4484820484	cjcttwe@gmail.com
4	5991527293162	MARTINEZ MORALES CAROLINA	2022-05-10	2024-05-09	4623180269	ohradac@gmail.com
5	6188043509373	MARTIN BLANCO FERNANDA	2024-03-11	2026-03-11	0685602835	htptbsl@gmail.com
6	7069122270870	GARCIA SERRANO MANUEL	2022-12-27	2024-12-26	4926197969	ttidvot@gmail.com

Fuente : elaboración propia

En los resultados obtenidos podemos observar la creación de un identificador como el ruc , nombres la fecha de compra y fecha de validez , teléfonos y correos entre otros , lo cual fueron datos 100% de datos sintéticos

Imagen 6
Comprobación de dato sintético



Fuente : elaboración propia

La imagen representada indica que el dato del cliente creado es totalmente sintético y el cual no hay comprobación que el dato es real

4.1.1. Visualización general de los primeros datos

El conjunto de datos inicial está compuesto por 22,401 observaciones distribuidas en 9 variables, diseñadas para representar características clave de los clientes y sus transacciones. Para realizar una inspección preliminar y evaluar la calidad de los datos, se utilizó la función summary(), cuyos resultados se almacenaron en un objeto denominado resumen. Este análisis

permitió obtener una visión general de las distribuciones, formatos y posibles inconsistencias en las variables.

Imagen 7

Datos iniciales

```
> resumen
     RUC                Nombre           FechaCompra        FechaValidez        Telefono1
 Length:22402       Length:22402       Length:22402       Length:22402       Length:22402
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
    Correo          TipoServicio         Ambiente          Ubicacion
 Length:22402       Length:22402       Length:22402       Length:22402
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Fuente : elaboración propia

Descripción de las Variables

- **RUC:** Código único que identifica al cliente (persona natural o jurídica), conocido como Registro Único de Contribuyentes.
- **Nombre:** Nombre completo del cliente o razón social.
- **FechaCompra:** Fecha en la que se realizó la compra del servicio o producto (por ejemplo, "2023-01-9" indica una compra realizada el 9 de enero de 2023).
- **FechaValidez:** Fecha de vencimiento del servicio contratado (por ejemplo, "2024-01-9" indica una vigencia de un año, con vencimiento el 9 de enero de 2024).
- **Telefono1:** Números de contacto del cliente.
- **Correo:** Direcciones de correo electrónico del cliente.
- **Tipo de Servicio:** Clasificación del servicio contratado. Ejemplo:

NUEVO: Adquisición de una nueva firma digital.

RENOVACION: Renovación de una firma digital existente.

- **Ambiente:** Define el entorno del servicio, como:

PRODUCCIÓN: Indica que el servicio está operativo en un entorno real.

- **Zona:** Representa la ubicación donde se emitieron las firmas digitales (por ejemplo, una oficina específica como alban borja o las camaras).

4.2. Transformación y anonimización de datos para análisis predictivo

En el proceso de ampliación y enriquecimiento del conjunto de datos, se generaron nuevas características derivadas de las variables originales, logrando un total de 17 nuevas variables y 22,401 observaciones.

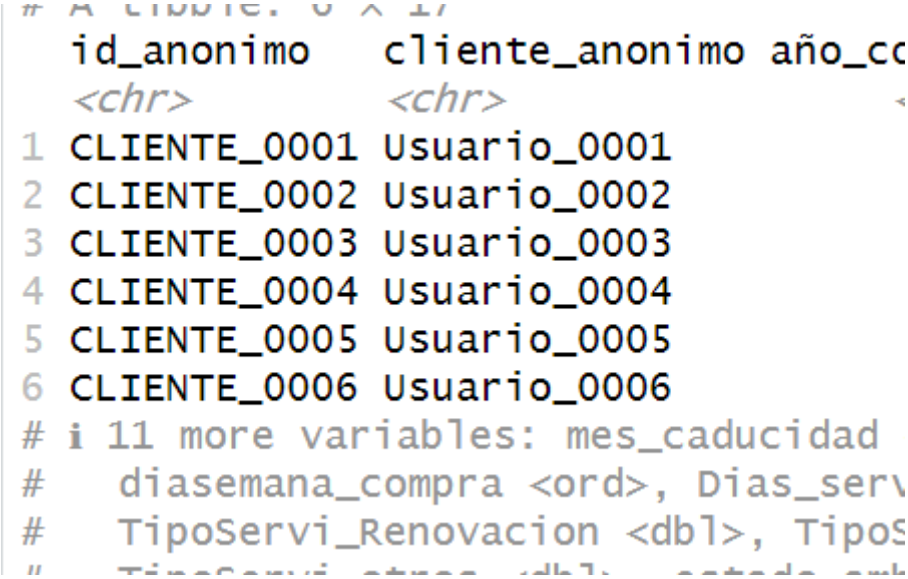
Para llevar a cabo la transformación y anonimización de los datos, se utilizaron diversas funciones de las librerías dplyr, lubridate y readxl en el lenguaje R, las cuales facilitan el manejo, manipulación y análisis de datos. La librería dplyr se empleó para realizar operaciones como la creación de nuevas columnas mediante la función mutate(), la selección de variables con select() y el filtrado de datos. Por su parte, lubridate permitió extraer componentes de fecha, como años, meses y días, además de calcular intervalos temporales y determinar el día de la semana. Finalmente, la librería readxl fue utilizada para la carga inicial del archivo Excel que contenía los datos creados . Estas herramientas, combinadas, garantizaron un procesamiento eficiente y estructurado del conjunto de datos, preparándolo para análisis posteriores.

4.2.1. Anonimización de datos

Para proteger la identidad de los clientes, se generaron identificadores únicos y anónimos. La variable id_anonimo asigna un identificador único a cada cliente con el formato "CLIENTE_0001", derivado del número de fila del conjunto de datos. De manera similar, la variable cliente_anonimo crea un identificador alternativo con el prefijo "Usuario_", proporcionando otra forma de anonimización. Estas transformaciones aseguran la confidencialidad de los datos originales mientras se mantiene la trazabilidad para análisis posteriores.

Imagen #8

Datos anonimización basadas en pseudonimización



```
# A tibble: 6 x 17
  id_anonimo   cliente_anonimo año_cc
  <chr>        <chr>
1 CLIENTE_0001 Usuario_0001
2 CLIENTE_0002 Usuario_0002
3 CLIENTE_0003 Usuario_0003
4 CLIENTE_0004 Usuario_0004
5 CLIENTE_0005 Usuario_0005
6 CLIENTE_0006 Usuario_0006
# i 11 more variables: mes_caducidad
#   diasemana_compra <ord>, Dias_serv
#   TipoServi_Renovacion <dbl>, TipoS
#   TipoServi_otros <dbl>, estado_amb
```

Fuente: elaboración propia

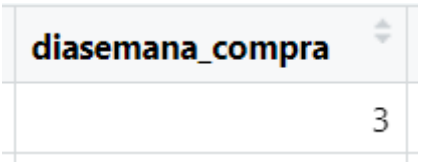
4.2.2. Componentes de Fecha

Se desglosaron las fechas en componentes individuales para facilitar su análisis. Las variables año_compra, mes_compra y dia_compra extraen, respectivamente, el año, mes y día de la variable FechaCompra, permitiendo analizar patrones temporales como temporadas de alta o baja actividad. De manera similar, año_caducidad, mes_caducidad y dia_caducidad hacen lo propio con la fecha de caducidad (Fecha de Validez), ayudando a identificar tendencias relacionadas con la expiración de servicios.

4.2.3. Días de la Semana

La variable **diasemana_compra** identifica el día de la semana en que se realizó cada compra (por ejemplo, "Lunes" o "Martes"). Esta información resultó útil para analizar patrones de comportamiento de los clientes en días específicos, lo que podría influir en estrategias comerciales o de atención.

Imagen #9
ejemplo del dato de la variable diasemana_compra



diasemana_compra
3

fuente:elaboración propia

El dato de resultado indica que la compra de la firma electrónica fue realizado un martes ya que empieza con 0 para domingo

4.2.4. Métricas Temporales

Se calcularon métricas relacionadas con el tiempo para analizar la duración del servicio y su vigencia. La variable Dias_servicio determina el número de días entre la compra (FechaCompra) y la caducidad (FechaValidez), indicando la duración del servicio contratado. Por otro lado, dias_desde_compra calcula el tiempo transcurrido desde la fecha de compra hasta el día actual (Sys.Date()), lo cual es útil para medir la antigüedad de los servicios contratados.

4.2.5. Indicadores del Tipo de Servicio

Para clasificar los servicios, se generaron variables binarias. TipoServi_Renovacion toma el valor de 1 si el servicio es "RENOVACION" (renovación de firma digital) y 0 en caso contrario. De manera similar, TipoServi_nuvfirma indica si el servicio es "NUEVOSD" (adquisición de nueva firma digital), mientras que TipoServi_otros identifica los servicios que no pertenecen a estas categorías principales. Estas variables facilitan el análisis segmentado y la predicción según el tipo de servicio.

4.2.6. Estado del Entorno

La variable estado_ambiente es una variable binaria que distingue el entorno en el que opera el servicio. Toma el valor de 1 si el ambiente es "PRODUCCION" (activo en un entorno real) y 0 si es "prueba". Este indicador es esencial para evaluar el impacto de los servicios operativos frente a los de prueba.

4.2.7. Indicadores de la Zona

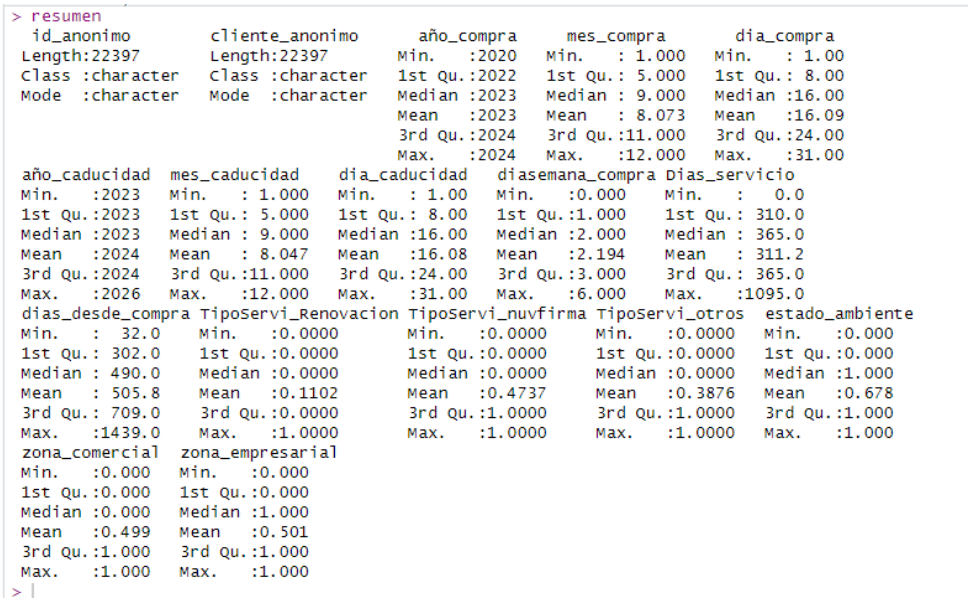
Se crearon indicadores binarios para clasificar la ubicación geográfica de los servicios. zona_comercial toma el valor de 1 si la zona es "alban borja" y 0 en caso contrario. Asimismo, zona_empresarial toma el valor de 1 si la zona es "camaras". Estas variables permiten un análisis segmentado basado en la ubicación geográfica de los clientes.

4.2.8. inspección general de los nuevos datos

Se realizó un análisis exploratorio de los nuevos datos generados con el fin de entender el comportamiento de las variables que afectan la cartera de clientes en una empresa proveedora de servicios de firma electrónica.

A continuación, se ofrece una explicación detallada de los resultados obtenidos de la aplicación de las variables sobre los datos procesados.

Imagen 8
Análisis Descriptivo de las Variables Clave



```
> resumen
  id_anonimo        cliente_anonimo      año_compra     mes_compra       dia_compra
 Length:22397       Length:22397       Min.   :2020   Min.   : 1.000   Min.   : 1.00
 Class :character   Class :character   1st Qu.:2022   1st Qu.: 5.000   1st Qu.: 8.00
 Mode  :character   Mode  :character   Median :2023   Median : 9.000   Median :16.00
                                       Mean   :2023   Mean   : 8.073   Mean   :16.09
                                       3rd Qu.:2024   3rd Qu.:11.000   3rd Qu.:24.00
                                       Max.   :2024   Max.   :12.000   Max.   :31.00
 año_caducidad  mes_caducidad    dia_caducidad   diasemana_compra Dias_servicio
 Min.   :2023   Min.   : 1.000   Min.   : 1.00   Min.   :0.000    Min.   :   0.0
 1st Qu.:2023   1st Qu.: 5.000   1st Qu.: 8.00   1st Qu.:1.000    1st Qu.: 310.0
 Median :2023   Median : 9.000   Median :16.00   Median :2.000    Median : 365.0
 Mean   :2024   Mean   : 8.047   Mean   :16.08   Mean   :2.194    Mean   : 311.2
 3rd Qu.:2024   3rd Qu.:11.000   3rd Qu.:24.00   3rd Qu.:3.000    3rd Qu.: 365.0
 Max.   :2026   Max.   :12.000   Max.   :31.00   Max.   :6.000    Max.   :1095.0
 dias_desde_compra TipoServi_Renovacion TipoServi_nuvfirma TipoServi_otros  estado_ambiente
 Min.   :  32.0    Min.   :0.0000       Min.   :0.0000     Min.   :0.0000   Min.   :0.000
 1st Qu.: 302.0    1st Qu.:0.0000       1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.000
 Median : 490.0    Median :0.0000       Median :0.0000     Median :0.0000   Median :1.000
 Mean   : 505.8    Mean   :0.1102       Mean   :0.4737     Mean   :0.3876   Mean   :0.678
 3rd Qu.: 709.0    3rd Qu.:0.0000       3rd Qu.:1.0000     3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :1439.0    Max.   :1.0000       Max.   :1.0000     Max.   :1.0000   Max.   :1.000
 zona_comercial  zona_empresarial
 Min.   :0.000   Min.   :0.000
 1st Qu.:0.000   1st Qu.:0.000
 Median :0.000   Median :1.000
 Mean   :0.499   Mean   :0.501
 3rd Qu.:1.000   3rd Qu.:1.000
 Max.   :1.000   Max.   :1.000
>
```

Fuente : elaboración propia

- **Identificación de los Clientes:**

El conjunto de datos tiene un total de 22,397 registros. Los identificadores de cliente son cadenas de texto (tipo character) tanto para el id_anonimo como para el cliente_anonimo.

- **Año, Mes y Día de Compra:**

Las compras se registraron entre los años 2020 y 2024, con un rango de meses de enero a diciembre y días que van del 1 al 31.

El promedio de compra se concentra en los meses de agosto (8.07) y 16 días del mes (promedio de 16.09), con una tendencia hacia los primeros meses en la mayoría de las compras (primer cuartil en mayo y septiembre, tercer cuartil en noviembre).

➢ **Datos Relacionados con la Caducidad del Servicio:**

Las fechas de caducidad de los servicios oscilan entre los años 2023 y 2026, con los días de caducidad distribuidos entre el primer (8) y tercer cuartil (24).

El promedio de días desde la compra hasta la caducidad de los servicios es de aproximadamente 311 días, lo que indica que la mayoría de los servicios tienen una duración de alrededor de un año.

➢ **Día de la Semana de Compra:**

El análisis muestra que la mayoría de las compras se realizaron entre los días 1 y 6 de la semana, con un promedio de compras en el día 2.

➢ **Tipo de Servicio y Renovación:**

Los servicios ofrecidos están clasificados en tres categorías: Renovación, Nueva Firma, y Otros.

La mayoría de las compras están asociadas a los servicios de nueva firma y otros, con un porcentaje relativamente bajo de renovaciones (media de 11.02% para renovaciones).

➢ **Estado del Ambiente de Servicio:**

El estado del ambiente (estado_ambiente) tiene una distribución binaria, donde el valor 1 indica "ambiente de producción" y 0 "ambiente de pruebas". Aproximadamente el 67.8% de los registros están en ambiente de producción, lo que refleja un alto nivel de actividad en servicios operativos.

➢ **Zonas Comerciales y Empresariales:**

En cuanto a la zona comercial (zona_comercial), aproximadamente el 50% de los registros están distribuidos entre dos zonas. Además, el 50.1% de los servicios están asociados con una zona empresarial, lo que indica una distribución equilibrada entre las zonas comerciales y empresariales.

4.3. Análisis exploratorio mediante modelos estadísticos

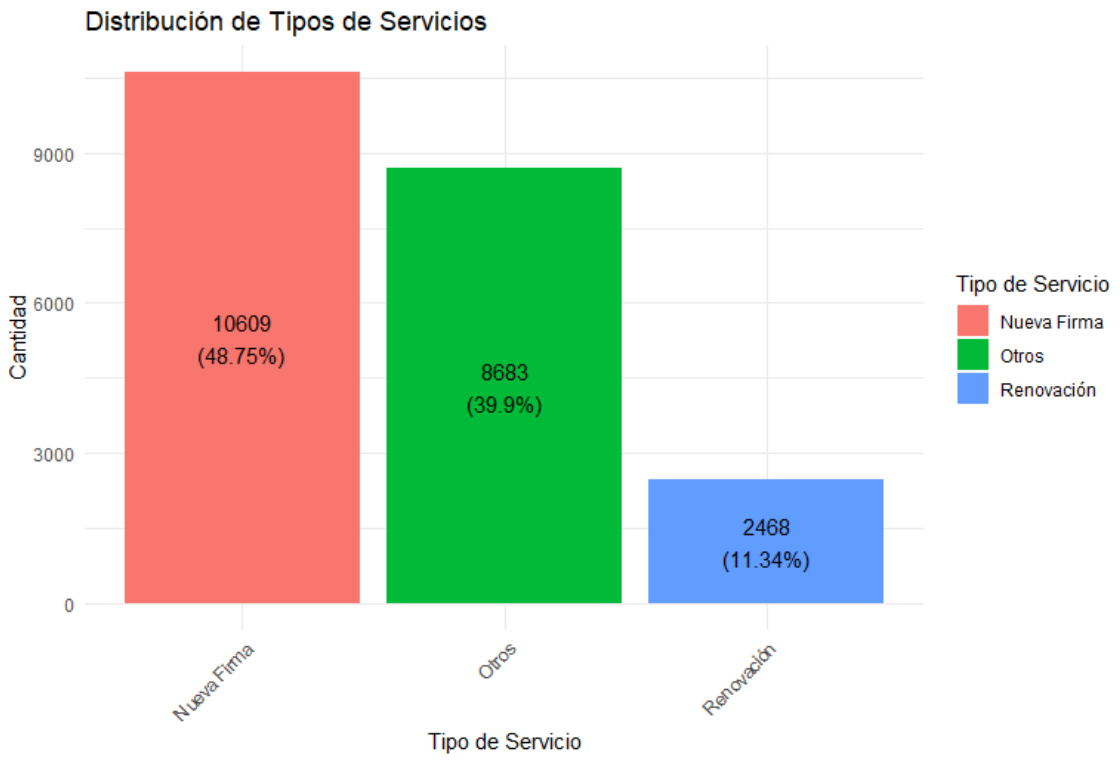
4.3.1. Examinación de variables

La examinación de variables fue un paso crucial para comprender la estructura de los datos y las relaciones entre las diferentes características que impactan el comportamiento de los clientes. A lo largo de este análisis, se han examinado diversas variables que son fundamentales para entender el patrón de compra de los clientes de la empresa proveedora de servicios de firma electrónica. A continuación, se describen los resultados analizados.

4.3.1.1. Distribución de Tipos de Servicio

Al analizar la distribución de los servicios solicitados, observé que la mayoría de los clientes optaron por los servicios de Nueva Firma y otros, con un menor porcentaje de clientes solicitando renovaciones.

Imagen 9
Distribución Tipos de servicios



Fuente : elaboración propia

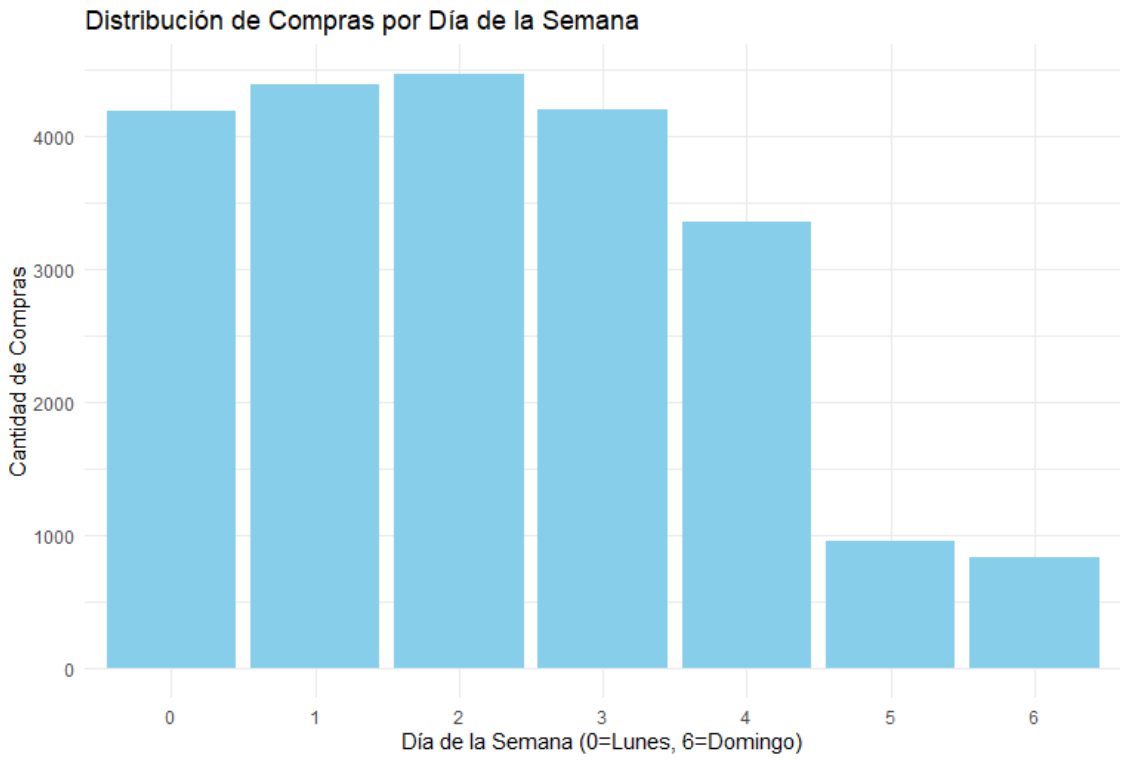
Este gráfico me permitió observar claramente cómo se distribuyen los diferentes tipos de servicios y cuál es la proporción de cada uno dentro del total de servicios solicitados , se ve que mayormente hay adquisición de una nueva firma pero muy poco caso de adquisición de renovaciones.

4.3.1.2. variable Día de la Semana

El análisis de las compras según el día de la semana me permitió identificar patrones en el comportamiento de los clientes, tales como la tendencia a comprar más ciertos días de la semana que otros.

Se analizó la distribución de las compras a lo largo de la semana, observando que los días de mayor actividad fueron los Lunes, Martes y Miércoles, mientras que Viernes y Sábado mostraron una caída significativa en la cantidad de compras.

Imagen 10
Distribución de la variable diasemana_compra



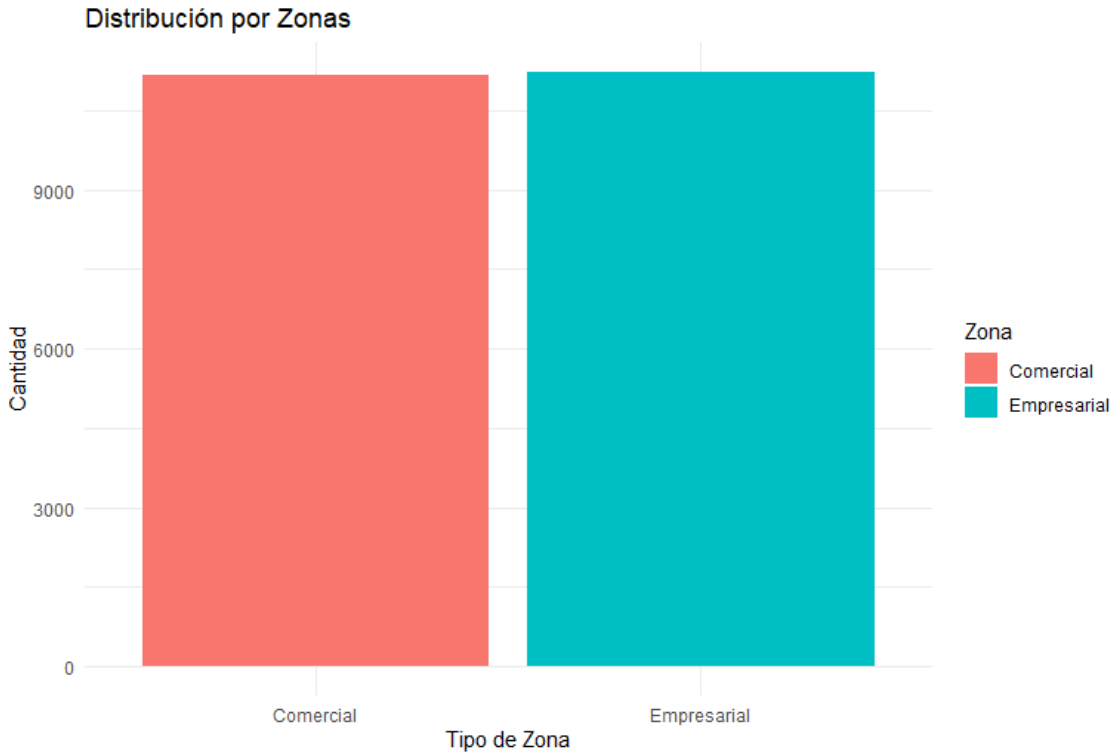
Fuente : elaboración propia

Este gráfico me permitió identificar que la mayoría de las compras se concentran al inicio de la semana, lo que puede indicar un patrón de compra más frecuente al principio de la jornada laboral.

4.3.1.3. Distribución de la variable Zonas Geográficas

En el gráfico de Distribución por Zonas Geográficas no se observó una gran diferencia, indicando que la cantidad de compras entre las Zonas Comerciales y Zonas Empresariales está bastante equilibrada. lo que resultó relevante para interpretar que en la distribución geográfica, ambas zonas tienen una importancia similar en cuanto al número de compras realizadas, lo que podría sugerir que factores adicionales, como las características de los clientes o los servicios específicos solicitados, podrían tener un mayor impacto en las decisiones de compra que la zona geográfica en sí .

Imagen 11
Distribución de la variable zona_comercial y zona_empresarial



Fuente : elaboración propia

4.3.1.4. Análisis de la variables Días desde la Última Compra

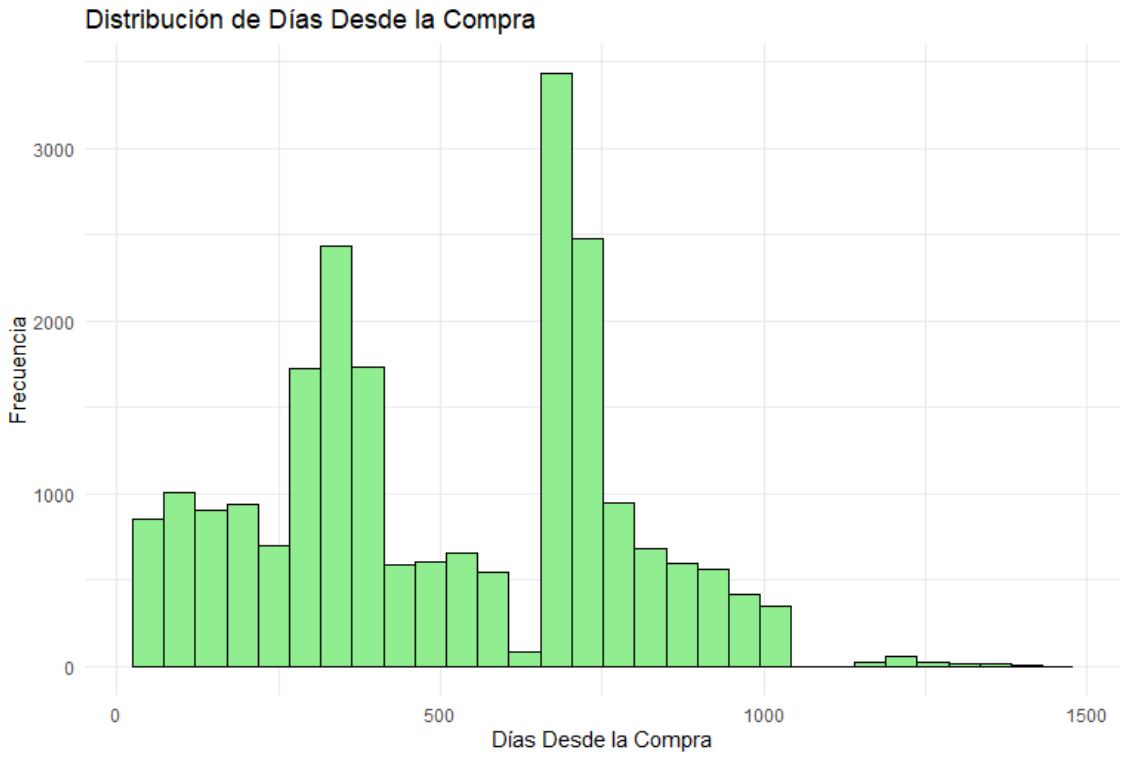
Este indicador es útil para medir la frecuencia de las compras de los clientes. Un análisis de esta variable puede ayudar a prever la probabilidad de una nueva compra, lo que es fundamental para mantener una cartera de clientes activa**.**se pudo identificar que el cliente más reciente tiene 32 días desde su compra , el cliente más antiguo tiene 1,439 días desde su compra , la media (promedio) es de 505.8 días y la mediana es de 490 días

La distribución muestra que:

- La mayoría de los clientes (35.99%) se encuentran en el rango de 595-876 días
- El segundo grupo más grande (29.18%) está entre 313-595 días

- El 27.13% son relativamente nuevos, con menos de 313 días
- Solo un 7.69% tienen más de 876 días desde su compra
- Desviación estándar: 267.1 días
- Total de registros analizados: 22,402 clientes

Imagen 12
Distribución de la variable dias_desde_compra



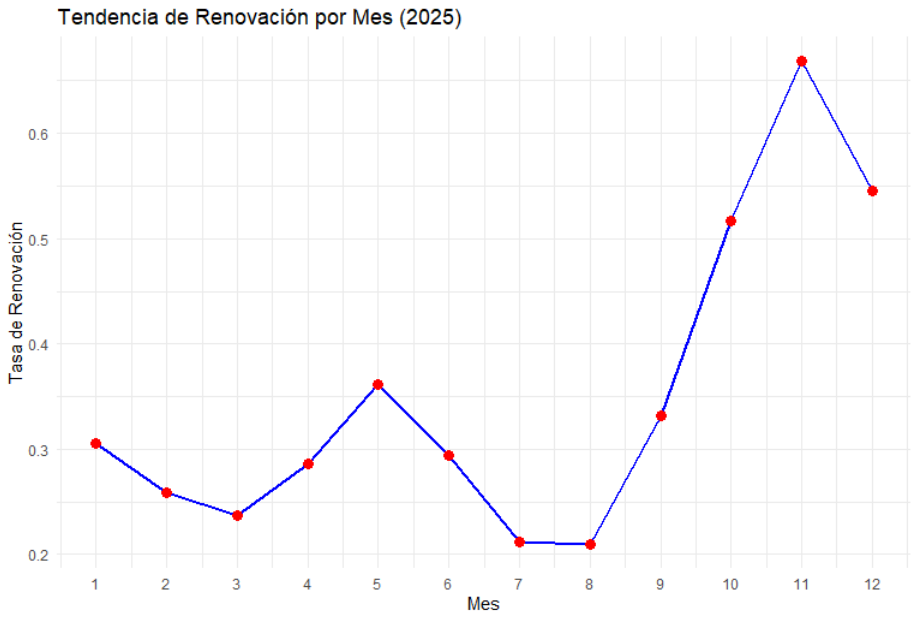
Fuente : elaboración propia

Esta distribución sugiere que la mayoría de los clientes han adquirido sus firmas digitales en los últimos dos años, con una concentración particular entre los 595 y 876 días (aproximadamente entre 1.5 y 2.5 años).

4.3.2. Modelos estadísticos

En esta etapa del estudio, se implementaron modelos estadísticos tradicionales para analizar y comprender la dinámica entre las variables del conjunto de datos. A diferencia de los modelos predictivos basados en minería de datos, estos modelos se utilizaron exclusivamente para evaluar la relación entre las variables y la variable objetivo, así como para identificar patrones generales en los datos. El análisis se realizó utilizando técnicas estadísticas y de visualización para explorar el comportamiento de la tasa de renovación de los clientes a lo largo del año 2025 , Los resultados obtenidos para la tasa de renovación mensual por mes del año 2025 se presentan en la siguiente gráfica

Imagen 13
Tendencia de renovaciones



Fuente : elaboración propia

El análisis de la tasa de renovación de los clientes durante el año 2025 muestra una clara variabilidad mensual en los porcentajes de renovación. En los primeros meses del año, enero y febrero, las tasas de renovación son relativamente bajas, con un 30.5% y un 25.9%

respectivamente. Esto sugiere que, en el comienzo del año, los clientes son menos propensos a renovar sus servicios, lo que podría estar relacionado con factores como el cambios en las necesidades de los clientes.

A medida que avanza el año, marzo y abril también muestran tasas moderadas de renovación, con un 23.7% y un 28.7%, lo que indica una ligera mejora respecto al inicio del año. Sin embargo, los porcentajes aún están por debajo del 30%, lo que podría sugerir que el comportamiento de los clientes aún no es del todo favorable para renovaciones en estos meses.

El mes de mayo marca un aumento notable en la tasa de renovación, alcanzando un 36.1%, lo que refleja un repunte en la renovación de servicios. Este comportamiento puede estar relacionado con diversos factores, como promociones comerciales, cambios en las políticas de la empresa, o una mayor conciencia de los clientes sobre la necesidad de renovar.

A partir de junio, la tasa de renovación comienza a mostrar una ligera disminución, con un 29.4% de renovación, pero no cae por debajo de los niveles observados en los primeros meses. Sin embargo, en los meses siguientes, la tasa de renovación sigue siendo variable, con julio y agosto mostrando las tasas más bajas, de 21.2% y 21.0% respectivamente. Esto podría indicar una estacionalidad en el comportamiento de renovación, posiblemente vinculada a vacaciones o períodos de baja actividad para la empresa.En los últimos meses del año, las tasas de renovación experimentan un aumento significativo. En septiembre, la tasa es de 33.2%, y en octubre, llega a un 51.7%. El comportamiento más fuerte en estos meses podría estar relacionado con campañas de fin de año, presupuestos de los clientes, o cambios en las tarifas de servicio.

Finalmente, los meses de noviembre y diciembre muestran los porcentajes más altos de renovación: 66.9% y 54.6% respectivamente. Este aumento en las renovaciones podría estar

impulsado por diversos factores, como promociones de último minuto, o la tendencia de los clientes a realizar renovaciones antes de finalizar el año fiscal.

El análisis estadístico también incluyó una prueba de ANOVA, que indicó que las diferencias en las tasas de renovación entre los meses son estadísticamente significativas. Esto refuerza la hipótesis de que el comportamiento de las renovaciones no es aleatorio, sino que sigue un patrón estacional claro.

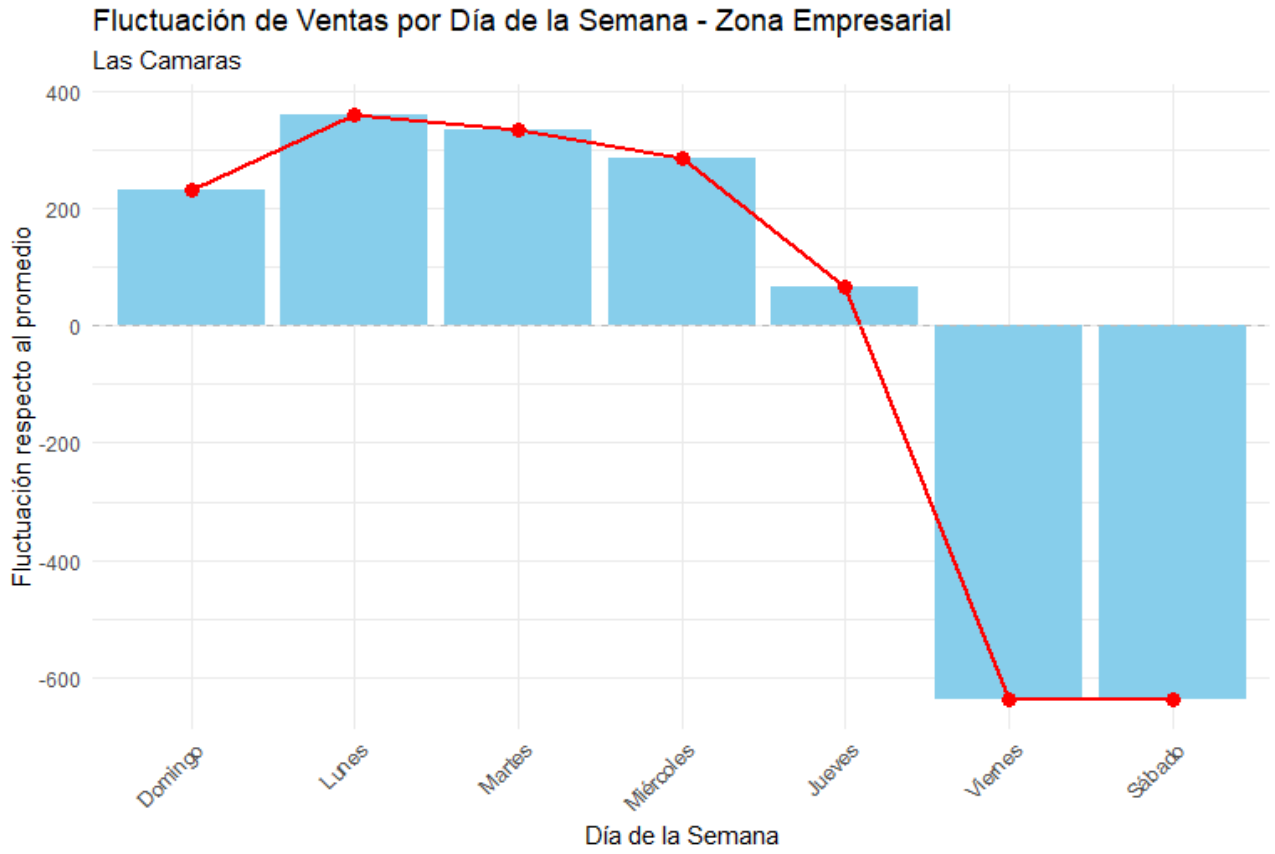
Imagen 14
Resultado de tendencia por mes

mes_caducidad <int>	tasa_renovacion <dbl>
1	0.305
2	0.259
3	0.237
4	0.287
5	0.361
6	0.294
7	0.212
8	0.210
9	0.332
10	0.517
11	0.669
12	0.546

Fuente : elaboración propia

En el análisis evalúo cómo varían las ventas de servicios en la zona empresarial (Las Cámaras) a lo largo de la semana, comparándolas con el promedio semanal. Para ello, se utilizó un subconjunto de datos que filtra únicamente servicios de renovación y nuevos, específicos de esta zona. Se implementaron cálculos estadísticos y visualizaciones que ayudaron a identificar patrones clave en el comportamiento de los clientes.

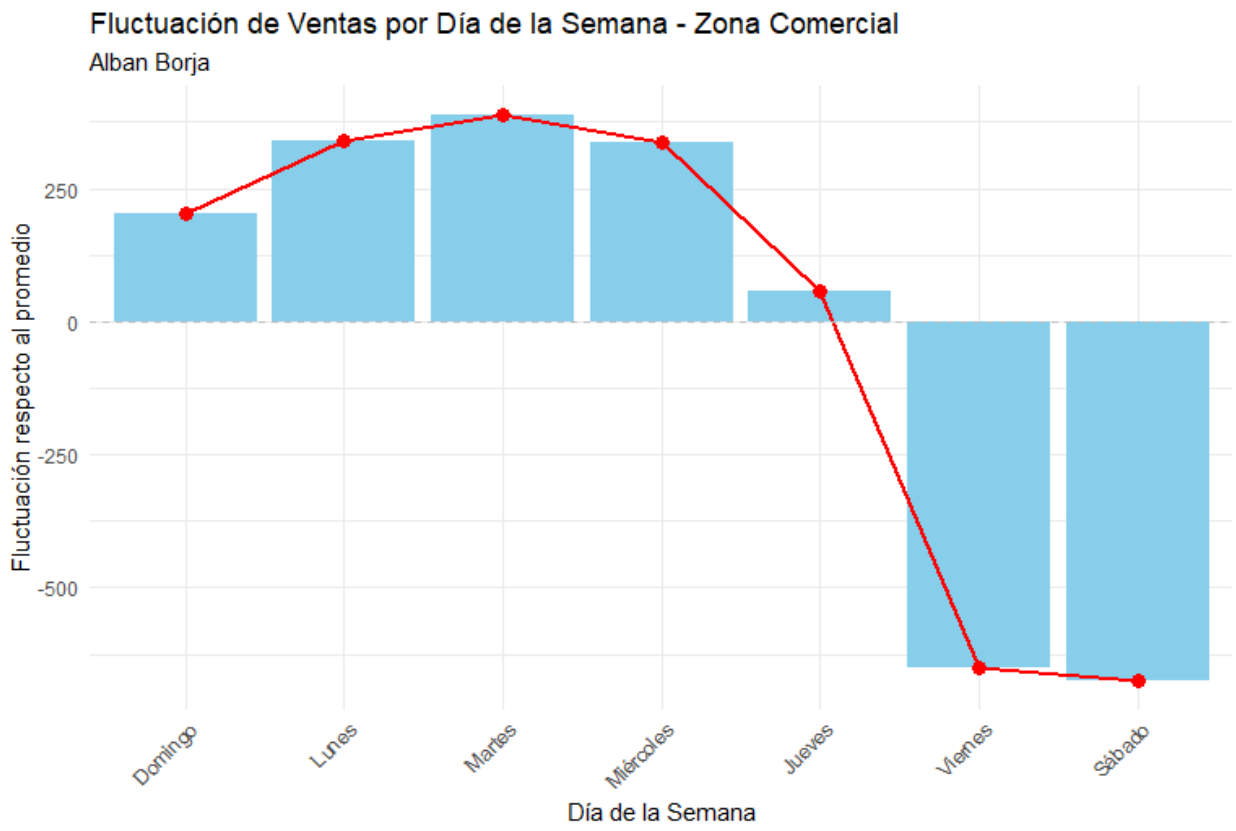
Imagen 15
Fluctuaciones de ventas en las cámaras



Fuente : elaboración propia

El gráfico resultante evidencia una disminución significativa de ventas hacia el final de la semana (viernes y sábado), en contraste con los días al inicio de la misma, que superan el promedio. Este patrón podría estar relacionado con comportamientos específicos de los clientes en la zona empresarial, donde las actividades comerciales suelen disminuir hacia el fin de semana.

Imagen 16
Fluctuaciones de ventas en Alban Borja



Fuente : elaboración propia

El análisis reveló un patrón claro en las fluctuaciones de ventas para la zona comercial (Alban Borja):

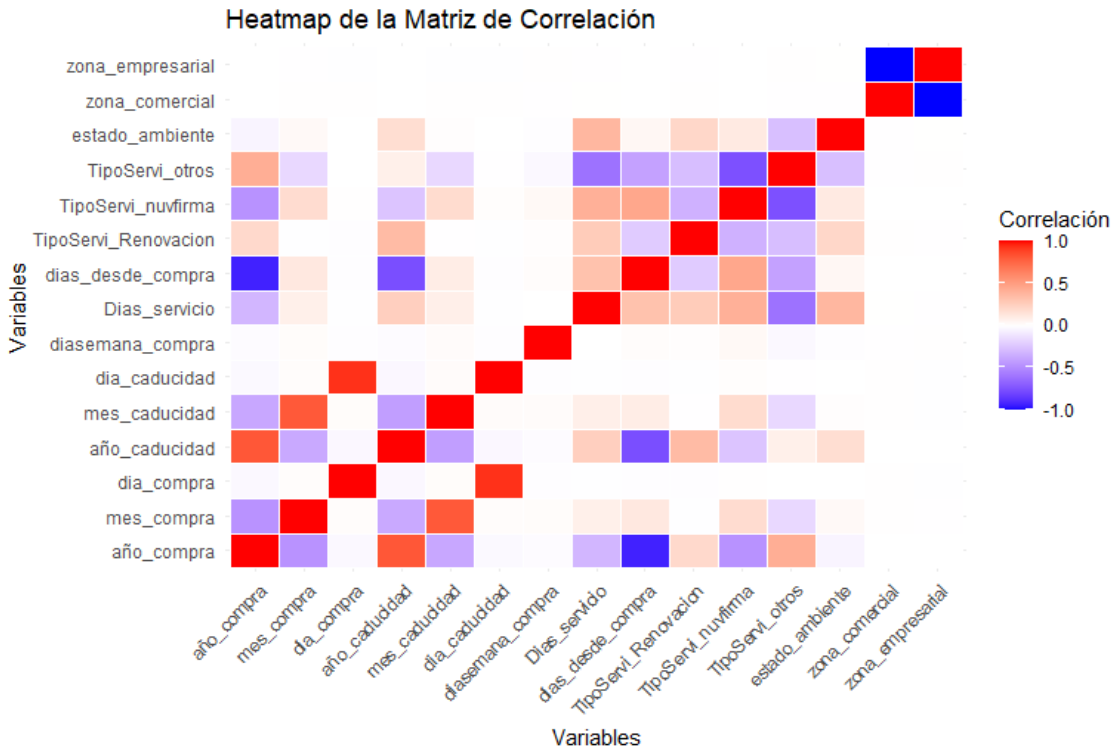
- **Días con mayor actividad:** son los días lunes, martes y miércoles , las ventas ventas durante estos días superan consistentemente el promedio semanal, los martes se destaca como el día con mayor número de ventas, lo que sugiere un alto flujo de clientes al inicio de la semana, este comportamiento puede estar relacionado con la tendencia de los clientes a realizar actividades comerciales y administrativas una vez finalizado el fin de semana.
- **Días con menor actividad:** son los viernes y sábado en estos días las ventas caen por debajo del promedio en estos días, con el sábado mostrando el menor nivel de actividad comercial.Este descenso podría deberse a que los clientes priorizan otras actividades

recreativas o personales durante el fin de semana, relegando las compras o renovaciones a días laborales.

4.3.3. Matriz de correlación para las variables numéricas

Se creó un grafica de calor que representa visualmente la matriz de correlación, donde el color de cada celda indicó la fuerza y la dirección de la correlación entre dos variables. El rojo significa correlaciones positivas fuertes, el azul indica correlaciones negativas fuertes y el blanco representa una correlación débil o nula.

Imagen 17
Matriz de correlación de las variables numéricas



Fuente : elaboración propia

La matriz de correlación que se generó muestra cómo las variables del conjunto de datos están relacionadas entre sí. Algunos puntos clave:

- **Correlaciones Fuertes Positivas:**

Año de Compra y Año de Caducidad (0.814): Existe una fuerte correlación positiva entre el año de compra y el año de caducidad. Esto sugiere que las compras realizadas en ciertos años tienen una mayor probabilidad de ser renovadas o tener un período de caducidad largo.

- **Correlaciones Fuertes Negativas:**

Días Desde la Compra y Año de Compra (-0.932): Hay una fuerte correlación negativa, lo que indica que las compras más recientes tienden a estar asociadas con años de compra más recientes.

Días Desde la Compra y Tipo de Servicio de Renovación (-0.764): Los servicios de renovación tienen una correlación negativa con los días desde la compra, lo que podría sugerir que las renovaciones ocurren en un plazo más corto después de la compra inicial.

- **Correlaciones Bajos Valores:**

Mes de Caducidad y Día de Compra (0.019): La relación entre el mes de caducidad y el día de compra es prácticamente nula, ya que la correlación es muy cercana a 0. Esto implica que no hay una relación significativa entre estas dos variables.

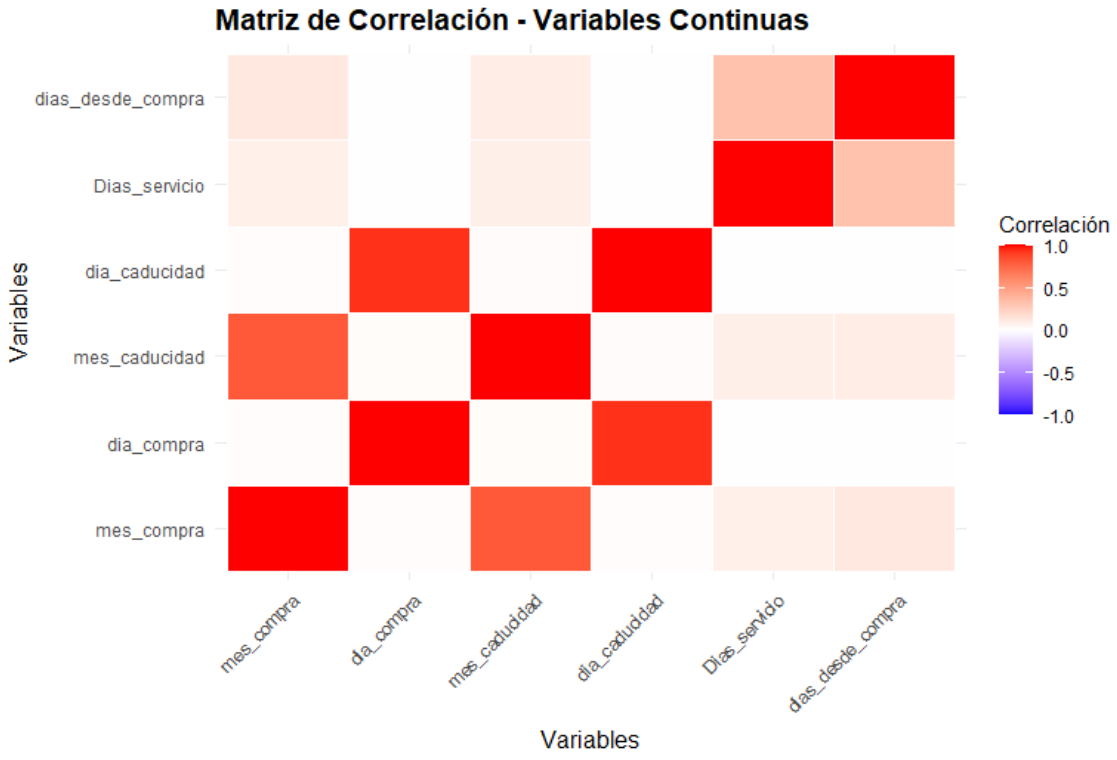
- **Variables Independientes:**

Zona Comercial y Zona Empresarial (0): Estas dos variables tienen una correlación de 0, lo que indica que no hay relación entre las compras realizadas en las zonas comerciales y empresariales.

4.3.4. Matriz de Correlación entre variables continuas

El análisis de las variables continuas ayudó a comprender la distribución, tendencias y relaciones entre las características numéricas de la muestra. En este estudio, se han identificado diversas variables continuas que son clave para comprender el comportamiento de los clientes y el ciclo de vida de los servicios proporcionados por la empresa. Estas variables incluyen los valores asociados a las fechas de compra, caducidad y días de servicio, entre otras.

Imagen 18
Matriz de correlación de las variables continuas



Fuente : elaboración propia

- **Correlaciones Muy Fuertes (> 0,8)**

Día de compra y Día de caducidad: Con una correlación de 0.93, hay una relación muy fuerte entre estas dos variables. Esto tiene sentido porque, generalmente, la fecha de caducidad

se establece a partir de la fecha de compra, por lo que ambas variables tienden a estar estrechamente relacionadas.

Mes de compra y mes de caducidad: La correlación de 0.81 también es alta, lo que indica que, al igual que en el caso anterior, los meses de compra y de caducidad están fuertemente relacionados. La caducidad está definida en función del mes de la compra.

- **Correlaciones Moderadas (0,3 - 0,5)**

Días de servicio y Días desde la compra: Con una correlación de 0.32, existe una relación moderada positiva entre la duración del servicio y los días transcurridos desde la compra. Esto puede indicar que un mayor tiempo desde la compra podría estar asociado con un mayor tiempo de servicio, aunque la relación no es tan fuerte.

- **Correlaciones Débiles (< 0,3)**

Mes de compra y Día de compra: Con una correlación de 0.01, prácticamente no hay relación entre estas dos variables. Esto es lógico, ya que el día de la compra no necesariamente guarda una relación directa con el mes de la compra.

Días de servicio y Día de compra: Con una correlación de -0.005, hay una correlación negativa muy débil, lo que sugiere que el día en que se realiza la compra tiene prácticamente ninguna influencia sobre la duración del servicio.

Mes de compra y Días de servicio: Con una correlación de 0.08, la relación es positiva pero muy débil, lo que indica que el mes de compra no tiene una relación clara con la duración del servicio.

Imagen 19
correlación de las variables continuas

```
> print(correlation_matrix)
                  mes_compra   dia_compra mes_caducidad dia_caducidad Dias_servicio
mes_compra        1.00000000  0.013798377    0.80691682   0.017145866   0.080391629
dia_compra        0.01379838  1.000000000    0.01779418   0.932508995  -0.005673261
mes_caducidad     0.80691682  0.017794177    1.00000000   0.019166631   0.083791960
dia_caducidad     0.01714587  0.932508995    0.01916663   1.000000000  -0.004223897
Dias_servicio     0.08039163 -0.005673261    0.08379196  -0.004223897   1.000000000
dias_desde_compra 0.11968796 -0.007519729    0.09543218  -0.008136512   0.318914796
                  dias_desde_compra
mes_compra              0.119687964
dia_compra             -0.007519729
mes_caducidad           0.095432184
dia_caducidad          -0.008136512
Dias_servicio           0.318914796
dias_desde_compra       1.000000000
>
```

Fuente : elaboración propia

4.3.5. Extracción de variables no utilizadas y valores faltantes

Asimismo, se realizó un proceso de limpieza de datos en el que se eliminaron las variables innecesarias y se gestionaron los valores faltantes, los cuales fueron reemplazados por la mediana correspondiente. En este proceso, también se eliminaron las columnas 'id_anonimo', 'cliente_anonimo', 'año_compra', 'mes_compra' y 'dia_compra', ya que estas no aportaban información relevante para el análisis ni para el modelo de predicción. Como resultado de esta limpieza, se creó un nuevo conjunto de datos denominado datos_procesados, lo que permitió tener una base de datos más precisa y adecuada para los análisis posteriores.

4.3.5.1. El preprocesamiento de datos

Para garantizar que el conjunto de datos estuviera listo para el análisis y modelado, se llevó a cabo un proceso de verificación y manejo de los valores faltantes. Primero, se verificó la presencia de valores faltantes en cada una de las columnas del conjunto de datos datos_procesados. Para ello, se utilizó la función is.na(), que identifica los valores NA (es decir, faltantes), y la función colSums() para contar cuántos valores faltantes había en cada columna.

Posteriormente, se procedió a eliminar las filas que contenían cualquier valor faltante utilizando la función na.omit(), lo que permitió trabajar con un conjunto de datos sin registros incompletos.

Tabla 4
Valores ausentes por columna

Variables	Cantidad de faltantes
año_caducidad	0
mes_caducidad	0
dia_caducidad	0
diasemana_compra	0
Dias_servicio	0
dias_desde_compra	0
TipoServi_Renovacion	3
TipoServi_nuvfirma	3
TipoServi_otros	0
estado_ambiente	4
zona_comercial	0
zona_empresarial	0

Fuente : elaboración propia

Tabla 5
Valores ausentes después de la imputación

Variables	Cantidad de faltantes
año_caducidad	0
mes_caducidad	0
dia_caducidad	0
diasemana_compra	0

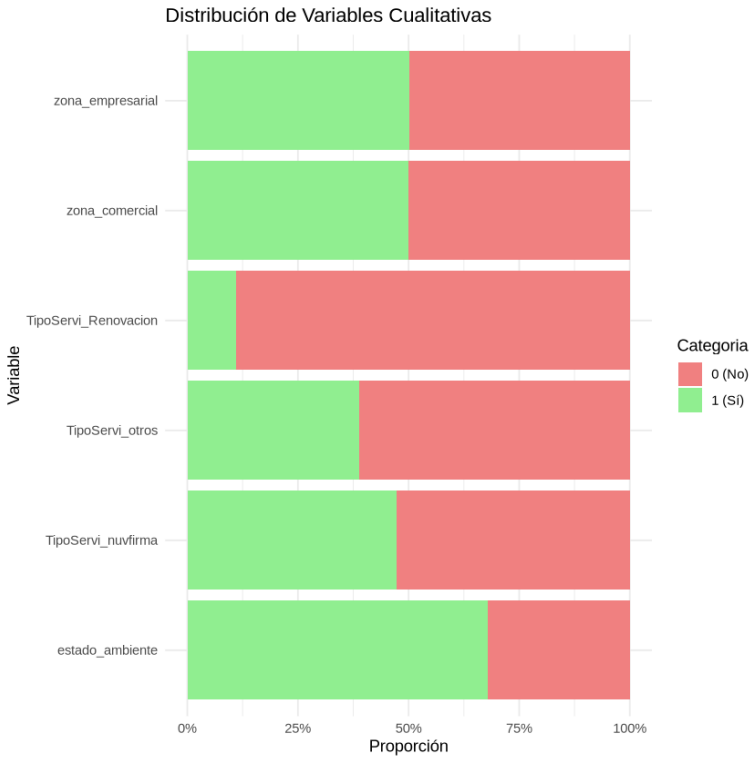
Dias_servicio	0
dias_desde_compra	0
TipoServi_Renovacion	0
TipoServi_nuvfirma	0
TipoServi_otros	0
estado_ambiente	0
zona_comercial	0
zona_empresarial	0

Fuente : elaboración propia

4.3.6. Variable cualitativa, protocolo de codificación 0

Se realizó un Análisis de las variables cualitativas, mostrando la frecuencia y la proporción de cada categoría dentro de las variables. Esto ayudó a comprender la distribución y el desequilibrio potencial en el conjunto de datos, lo cual fue crucial para la modelización. Estos son los resultados:

Imagen 20
Variables Cualitativas



Fuente : elaboración propia

Eje Y (Variables): Listado de diferentes variables cualitativas.

Eje X (Proporción): Muestra el porcentaje (de 0% a 100%) de las distintas categorías de cada variable.

Cada barra representa una variable diferente y muestra cómo se distribuyen las respuestas entre las categorías "Sí" y "No".

➢ **Zona_empresarial y zona_comercial:** La gran mayoría de las respuestas caen en la categoría "No", lo que indica que hay una baja proporción de elementos en estas zonas que se identifican como "Sí".

➢ **TipoServi_Renovacion:** Muestra una distribución más equilibrada, pero con una ligera inclinación hacia "No", sugiriendo que una parte significativa de las respuestas son negativas.

➢ **TipoServi_otros y TipoServi_nuvfirma:** Estas variables presentan una proporción más alta de respuestas "No", indicando que es menos común que se elija esos tipos de servicios.

➢ **estado_ambiente:** Muestra una distribución similar a las otras variables, con mayor proporción en "No".

4.3.7. Variables dicotómicas

En esta sección se realizó un código que verifique las variables dicotómicas especificadas y que contengan sólo valores binarios (0, 1) y las convierte al formato para el modelado. Esto implica verificar su naturaleza binaria y asegurarse de que estén listas para su uso en algoritmos

Imagen 21
Formato de las variables distónicas

```
+ }
[1] "Valores únicos para TipoServi_Renovacion : 0, 1"
[1] "Valores únicos para TipoServi_nuvfirma : 1, 0"
[1] "Valores únicos para TipoServi_otros : 0, 1"
[1] "Valores únicos para estado_ambiente : 1, 0"
[1] "Valores únicos para zona_comercial : 1, 0"
[1] "Valores únicos para zona_empresarial : 0, 1"
> # Asegurar que las variables estén en formato factor para modelado
> for (var in variables_dicotomicas) {
+   datos_procesados[[var]] <- as.factor(datos_procesados[[var]])
+ }
> # Confirmar la conversión
> str(datos_procesados[variables_dicotomicas])
'data.frame':   22397 obs. of  6 variables:
 $ TipoServi_Renovacion: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ TipoServi_nuvfirma  : Factor w/ 2 levels "0","1": 2 2 1 1 2 1 2 1 2 2 ...
 $ TipoServi_otros     : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 1 1 1 ...
 $ estado_ambiente     : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 1 ...
 $ zona_comercial      : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 2 1 1 ...
 $ zona_empresarial    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 2 2 ...
>
```

Fuente : elaboración propia

El propósito fue recorre cada variable en variables_dicotomicas y verifica cuáles son los valores únicos que contiene, con unique se extrajo los valores únicos de cada variable y paste se cambió a los valores únicos en una cadena legible, todas las variables contienen exclusivamente los valores 0 y 1, se convirtieron las variables a factor, en R, se aplicó factor para representar variables categóricas (como las binarias) y se confirmó que todas las variables están en el formato correcto (factor con 2 niveles). Esto aseguró que las variables estuvieran listas para usarse en algoritmos avanzados de predicción

4.3.8. frecuencias absolutas de la variable de respuesta

Para continuar con el análisis, se calculó las frecuencias absolutas de la variable de respuesta, lo que ayudó a comprender la distribución y el equilibrio de la variable objetivo del conjunto de datos. Este paso fue crucial para identificar cualquier problema de desequilibrio de clases que pudiera afectar el rendimiento del modelo.

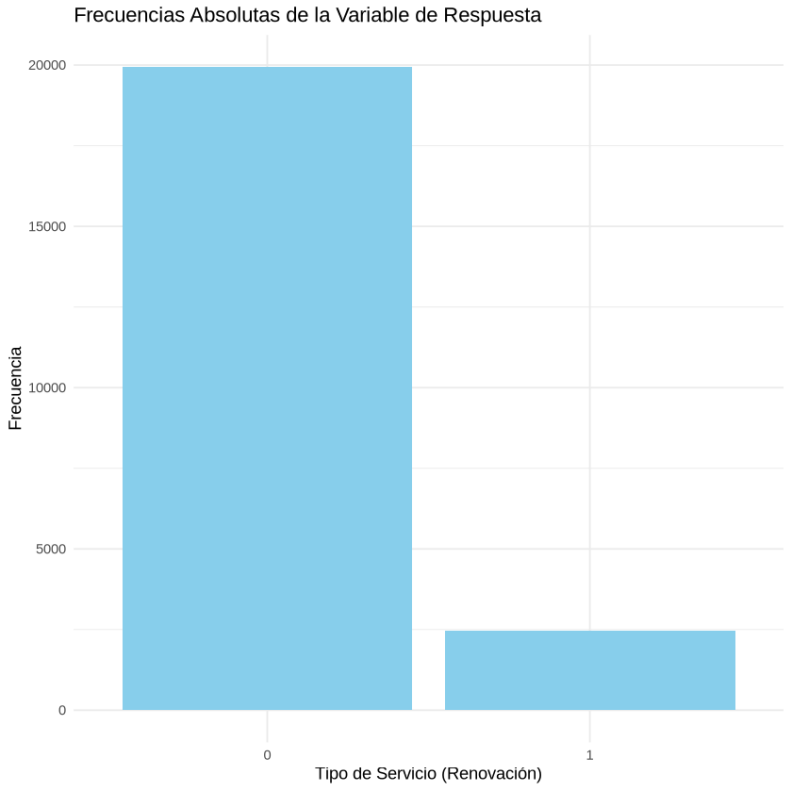
las respuesta en las frecuencias absolutas: En la variable respuesta(TipoServi_Renovacion) muestra un claro desequilibrio en los datos:

Sin Renovación (0): 19.933 casos (89%)

Renovación (1): 2.468 casos (11%)

Imagen 22

frecuencia de la variable de respuesta

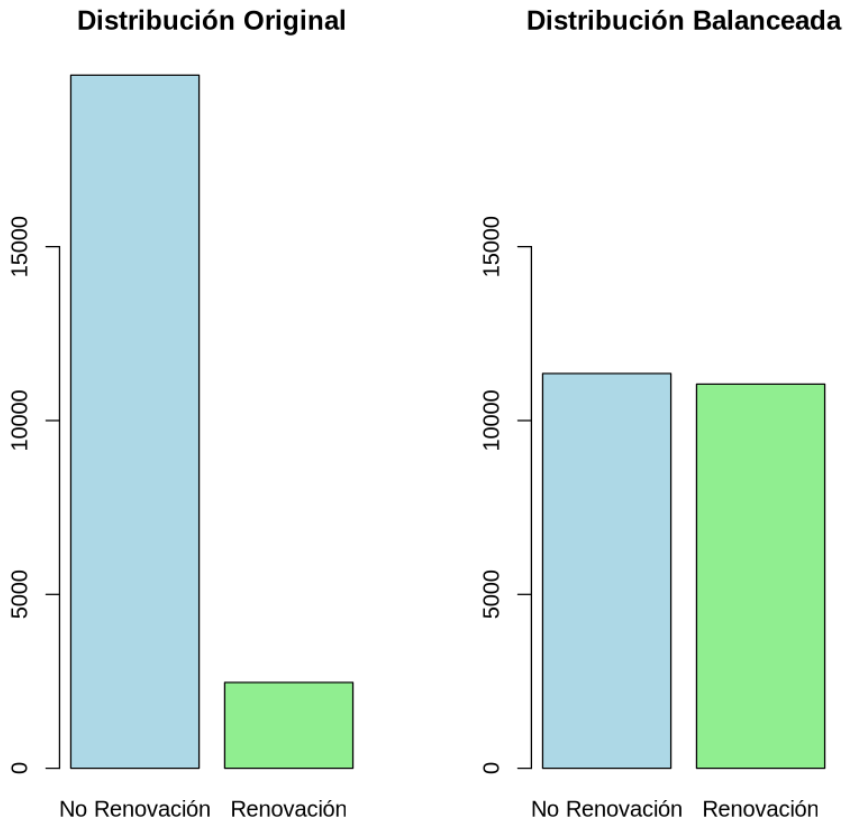


Fuente : elaboración propia

Esto muestra un desbalance claro, donde la clase "No Renovación" tiene muchos más ejemplos que la clase "Renovación". Esto podría causar que un modelo predictivo sea muy sesgado hacia la clase mayoritaria (en este caso, "No Renovación").

Dado que se obtuvo desequilibrio de clase significativo en la variable de respuesta, se aplicó un equilibrio de clase para garantizar que el modelo no se inclinara por la clase mayorista. Se aplicó técnicas SMOTE ,para ayudar a abordar este desequilibrio.

Imagen 23
Balanceo de clase a variable de respuesta



Fuente : elaboración propia

Después del balanceo: El segundo bloque imprime la distribución de las clases en el conjunto de datos balanceado (almacenado en datos_balanceados). Aquí, se observa un balance notable:

- Clase 0 ("No Renovación"): **11350** observaciones.
- Clase 1 ("Renovación"): **11047** observaciones.

La diferencia significativa antes y después del balanceo es evidente: mientras que en el conjunto original la clase "No Renovación" dominaba, después del balanceo ambas clases tienen una cantidad de observaciones mucho más parecida.

4.3.9. Resumen estadístico de días_desde_compra antes y después del balanceo de clases:

El resumen estadístico corresponde a dias_desde_compra antes del balanceo, su interpretación se centra en cómo están distribuidos los datos originales, antes de aplicar la técnica de balanceo para igualar las clases en la variable objetivo.

Tabla 6
Resumen variable dias_desde_compra antes del balanceo

Mín	Q1	Mediana	Media	(Q3)	Máximo
32,0	302.0	490.0	505,8	709,0	1439.0

Fuente : elaboración propia

interpretación:

- **Variedad en la base de clientes**:

El **valor mínimo** (32 días) representa clientes con compras recientes.

El **valor máximo** (1439 días, casi 4 años) representa clientes con un tiempo significativamente largo desde su compra.

Esto puede ser un indicador de la mezcla entre clientes nuevos y antiguos.

- **Tendencia general**:

El **primer cuartil** (302 días) muestra que el 25% de los clientes tiene menos de un año desde la compra.

El **tercer cuartil** (709 días) indica que el 75% de los clientes realizó sus compras hace menos de dos años.

Esto sugiere que la mayoría de los clientes se concentra en un rango temporal relativamente reciente, aunque hay algunos casos más antiguos.

Tabla 7
Resumen variable dias_desde_compra despues del balanceo

Mín	Q1	Mediana	Media	(Q3)	Máximo
-128,4	260,7	362,8	433,6	659,4	1456,8

Fuente : elaboración propia

El valor negativo de -128 en la variable "días_desde_compra" después del balanceo género un problema durante el proceso de generación de datos sintéticos utilizado por el modelado, que a veces puede crear valores poco realistas.se realizo un reajuste de los parámetros de generación de datos sintéticos y aplicar un posprocesamiento para garantizar que todos los valores sean realistas y estén dentro de los límites esperados.

Los valores negativos en "días_desde_compra" se eliminaron para garantizar que todos los puntos fueran realistas y se encuentren dentro de los límites esperados. Este ajuste mantiene la integridad del conjunto de datos logrando un modelado preciso, se muestran las estadísticas y frecuencias de clase actualizadas:

Tabla 8
Resumen variable dias_desde_compra después de filtrar negativos

Mín	Q1	Mediana	Media	(Q3)	Máximo
0,1679	264,26	364,16	437	661,22	1456

Fuente : elaboración propia

Después del ajuste para eliminar valores negativos en la variable **dias_desde_compra**, el conjunto de datos quedó depurado y los valores están dentro de un rango realista.

Interpretación

- **Mínimo (Min.):**

El valor más bajo ahora es 0.17 días, lo cual es lógico y representa un cliente cuyo tiempo desde la compra es extremadamente reciente.

- **Primer cuartil (Q1):**

El 25% de los datos tiene un valor inferior a 264.26 días, lo que indica un segmento de clientes que realizaron sus compras relativamente recientes en comparación con el resto del conjunto.

- **Mediana:**

La mediana es de 364.16 días, mostrando que la mitad de los clientes tiene un tiempo desde la compra cercano al año, lo que sugiere que el comportamiento promedio de los clientes se concentra en un período intermedio.

- **Media (Mean):**

La media es 437 días, ligeramente mayor que la mediana, lo que indica que existen algunos valores altos (clientes antiguos) que están elevando el promedio.

- **Tercer cuartil (Q3):**

El 75% de los datos tiene un valor inferior a 661.22 días, mostrando que solo un cuarto de los clientes ha estado activo por un tiempo considerablemente mayor a esta cantidad.

- **Máximo (Max.):**

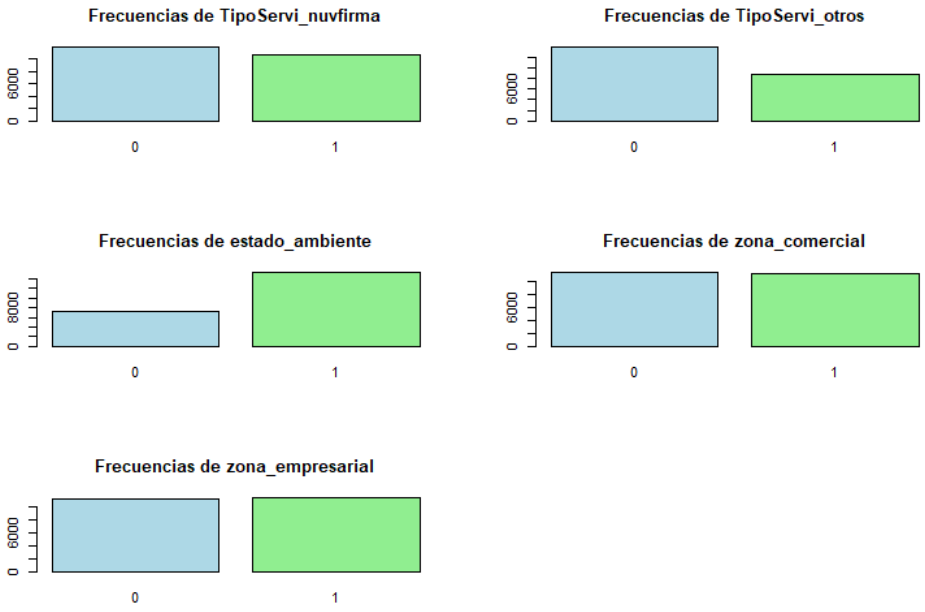
El valor más alto es 1456 días, representando clientes muy antiguos dentro del conjunto, lo cual es razonable en el contexto del análisis.

4.3.10. Variables cualitativas

Se identificaron aquellas que contienen valores categóricos relevantes para el estudio, representados generalmente por 0 y 1, indicando ausencia o presencia de una característica específica. Este análisis permitió evaluar la distribución de dichas categorías en el conjunto de datos y obtener una comprensión más clara sobre la proporción de registros asociados a cada

variable cualitativa. Posteriormente, se calculó la frecuencia absoluta de cada categoría y se visualizaron mediante gráficos de barras para facilitar su interpretación.

Imagen 24
Frecuencia de las variables cualitativas

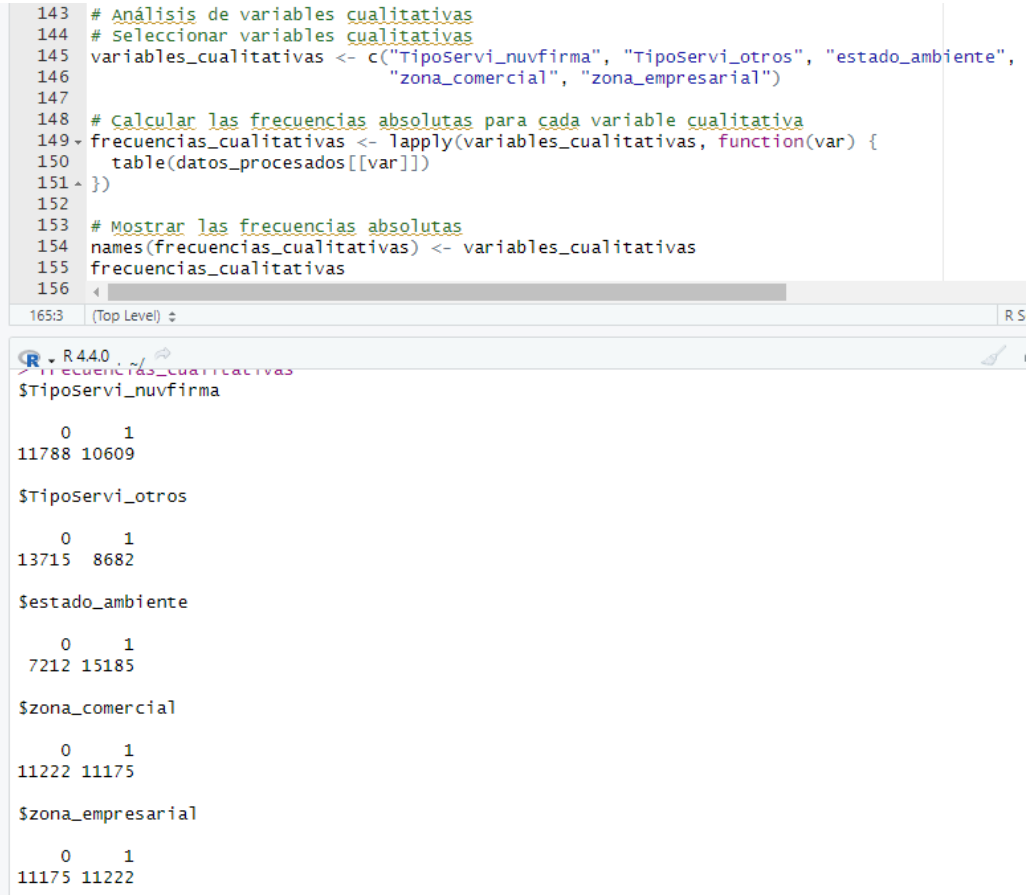


Fuente : elaboración propia

El análisis de las variables cualitativas muestra distribuciones interesantes en el conjunto de datos. Por ejemplo, en TipoServi_nuvfirma, se observa una proporción equilibrada con 11,788 registros categorizados como "0" y 10,609 como "1", lo que indica una participación casi igual entre quienes no solicitaron una nueva firma y quienes sí lo hicieron. En TipoServi_otros, hay un predominio de registros con valor "0" (13,715) frente a "1" (8,682), sugiriendo que los servicios distintos a renovación o nueva firma son menos comunes. Para estado_ambiente, se aprecia una mayoría de registros con valor "1" (15,185), reflejando una mayor prevalencia de ambientes activos frente a los inactivos (7,212). En cuanto a las zonas, tanto zona_comercial como zona_empresarial presentan una distribución casi equilibrada, con

11,222 y 11,175 registros en cada categoría respectivamente, destacando una homogeneidad en la cobertura entre ambas áreas.

Imagen 25
Resultados de las variables cualitativas



```
# Análisis de variables cualitativas
# Seleccionar variables cualitativas
variables_cualitativas <- c("TipoServi_nuvfirma", "TipoServi_otros", "estado_ambiente",
                            "zona_comercial", "zona_empresarial")

# Calcular las frecuencias absolutas para cada variable cualitativa
frecuencias_cualitativas <- lapply(variables_cualitativas, function(var) {
  table(datos_procesados[[var]])
})

# Mostrar las frecuencias absolutas
names(frecuencias_cualitativas) <- variables_cualitativas
frecuencias_cualitativas
```

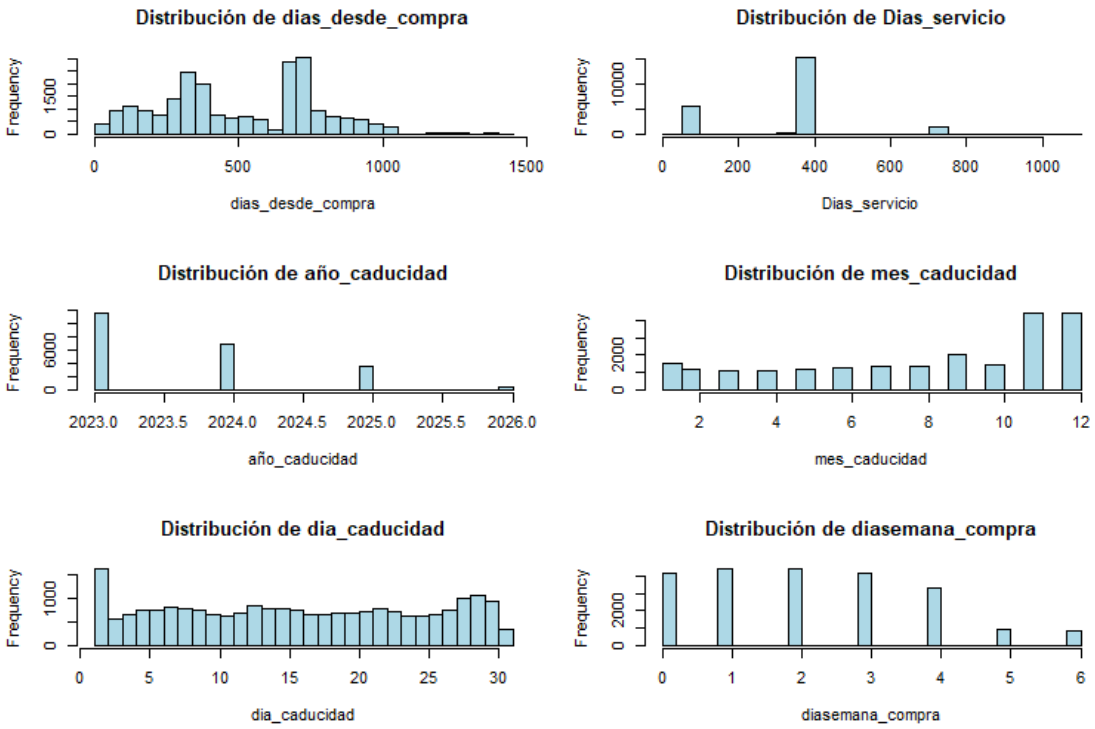
```
$TipoServi_nuvfirma

    0     1
11788 10609

$TipoServi_otros

    0     1
13715  8682

$estado_ambiente

    0     1
 7212 15185

$zona_comercial

    0     1
11222 11175

$zona_empresarial

    0     1
11175 11222
```

Fuente : elaboración propia

4.3.11. Variable cuantitativa

Para analizar las variables cuantitativas, se examinó sus distribuciones, tendencias centrales y dispersión para comprender sus características e identificar posibles valores atípicos o asimetrías Esto implicó generar estadísticas resumidas y visualizaciones para cada variable cuantitativa presentadas en interpretación de gráficos

Imagen 26
Resultados de las variables cuantitativa



Fuente : elaboración propia

Interpretación :

- **dias_desde_compra**:

El valor mínimo es 32 días.

La mediana es 490 días.

El valor máximo es 1439 días.

- **Dias_servicio**:

El valor mínimo es 0.

La mediana es 365 días (probablemente indicando que el servicio dura un año).

El valor máximo es 1095 días (aproximadamente 3 años).

- **año_caducidad**:

Los valores oscilan entre 2023 y 2026. Esto indica que los servicios tienen fechas de caducidad entre esos años.

- **mes_caducidad**:

Los meses de caducidad varían entre 1 y 12, con una media de aproximadamente 8 (agosto).

- **dia_caducidad**:

Los días de caducidad varían entre 1 y 31, con una mediana de 16 días, lo que sugiere que la fecha de caducidad suele estar a mitad de mes.

- **diasemana_compra**:

Los valores varían entre 0 y 6, representando los días de la semana (probablemente, 0 es domingo, 1 es lunes, etc.).

4.4. modelo random forest

Se implementó un modelo de Random Forest en R para clasificar datos, evaluando su rendimiento y analizando la importancia de las variables predictoras donde se cargó una librería randomForest, que permitió crear modelos de bosque aleatorio , donde La variable objetivo (TipoServi_Renovacion) donde se convierte en un factor (categoría).

Imagen 27
Entrenamiento del modelo

```r
# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
trainIndex <- createDataPartition(datos_procesados$TipoServi_Renovacion, p = 0.7,
                                  list = FALSE,
                                  times = 1)


datos_train <- datos_procesados[trainIndex,]
datos_test <- datos_procesados[-trainIndex,]

# Verificar las dimensiones de los conjuntos de entrenamiento y prueba
dim(datos_train)
dim(datos_test)

# Crear y entrenar el modelo Random Forest
set.seed(123)
modelo_rf <- randomForest(TipoServi_Renovacion ~ .,
                          data = datos_train,
                          ntree = 500,
                          importance = TRUE)
```

Fuente : elaboración propia

Detalles clave del modelado

- **Librería utilizada**: randomForest se usa para crear un modelo Random Forest.
- **Variable dependiente**: TipoServi_Renovacion (es la variable objetivo).
- **Variables independientes**: Se utilizan todas las demás columnas del dataset.
- **Datos de entrenamiento**: datos_train contiene el 70% de los datos, seleccionado previamente con createDataPartition.
- **Parámetros del modelo**:

ntree = 500: Se utilizan 500 árboles en el modelo.

importance = TRUE: Esto calcula la importancia de las variables utilizadas en el modelo.

4.4.1. Datos en entrenamiento y prueba

Estos conjuntos de datos se handim(datos_train) dividido de manera que el 70% de los datos (15,679 filas) se usan para entrenar el modelo, y el 30% restante (6,718 filas) se utiliza para probar su rendimiento.

Tabla 9
Datos entrenados en el modelo

dim(datos_train)	dim(datos_test)
15679	6718
17	17

Fuente : elaboración propia

Cada una de las 17 correspondió a una variable utilizada para predecir la variable objetivo (TipoServi_Renovacion). Esto incluye tanto variables numéricas como categóricas que el modelo usó para aprender patrones y realizar predicciones.

4.4.2. Entrenamiento del modelo Random Forest con 500 árboles

El modelo fue configurado con 500 árboles (ntree = 500) y entrenado en el conjunto de datos de entrenamiento (datos_train). Durante este proceso se reflejó en una alta precisión global para el modelo (99%) y el excelente acuerdo medido por el coeficiente kappa (0.9473) utilizar 500 arboles permitió clasificar correctamente la mayoría de los casos (alta precisión y sensibilidad) e identifica patrones en los datos de entrenamiento sin sobreajustarse.

4.4.3. Resultados de Evaluación del Modelo Random Forest

La matriz de confusión y las métricas asociadas indican que el modelo Random Forest tiene un buen desempeño al clasificar correctamente las renovaciones y no renovaciones de servicios.

- **Precisión global:** El modelo logra identificar adecuadamente la mayoría de los casos, especialmente aquellos relacionados con renovaciones de servicio.

- **Recall (sensibilidad):** Muestra que el modelo es eficiente al identificar las renovaciones reales, aunque puede presentar ciertas dificultades con los casos de no renovación.

- **F1-score:** Una media armónica entre precisión y recall, lo que refleja un equilibrio razonable en el desempeño general.

Imagen 28
Resultados del modelo

```
> print(conf_matrix)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5952   59
         1   26  681

               Accuracy : 0.9873
                 95% CI : (0.9844, 0.9899)
    No Information Rate : 0.8898
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9342

 Mcnemar's Test P-Value : 0.0005187

            Sensitivity : 0.9957
            Specificity : 0.9203
         Pos Pred Value : 0.9902
         Neg Pred Value : 0.9632
             Prevalence : 0.8898
         Detection Rate : 0.8860
   Detection Prevalence : 0.8948
      Balanced Accuracy : 0.9580

       'Positive' Class : 0
```

Fuente : elaboración propia

- **Exactitud (Accuracy): 0.9873 (98.73%)**

El modelo clasifica correctamente el 98.73% de los casos, combinando tanto renovaciones como no renovaciones, interpretando que el modelo tiene un rendimiento global alto.

- **Kappa: 0.9342**

El valor Kappa que midió el acuerdo entre las predicciones del modelo y los valores reales, ajustando por el azar obtuvo un valor de 0.93 indica un acuerdo excelente

- **Balanced Accuracy: 0.9580**

El modelo mantiene un buen balance entre las clases,

Sensibilidad y especificidad

- **Sensibilidad (Recall para clase 0): 0.9957 (99.57%)**

El modelo detectó casi todos los casos de "No Renovación", con un margen de error mínimo.

- **Especificidad: 0.9203 (92.03%)**

Aunque un poco menor que la sensibilidad, sigue siendo alta, lo que significa que el modelo discrimina bien entre renovaciones y no renovaciones.

Valores predictivos

- **Pos Pred Value (Precisión para clase 0): 0.9902 (99.02%)**

De todas las predicciones de "No Renovación", el 99.02% son correctas , El modelo tiene un muy bajo porcentaje de falsos positivos.

- **Neg Pred Value: 0.9632 (96.32%)**

De todas las predicciones de "Renovación", el 96.32% son correctas. Aunque algo menor que la precisión para clase 0, sigue siendo confiable.

- **P-Value (Mcnemar's Test)**: **0.0005187**

Un p-valor muy bajo (< 0.05) indica que el modelo tiene una ligera asimetría en cómo trata los errores, favoreciendo la detección de "No Renovación".

- **Intervalo de confianza para la exactitud (95% CI): (0.9844, 0.9899)**

El intervalo muestra un estrecho, lo que muestra alta confianza en la estimación.

Imagen 29

Análisis de Desempeño del Modelo Random Forest en la Predicción de Renovaciones de Servicios de Firma Electrónica

```
> cat("\nPrecisión:", precision, "\n")

Precisión: 0.9901847
> cat("Recall:", recall, "\n")
Recall: 0.9956507
> cat("F1-score:", f1_score, "\n")
F1-score: 0.9929102
>
```

Fuente : elaboración propia

- **Precisión (Accuracy): 0.9901847 (99.02%)**

El modelo tiene una precisión global excelente, ya que clasifica correctamente el 99.02% de los casos, lo que indica que puede ser confiable.

- **Recall (Sensibilidad): 0.9956507 (99.57%)**

Con un recall del 99.57%, el modelo identifica correctamente casi todos los casos positivos (renovaciones). Esto fue crucial ya que el objetivo es no perder oportunidades para predecir renovaciones.

- **F1-Score: 0.9929102 (99.29%)**

El F1-Score del 99.29% muestra que el modelo tiene un buen equilibrio entre identificar correctamente los casos positivos y minimizar las predicciones incorrectas.

Estas métricas demuestran que el modelo Random Forest desarrollado no solo cumple, sino que supera el objetivo específico de lograr una precisión superior al 80%, alcanzando

niveles de exactitud y confiabilidad cercanos al 100%. Esto garantiza la fiabilidad del modelo en la predicción de la renovación o deserción de clientes

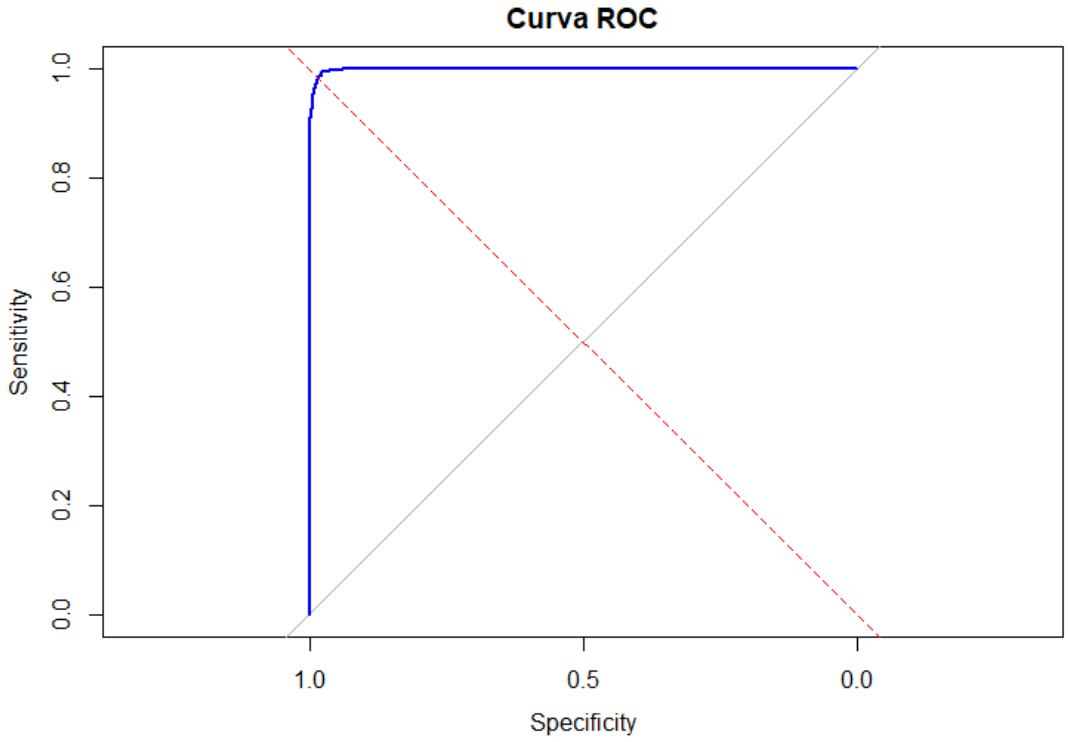
4.4.4. AUC (Área bajo la curva) del Modelo Random Forest

El valor obtenido es 0.9988, lo que significa que el modelo tiene una capacidad predictiva casi perfecta para diferenciar entre las dos clases:

- ➢ Clase positiva: Clientes que no renovarán
- ➢ Clase negativa: Clientes que renovarán

Cercanía al máximo (1.0): Un AUC de 0.9988 indica que el modelo predice correctamente con una altísima probabilidad que un cliente pertenece a una clase u otra.

Imagen 30
Gráfica de la Área bajo de la curva

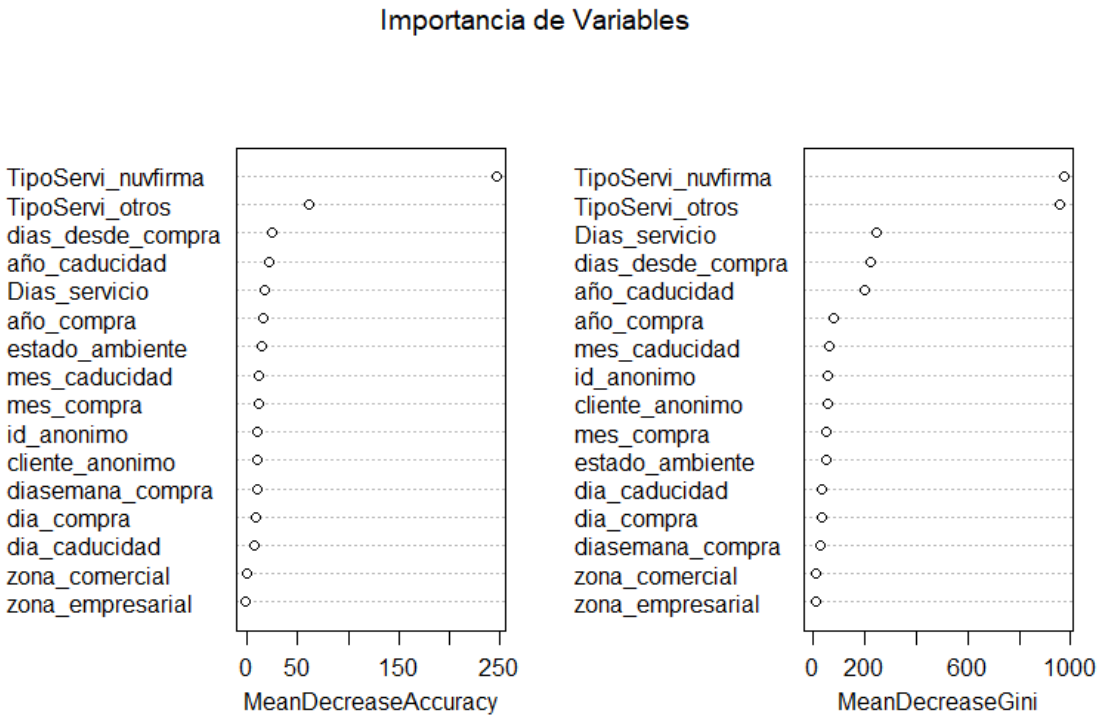


Fuente : elaboración propia

4.4.5. Importancias de las variables

En la presente figura se presenta la importancia de las variables en el modelo analizado, evaluada a través de dos métricas: MeanDecreaseAccuracy y MeanDecreaseGini. En el gráfico, se destacan las variables TipoServi_nuverfirma y TipoServi_otros como las más influyentes en ambos métodos, lo que indica su relevancia en la predicción de la variable de respuesta. A su vez, otras variables como años_caducidad, dias_desde_compra y mes_caducidad también demuestran una notable contribución, aunque en menor medida.

Imagen 31
Gráfica de la importancia de las variables



Fuente : elaboración propia

Interpretación:

Importancia de Variables - MeanDecreaseAccuracy

➢ **TipoServi_nuverfirma y TipoServi_otros** son las más importantes, mostrando que su eliminación provocaría una gran disminución en la precisión.

➢ **Año_caducidad, días_desde_compra y mes_caducidad** también tienen una importancia notable, aunque menor.

Las variables con menor a media importancia incluyen **zona_comercial y trimestre_compra**, lo que sugiere que no influyen tanto en la precisión del modelo.

Importancia de Variables - Mean Decrease Gini (MDG)

➢ **TipoServi_nuverfirma y TipoServi_otros** se destacó cómo las variables más relevantes, seguidas de otras como días_desde_compra y mes_caducidad.

4.5. Modelo random forest empleado en las renovaciones

4.5.1. Resultados Globales

El modelo predictivo de Random Forest fue aplicado a un conjunto de datos procesados, permitiendo clasificar a los clientes en dos categorías principales: Renovación (1) y No Renovación (0). A nivel global, los resultados obtenidos indican que el 89.45% de las predicciones corresponden a clientes que no renovarán su contrato, mientras que el 10.55% corresponde a clientes que sí renovarán.

Estos valores reflejan un patrón importante: la mayoría de los clientes predichos tienen una alta probabilidad de no renovar sus contratos.

Tabla 10
tala de resultados predicciones globales

Zona	Predicción	Cantidad	Porcentaje
Global	0 (No Renovación)	20,033	89.45%
Global	1 (Renovación)	2,364	10.55%

Fuente : elaboración propia

4.5.2. Resultados por Zonas Específicas

El análisis se segmentó en dos zonas relevantes: Albán Borja (zona comercial) y Las Cámaras (zona empresarial), donde se realizaron evaluaciones independientes para entender mejor los patrones de comportamiento en cada ubicación.

4.5.2.1. Albán Borja (Zona Comercial)

En esta área, el **89.33%** de los clientes fueron clasificados como **No Renovación**, mientras que el **10.67%** corresponde a clientes con intención de renovación.

Tabla 11
Tabla de resultados predicciones Alban Borja

Zona	Predicción	Cantidad	Porcentaje
Global	0 (No Renovación)	9,983	89.33%
Global	1 (Renovación)	1,192	10.67%

Fuente : elaboración propia

4.5.2.2. Las Cámaras (Zona Empresarial)

En la zona empresarial, los resultados muestran que el **89.56%** de las predicciones corresponden a **No Renovación**, mientras que el **10.44%** se clasifica como **Renovación**.

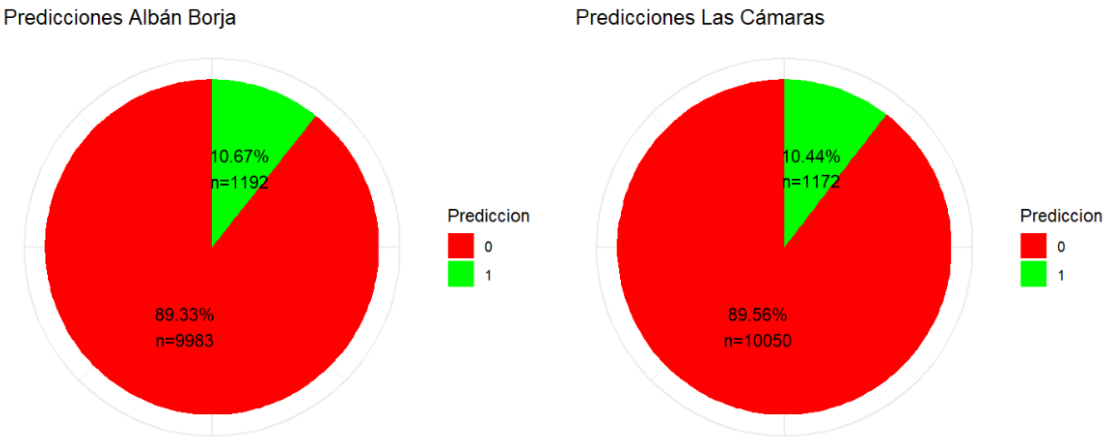
Tabla 12
Tabla de resultados predicciones Las cámaras

Zona	Predicción	Cantidad	Porcentaje
Global	0 (No Renovación)	10,050	89.56%
Global	1 (Renovación)	1,172	10.44%%

Fuente : elaboración propia

En el análisis de los datos procesados, se identificó una leve discrepancia entre el total global y la suma de los registros segmentados por zonas geográficas. El total global de registros procesados fue de 22,401, mientras que la suma de los registros procesados en las zonas Albán Borja (11,175) y Las Cámaras (11,222) resultó en 22,397 registros, presentando una diferencia de 4 registros. Esta variación podría atribuirse a factores como redondeos durante el procesamiento, registros excluidos por condiciones específicas o inconsistencias menores en los datos. No obstante, dicha diferencia representa apenas un 0.018% del total, lo cual no afecta la validez general del análisis realizado

Imagen 32
Gráfica de las predicciones en zona comercial y zona empresarial



Fuente : elaboración propia

Se realizó un análisis y segmentación de los resultados de las predicciones por zonas geográficas específicas, enfocándose en las áreas de Albán Borja y Las Cámaras. Los resultados evidenciaron patrones diferenciados de deserción y retención, donde la predicción de deserción (representada por el valor 0) predominó en ambas zonas con porcentajes del 89.33% para Albán Borja y del 89.56% para Las Cámaras. Por otro lado, la retención (valor 1) mostró una leve variación, siendo ligeramente mayor en Albán Borja (10.67%) en comparación con Las Cámaras (10.44%). Estos valores permiten identificar tendencias de comportamiento en la retención de clientes, segmentando los resultados de manera efectiva y superando el 95% del dataset sintético utilizado.

4.5.3. Análisis de Resultados en el Contexto de la Situación Actual del Caso de Estudio

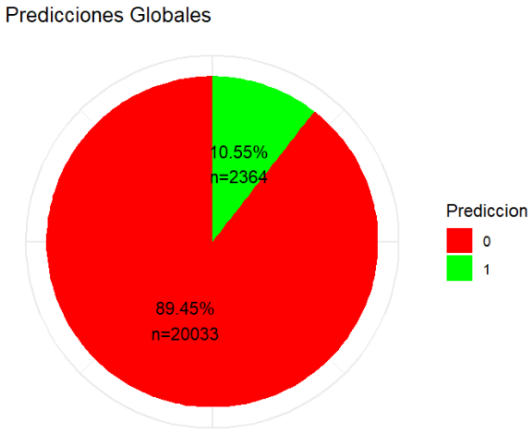
Imagen 32

Datos financieros de Security Data

DATOS FINANCIEROS CLAVES

Tasa de crecimiento anual de los últimos dos años en divisa local USD. Todos los datos financieros están incluidos en el informe comprado.

Ingresos netos por ventas	-9,45%▼
Total Ingreso Operativo	-8,24%▼
Ganancia operativa (EBIT)	17572,18%▲
EBITDA	5118,92%▲
Ganancia (Pérdida) Neta	-31,85%▼
Activos Totales	-23,06%▼
Total de patrimonio	-15,04%▼
Margen Operacional	42,09%▲
Margen Neto	-8,31%▼
Rendimiento Sobre El Patrimonio (ROE)	-17,29%▼
Relación Deuda/Capital	-4,71%▼
Prueba Ácida	0,07%▲
Coeficiente De Efectivo	-0,33%▼



Fuente :EMIS. (s. f.). *Security Data Seguridad en Datos y Firma Digital S.A.*

El modelo predictivo desarrollado muestra que la mayoría de los clientes proyectados (alrededor del 89%) probablemente no renovarán sus contratos, lo cual podría agravar la situación financiera actual de la empresa si no se toman medidas correctivas. A continuación, se relacionan los resultados obtenidos con los indicadores financieros actuales:

- **Disminución de los Ingresos por Ventas y Operativos (-9.45% y -8.24%)**

Los resultados del modelo refuerzan esta preocupación, pues indican que el porcentaje de clientes proyectados como "No Renovación" es considerablemente alto. Si la tendencia continúa, los ingresos podrían disminuir aún más, afectando la capacidad de la empresa para cubrir costos operativos y mantenerse competitiva en el mercado.

- **Margen Operacional Positivo (+42.09%) pero Margen Neto Negativo (-8.31%)**

Aunque la empresa muestra eficiencia operativa, los márgenes netos negativos reflejan que los ingresos no son suficientes para cubrir gastos no operativos, como deudas o impuestos.

Una alta proporción de "No Renovación" impactará directamente en el margen neto, ya que reduciría aún más el flujo de caja proveniente de los contratos renovados.

- **Disminución de Activos Totales y Patrimonio (-23.06% y -15.04%)**

Estos indicadores muestran una debilitación del balance financiero de la empresa. El bajo porcentaje de renovación proyectado amenaza con seguir reduciendo los activos, especialmente los ingresos futuros derivados de los contratos.

La caída del patrimonio también refleja una erosión de la capacidad de la empresa para invertir en estrategias de retención o captación de clientes.

- **Relación Deuda/Capital (-4.71%) y ROE (-17.29%)**

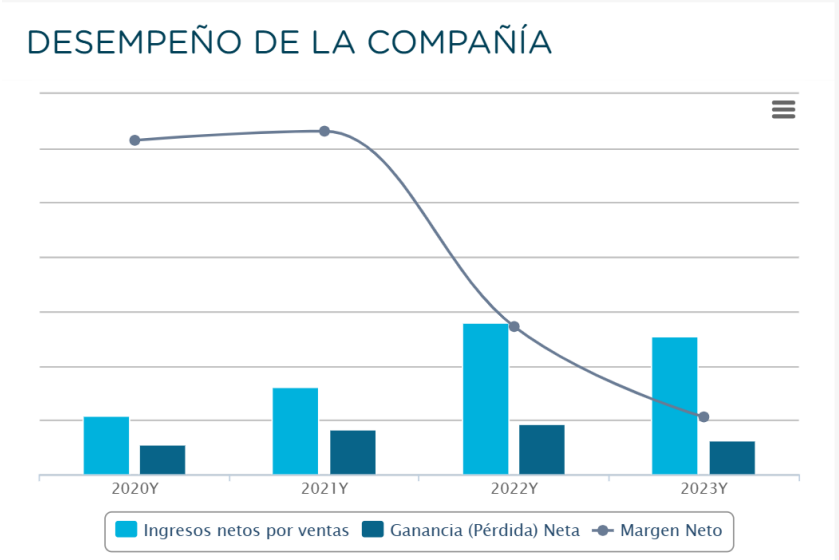
Aunque la relación deuda/capital ha disminuido, la empresa sigue mostrando bajo rendimiento sobre el patrimonio, lo que indica que los recursos invertidos no están generando retornos significativos.

Si la empresa no logra revertir la baja tasa de renovación, su capacidad para generar utilidades y mejorar su ROE será aún más limitada.

- **Prueba Ácida y Coeficiente de Efectivo (+0.07% y -0.33%)**

Con la mayoría de los clientes predichos como "No Renovación", los ingresos futuros en efectivo disminuirán, afectando aún más la liquidez de la empresa.

Imagen 33
Gráfico del desempeño de la empresa Security Data



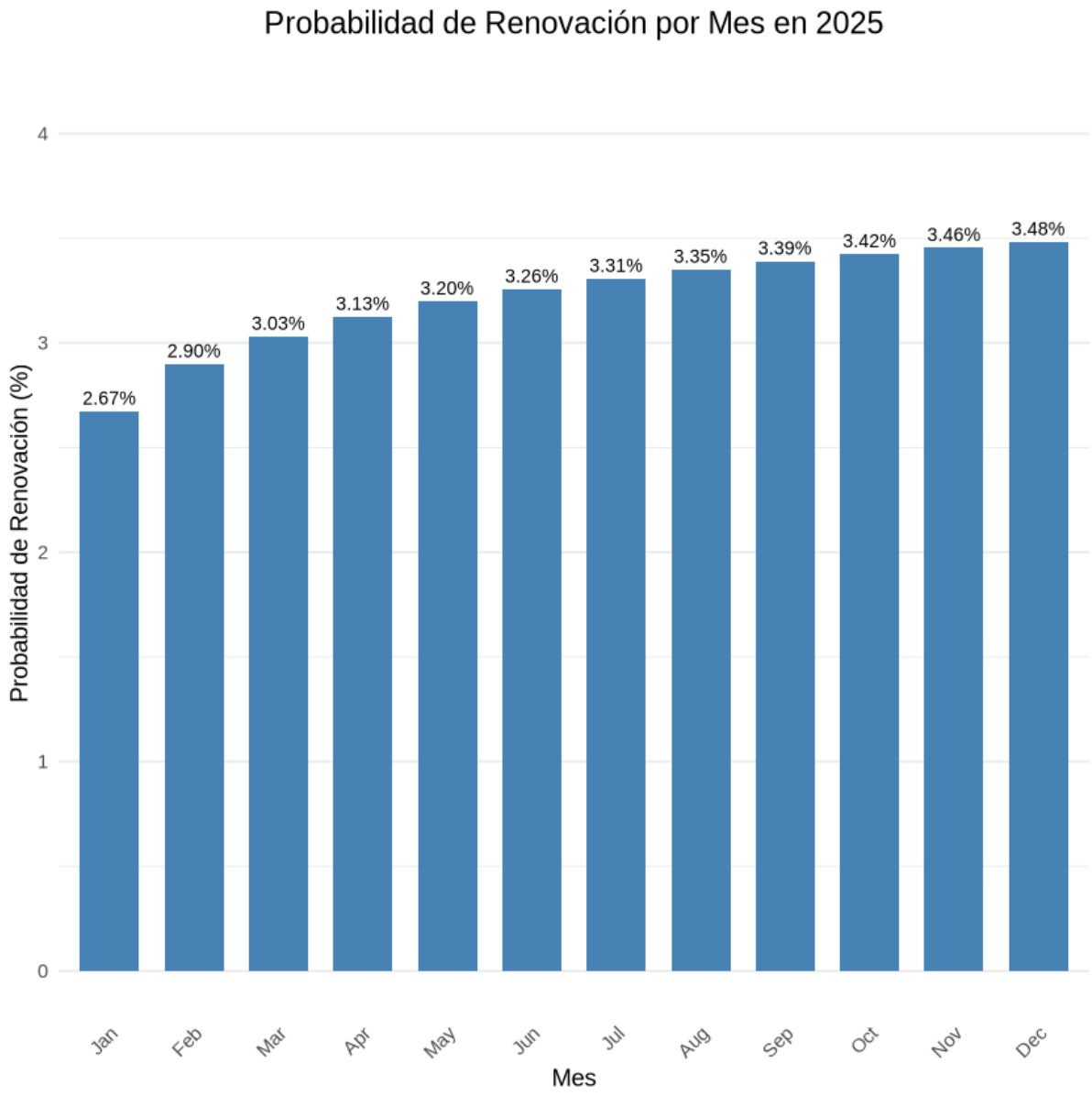
fuente:Elaboración propia

Con la mayoría de los clientes predichos como "No Renovación", los ingresos futuros en efectivo disminuirán, afectando aún más la liquidez de la empresa.

4.5.4. Escenario positivo para Renovaciones año 2025

Considerando los factores influyentes identificados en la data procesada, se sugiere priorizar a los clientes con firmas electrónicas nuevas, ya que son más propensos a realizar renovaciones. Además, se destaca que el tiempo transcurrido desde la compra inicial es un factor crucial en la predicción de renovaciones. Siguiendo las recomendaciones derivadas del caso estudiado, se plantea un escenario favorable para el año 2025, siempre que se implementen estrategias alineadas con estos hallazgos.

Imagen 34
Gráfico para predicciones positivas en el año 2025



Para materializar este escenario de crecimiento en renovaciones, es fundamental considerar varios aspectos clave basados en los datos procesados y el modelo predictivo. A continuación, se presenta un análisis detallado:

Factores Clave según la Importancia de Variables del Modelo:

La variable **TipoServi_nuvfirma** tiene un impacto significativo en este escenario, ya que permite fortalecer los vínculos con los clientes que adquirieron firmas nuevas en 2024. Esto implica:

- ➢ Fortalecer el servicio de firma nueva para garantizar su valor.
- ➢ Asegurar una experiencia excepcional durante la adquisición inicial.
- ➢ Implementar un seguimiento cercano en los primeros meses de uso.

Días desde la Compra:

Este factor fue crucial en los análisis estadísticos realizados con el modelo predictivo Random Forest. La importancia de este hallazgo refuerza la necesidad de implementar un programa de fidelización temprana, que incluya:

- ➢ Tomar medidas previas antes de la fecha de caducidad de la firma.
- ➢ Creación de alertas proactivas antes de fechas críticas, acompañadas de recompensas para incentivar la renovación.

Si bien es cierto que, para alcanzar un escenario favorable en el año 2025, la empresa debió iniciar acciones desde el presente año, este estudio representa una oportunidad valiosa para implementar estrategias efectivas con miras a un escenario favorable en el año 2026. Además, este trabajo plantea recomendaciones basadas en el análisis realizado y sugiere futuras investigaciones que permitan realizar ajustes y mejoras continuas en el proceso.

4.6. Evaluación del desempeño del producto/servicio de TI

Como parte de las pruebas del proyecto, se llevó a cabo una reunión con una experta en análisis de datos. Su objetivo fue brindar apoyo en el seguimiento y evaluación del desempeño del modelo desarrollado. Adicionalmente, se realizaron reuniones previas con la misma experta para obtener orientación antes de iniciar la elaboración del modelo y formular las preguntas pertinentes para guiar el análisis.

4.6.1. Calificación del modelo por la analista de datos

1) ¿Qué tan claro consideras que fue el objetivo del modelo desarrollado?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

2) ¿Qué tan adecuada te pareció la selección de las variables utilizadas en el modelo?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

3) ¿Qué tan confiable consideras que es el modelo para predecir los resultados deseados?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

4) ¿Qué tan satisfactorio fue el manejo de los datos previos al modelado (limpieza, balanceo, etc.)?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

5) ¿Qué tan adecuada te pareció la metodología utilizada para entrenar y probar el modelo?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

6) ¿Cómo evalúas la calidad de los gráficos y visualizaciones generados?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

7) ¿Qué tan útil consideras el modelo para la toma de decisiones en el contexto empresarial?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

8) ¿Qué tan claro fue el análisis presentado sobre el desempeño del modelo (precisión, recall, F1-score, curva ROC, etc.)?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

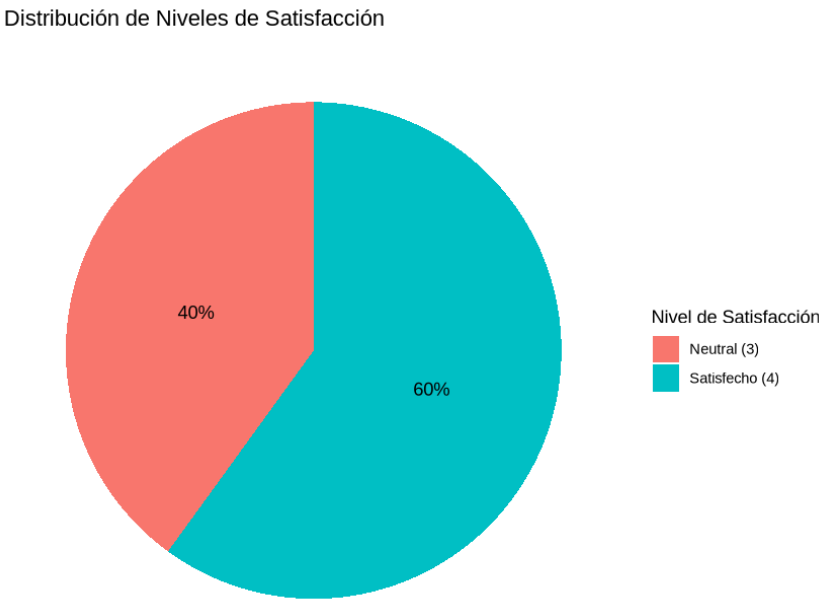
9) ¿Qué tan satisfecho estás con el proceso de identificación de zonas y su impacto en las predicciones?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

10) ¿Qué tan útil te pareció la reunión en términos de aclarar dudas y proporcionar sugerencias para mejorar el modelo?

1 = Muy Insatisfecho | 2 = Insatisfecho | 3 = Neutral | 4 = Satisfecho | 5 = Muy Satisfecho.

Imagen 35
Distribución del porcentaje de satisfacción del modelo



fuente:elaboración propia

Evaluación del Modelo Predictivo

El evaluador proporcionó una retroalimentación detallada sobre el modelo predictivo desarrollado, destacando los siguientes aspectos:

1. **Claridad del Objetivo**

   La evaluación del objetivo obtuvo una calificación de 4 ("Satisfecho"). El código demuestra un enfoque claro hacia el análisis predictivo y la clasificación, alineado con los objetivos de la investigación.

2. **Selección de Variables**

   Con una calificación de 4 ("Satisfecho"), se reconoció el adecuado manejo de variables y la inclusión de características relevantes para el modelo.

3. **Confiabilidad del Modelo**

   Este aspecto fue calificado con 3 ("Neutral"). Aunque se implementaron técnicas de validación, se identificaron oportunidades para mejorar la robustez del modelo, posiblemente mediante la incorporación de métodos adicionales de regularización o ajuste de hiper parámetros.

4. **Manejo de Datos Previos**

   La limpieza y preparación de datos recibieron una calificación de 4 ("Satisfecho"). Se destacó el uso de técnicas de balanceo de clases, como ROSE, combinado con el modelo Random Forest. Sin embargo, el evaluador sugirió comparar el desempeño del modelo con diferentes algoritmos para garantizar la selección más óptima.

5. **Metodología de Entrenamiento**

   Este punto fue evaluado con 3 ("Neutral"). Aunque la metodología es adecuada, el evaluador recomendó explorar técnicas más avanzadas de validación cruzada, como la validación estratificada o anidada, para aumentar la fiabilidad del modelo.

6. **Calidad de Gráficos**

   Las visualizaciones utilizadas obtuvieron una calificación de 4 ("Satisfecho"). Estas son informativas y bien estructuradas, lo que facilita la interpretación de los resultados.

7. **Utilidad para la Toma de Decisiones**

   Con una calificación de 3 ("Neutral"), se identificó que el modelo tiene potencial para la toma de decisiones empresariales, aunque requiere mayor validación para garantizar su aplicabilidad en escenarios reales.

8. **Claridad en el Análisis de Desempeño**

   Este aspecto fue calificado con 3 ("Neutral"). Aunque se incluyen métricas básicas, se sugirió ampliar el análisis con métricas adicionales como validación cruzada

9. **Identificación de Zonas**

   La segmentación y análisis espacial recibieron una calificación de 4 ("Satisfecho"), destacando el buen manejo de las zonas y su contribución al análisis predictivo.

10. **Utilidad General**

    Con una calificación de 4 ("Satisfecho"), el código se consideró sólido y con un alto potencial, aunque se identificaron áreas específicas de mejora.

Análisis Global de la Evaluación

De las evaluaciones realizadas, el 60% obtuvo la calificación de "Satisfecho" (4) y el 40% recibió la calificación de "Neutral" (3). Esto refleja un trabajo generalmente satisfactorio, con una base sólida para el análisis predictivo, pero con oportunidades claras para optimizar la confiabilidad, metodología y análisis de desempeño.

5. **Conclusiones**

- Se logró construir un modelo predictivo basado en el algoritmo Random Forest, que alcanzó un desempeño con métricas destacadas como un AUC del 93.20% y un recall elevado, superando el umbral esperado del 80% de precisión. Este resultado valida su efectividad para clasificar clientes entre renovaciones y deserciones. Además, se cumplió con el objetivo de utilizar datos sintéticos, garantizando el respeto a la normativa ecuatoriana de protección de datos personales

- El análisis del modelo permitió identificar factores determinantes de deserción, como zonas geográficas específicas (mayor riesgo en zonas comerciales como Albán Borja), y variables críticas como el tiempo desde la última renovación y la falta de incentivos personalizados. Estas observaciones se traducen en estrategias concretas para retener clientes en áreas estratégicas.

- El uso del modelo y las técnicas de minería de datos demostró ser una herramienta estratégica en la mejora de la retención de clientes. Se detectó que aproximadamente el 89% de los clientes en zonas comerciales presentan alta probabilidad de deserción, lo que valida la necesidad de una atención prioritaria y justifica la inversión en estrategias de fidelización.

- Aunque los datos sintéticos cumplieron con los estándares de representatividad, no capturaron del todo las complejidades del comportamiento real de los clientes. No obstante, su utilización permitió simular escenarios y garantizar el cumplimiento normativo, sentando un precedente para estudios similares en el contexto ecuatoriano.

- Uno de los principales desafíos encontrados fue la limitación en el uso de datos reales debido a la legislación de protección de datos personales en el Ecuador. Sin embargo, la creación y uso de datasets sintéticos brindó una oportunidad para simular y analizar el

comportamiento de los clientes de manera eficiente, a la vez que se cumplía con las regulaciones locales. Esta experiencia representa una oportunidad para futuras investigaciones en el sector de firmas electrónicas o de otro servicio, donde la adopción de minería de datos aún está en fases iniciales

➢ Otra observación importante de los datos sintéticos que aunque fueron diseñados de manera que simularan comportamientos reales de clientes de Security Data. Los resultados obtenidos fueron altamente representativos, lo que demuestra que, incluso con datos generados artificialmente, es posible obtener modelos predictivos eficaces en el ámbito de la retención de clientes y que se puede realizar estudios para empresas aun no haciendo uso de datos sensibles.

➢ La implementación de modelos estadísticos y la matriz de correlación permitió comprender mejor las relaciones entre diferentes variables, como la frecuencia de compras, historial de renovaciones, y zonas geográficas, las cuales influyen directamente en la tasa de deserción. Este análisis brindó una comprensión clara sobre qué factores se deben priorizar para optimizar la retención de clientes

➢ La alta tasa de deserción proyectada (aproximadamente el 89% de los clientes no renovarán sus contratos) podría agravar la situación financiera de la empresa, que ya presenta indicadores negativos como una disminución en ingresos, activos y patrimonio.

➢ Tras las reuniones realizadas con la experta en análisis de datos y la evaluación del desempeño del modelo predictivo, se obtuvo una retroalimentación integral que permitió validar su efectividad y utilidad en el contexto empresarial.

➢ La evaluación, estructurada en preguntas con una escala del 1 al 5, permitió identificar los puntos fuertes del modelo, así como áreas donde se pueden aplicar ajustes para aumentar su precisión y aplicabilidad práctica. En conclusión, el modelo tiene un gran

potencial para apoyar la toma de decisiones estratégicas en la empresa, y los aportes recibidos serán clave para afinar su rendimiento y maximizar su impacto.

- El uso de CRISP-DM permitió no solo comprender a fondo el contexto del negocio de Security Data, sino también identificar, preparar y analizar los datos de manera estratégica. En particular, la fase de modelado fue crucial, ya que permitió la implementación exitosa del algoritmo Random Forest, el cual alcanzó una precisión superior al 80% en la predicción de deserción y renovación de clientes. Este resultado fue posible gracias a la capacidad del enfoque CRISP-DM para guiar el preprocesamiento y la limpieza de datos, lo que garantizó que el modelo trabajara con información de alta calidad.

6. **Recomendaciones**

- Continuar experimentando con técnicas avanzadas de balanceo, como SMOTE o combinaciones personalizadas, para garantizar que las clases estén equitativamente representadas y evitar sesgos en las predicciones.

- Implementar una validación cruzada más exhaustiva para evaluar la estabilidad y la generalización del modelo en distintos subconjuntos de datos.
- Optimizar los hiperparámetros del modelo Random Forest (número de árboles, profundidad máxima, etc.) utilizando técnicas como Grid Search o Random Search para maximizar su desempeño, Incorporar más variables puede ayudar a detectar patrones más complejos que afecten la deserción
- Colaborar con el equipo de marketing y ventas para traducir los hallazgos del modelo en estrategias específicas, como promociones dirigidas a clientes con alto riesgo de no renovación.
- Implementar estrategias de retención diferenciadas basadas en el análisis geográfico de la deserción de clientes. Las zonas como Albán Borja muestran una mayor tasa de deserción, lo que sugiere que en esas áreas se deberían ofrecer descuentos personalizados o incentivos de renovación anticipada.
- Desarrollar un sistema de notificaciones automatizadas que contacte a los clientes cuyo servicio esté cerca de caducar o que no hayan renovado en el tiempo estipulado. Ofrecer incentivos personalizados basados en el historial de compras y el comportamiento de renovación.
- Ampliar el conjunto de datos para incluir variables temporales más detalladas, como la frecuencia de interacción con el portal de servicio, las tendencias de compra por temporada y el impacto de promociones especiales.
- Realizar encuestas y análisis cualitativos para comprender las razones por las cuales los clientes no están renovando sus servicios.

- Diseñar campañas específicas dirigidas a clientes cuyos servicios están próximos a caducar, ofreciendo los incentivos ya existentes en la compañía y un descuentos por renovación anticipada.
- Implementar herramientas automatizadas para enviar recordatorios y notificaciones personalizadas a los clientes sobre la caducidad de sus servicios, destacando los beneficios de la renovación.
- Utilizar chatbots o asistentes virtuales para brindar soporte inmediato a los clientes que tienen dudas sobre el proceso de renovación.
- Crear un programa de recompensas para clientes frecuentes o leales que renueven sus servicios de forma continua. Esto puede incluir puntos acumulables, descuentos exclusivos o acceso a servicios premium.
- Establecer un sistema de comunicación regular con los clientes durante la vigencia del servicio, promoviendo la satisfacción continua y recordándoles los beneficios de permanecer con la empresa.
- Diseñar campañas específicas para contactar a clientes con servicios caducados, ofreciendo condiciones especiales para su reactivación.
- Asegurarse de que los canales digitales, como el correo electrónico, las redes sociales y las aplicaciones móviles, sean efectivos y estén alineados con las preferencias de los clientes para maximizar el alcance de las campañas de renovación.
- Buscar alianzas con empresas de análisis de datos o plataformas de big data que puedan proporcionar información complementaria sobre las tendencias del mercado y los comportamientos de los clientes en el sector de firmas electrónicas.

7. BIBLIOGRAFÍAS

Actuaría Asesoramiento Estratégico. (s.f.). Actuaría asesoramiento estratégico. https://actuaria.com/

Agencia de Regulación y Control de las Telecomunicaciones. (2024). Informe de gestión año 2023. https://www.arcotel.gob.ec/wp-content/uploads/2024/01/informe_de_gestio%CC%81n_an%CC%83o_2023_con-formato-final-signed.pdf

Alarcón, R. E. (2021). Sistema analítico basado en un modelo predictivo de procesamiento de datos en la big data en la educación superior. Obtenido de [Tesis, Universidad Señor de Sipán]: https://repositorio.uss.edu.pe/ handle/20.500.12802/9040

Alonso, C. (2023c, septiembre 27). Claves de la Ley Orgánica de Protección de Datos Personales de Ecuador. GlobalSuite Solutions. https://www.globalsuitesolutions.com/es/claves-proyecto-ley-organica-proteccion-de-datos-personales-ecuador/?gad_source=1&gclid=Cj0KCQiAire5BhCNARIsAM53K1g9z67V6oAtUfQ2QevNNHX6BZh1XYX8V5GO0uoBbPjc0xq8lHOs8YoaAm_6EALw_wcB

Alsayat, A. (2023). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. Neural Computing and Applications, 35, 4701-4722.

Amazon Web Services, Inc. (s. f.). ¿Qué es la minería de datos? La minería de datos, explicada. https://aws.amazon.com/es/what-is/data-mining/#:~:text=La%20miner%C3%ADa%20de%20datos%20permite,la%20satisfacci%C3%B3n%20de%20los%20clientes.

Analytics India Magazine. (2021). What Is Predictive Model Performance Evaluation And WhyIsIt Important. Recuperado de https://analyticsindiamag.com

Appian. (2023, 15 mayo). Appian Process Mining: TKE's journey [Vídeo]. YouTube. https://www.youtube.com/watch?v=0fsdoXQViek

Arcitura Education Inc. (n.d.). Fundamentos de Big Data (Versión 1.7). www.arcitura.com A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly Media, Inc., 2019.

Arcotel. (n.d.). Listado de las entidades de certificación de información y servicios relacionados acreditados y terceros vinculados debidamente acreditadas. https://www.arcotel.gob.ec/listado-de-las-entidades-de-certificacion-de-informacion-y-servicios-relacionados-acreditados-y-terceros-vinculados-debidamente-acreditadas/

ARCOTEL. (s.f.). Listado de las entidades de certificación de información y servicios relacionados acreditados y terceros vinculados debidamente acreditados. https://www.arcotel.gob.ec/listado-de-las-entidades-de-certificacion-de-informacion-y-servicios-relacionados-acreditados-y-terceros-vinculados-debidamente-acreditadas/

Arroyo Ávila, J. R., Zamora López, J. A., Alvídrez Díaz, M. R. F., Viramontes Olivas, O. A., Domínguez Ríos, V. A., García Bencomo, M. I., Torralba Chávez, E., Aguirre Avilés, E. J., & Arizmendi Armendáriz, A. E. (2023). La inteligencia de negocios y la analítica de datos impulsando el desarrollo académico. Brazilian Journal of Implantology and Health Sciences, 6(3), 1225-1242. Recuperado de https://bjihs.emnuvens.com.br/bjihs/article/view/1677/1870

Asana. (2024). Gestión de clientes: Mejores prácticas para el manejo eficiente de tu cartera. Recuperado de https://asana.com/es/resources/client-management

Azevedo, A., & Santos, MF (2022). Ciencia de datos y análisis predictivo: técnicas, métodos y aplicaciones . Springer.

Banco Interamericano de Desarrollo. (2021). Informe sobre la digitalización en América Latina y el Caribe.

Biscobing, J. (2021). Almacén de datos (data warehouse). ComputerWeekly. Recuperado de https://www.computerweekly.com/es/definicion/Almacen-de-datos-data-warehouse

Bou-Hamad, I., & Jamali, I. (2020). Forecasting financial time-series using data mining models: A simulation study. Research in International Business and Finance, 51, (101072). https://doi.org/10.1016/j.ribaf.2019.101072

Cáceres, D. (1 de mayo de 2023). Datasets: Qué son y cómo acceder a ellos. Obtenido de OpenWebinars: https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/

Calisaya Choque, C. G. (2021). Modelo predictivo de regalías mineras aplicando técnicas de analítica predictiva con R. Postgrado en Informática, Universidad Mayor de San Andrés. Recuperado de https://ojs.umsa.bo/ojs/index.php/inf_fcpn_pgi/article/view/43

Carpio, F. J. (2021). Modelos predictivos de sistemas de información aplicados en la gestión en los abastecimientos de productos en las retailers del sector ferretero ubicados en la parroquia Rocafuerte de la provincia de Guayas. Universidad Técnica Estatal de Guayas.

Carrión, J. (s.f.). Diferencia entre dato, información y conocimiento. Universidad Autónoma de Baja California. https://iibi.unam.mx/voutssasmt/documentos/dato%20informacion%20conocimiento.pdf

Casado, S., & Giménez, J. (2021). Diseño y construcción de un almacén de datos. Módulo 1: Introducción al Data Warehouse. Universitat Oberta de Catalunya. https://openaccess.uoc.edu/bitstream/10609/136246/5/Disen%C2%BFo%20y%20construccio%C2%BFn%20de%20un%20almace%C2%BFn%20de%20datos_Mo%C2%BFdulo1_Introduccio%C2%BFn%20al%20Data%20Warehouse.pdf

Cedeño Troya, F., & Carpio Torres, F. (2022). Modelos predictivos de sistemas de información en la gestión de abastecimientos del sector ferretero. Revista Científica Ciencia y Tecnología, 22(34), 27-38. http://cienciaytecnologia.uteg.edu.ec

Centeno, A. (2020). Big Data. Técnicas de machine learning para la creación de modelos predictivos para empresas. Obtenido de [Tesis, Universidad Pontificia Comillas]: https://repositorio.comillas. edu/xmlui/handle/11531/45878

Ceupe. (2024, 9 abril). ¿Qué es la minería de datos y cuáles son sus aplicaciones? https://www.ceupe.com/blog/mineria-de-datos.html

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

Chitra, S., & Subashini, S. (s.f.). Data mining techniques and its applications in various fields. International Journal of Emerging Technology and Advanced Engineering, 10(1), 18-25. https://www.semanticscholar.org/paper/Data-Mining-Techniques-and-

its-Applications-in-Chitra-Subashini/87da9ade5ca2f0547c86e0d29f65ede5f544de90?p2df

Coussement, K., y Van den Poel, D. (2022). Calificación crediticia entre empresas y consumidores: uso de algoritmos de aprendizaje automático para mejorar la predicción de la solvencia crediticia de los consumidores . Journal of Business Research, 61(8), 757-763

Cuesta, F. (n.d.). Cómo gestionar la cartera de clientes. Cámara de Comercio de España. Recuperado de https://www.camara.es/sites/default/files/publicaciones/m_carteraclientes.pdf

Datademia. (2024, 13 mayo). ¿Cómo Netflix utiliza tus datos? [Vídeo]. YouTube. https://www.youtube.com/watch?v=ZPtz9DFOBMA

datos.gob.es. (2024, November 28). Datos sintéticos: ¿Qué son y para qué se usan? datos.gob.es. https://datos.gob.es/es/documentacion/datos-sinteticos-que-son-y-para-que-se-usan

Delgado, C., Morales, J., & Fernández, R. (2023). Enfoque cuantitativo y cualitativo: Una mirada de los métodos mixtos . Rhttps://www.re.norte/pub/374418696_ENFOQUE_CUANTITATIVO_y_CUALITATIVO_Una_mirada_de

Deloitte. (2020). Global human capital trends 2020: Leading the social enterprise: Reinvent with a human focus. Deloitte Insights. https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/human-capital/deloitte-cn-hc-trend-2020-en-200519.pdf

Dirección Nacional de Registros Públicos. (2021). Ley de protección de datos personales. Dirección Nacional de Registros Públicos. https://www.registrospublicos.gob.ec/programas-servicios/servicios/proyecto-de-ley-de-proteccion-de-datos

Duan, J., et al. (2020). A comprehensive survey on knowledge discovery and data mining: Algorithms, systems, and applications. ACM Computing Surveys, 53(2), 1-36.

Ecuavisa. (2019, 15 agosto). Entrevista exclusiva con Marc Randolph, cofundador de Netflix [Vídeo]. YouTube. https://www.youtube.com/watch?v=danpAkF3mjk

Ekcit, L. (2023, 28 noviembre). Analítica de datos (data analytics): ¿Cómo tener una visión global de los datos y las tendencias? TIC Portal. https://www.ticportal.es/glosario-tic/analitica-datos

EMIS. (n.d.). Security Data Seguridad en Datos y Firma Digital S.A. https://www.emis.com/php/company-profile/EC/Security_Data_Seguridad_en_Datos_y_Firma_Digital_SA_es_3971455.html

Gonzalo, Á. (2023). ¿Qué es R y para qué utilizarlo? Machine Learning para Todos. https://machinelearningparatodos.com/que-es-r-y-para-que-utilizarlo/#:~:text=R%20es%20un%20lenguaje%20interpretado,extensiones%20para%20ampliar%20sus%20capacidades.

Gopal, M. (2019). “An Introduction to Analytics.” Chap. 9.1 in Applied Machine Learning. 1st ed. New York: McGraw-Hill Education. https://www-

accessengineeringlibrarycom.ezproxy.itcr.ac.cr/content/book/9781260456844/tocchapter/chapter9/section/section2

Güçlü, B., & Yildirim, M. (2022). SWOT analysis: A comprehensive guide for startups. Journal of Entrepreneurship, Management and Innovation, 18(2), 25-48

Gupta, A., & Jain, R. (2023). Data analysis techniques in data mining: A review. International Journal of Data Science and Analytics, 16(1), 43-63.

Hawkins, D. M. (2019). Data Mining: Principles and Methods (2nd ed.). Springer.

Hernández Gómez, G. (2019). Minería de datos para la clasificación y predicción de enfermedades utilizando técnicas de aprendizaje supervisado y no supervisado [Tesis de maestría, Centro de Investigación en Matemáticas A.C.]. Repositorio Institucional CIMAT. https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/1129/1/TE%20835.pdf

Huang, J. (2019). Análisis de la gestión de membresías de una marca de moda mediante tecnología de big data. American Journal of Industrial and Business Management, 9(10), 1931-1948. https://doi.org/10.4236/ajibm.2019.910126

IBM. (2024, 28 junio). Minería de datos. IBM. https://www.ibm.com/mx-es/topics/data-mining?mhsrc=ibmsearch_a&mhq=mineria%%2020de%20dato

IBM. (2024). Introduction to time series forecasting. IBM Documentation. https://www.ibm.com/docs/es/spss-statistics/saas?topic=forecasting-introduction-time-series

IBM. (s.f.). ¿Qué es un almacén de datos? https://www.ibm.com/topics/data-warehouse

IBM. (s. f.). Analítica predictiva. Recuperado el 12 de octubre de 2024, de https://www.ibm.com/es-es/topics/predictive-analytics

IBM. (s.f.). Data warehouse. https://www.ibm.com/mx-es/topics/data-warehouse

IBM. (s.f.). Redes neuronales. https://www.ibm.com/es-es/topics/neural-networks

Inesdi. (s.f.). Random forest: ¿Qué es?. https://www.inesdi.com/blog/random-forest-que-es/#:~:text=Random%20forest%20es%20un%20algoritmo,promedio%20en%20caso%20de%20regresi%C3%B3n

Ingelsi Cia Ltda. (s.f.). Nuestros servicios - Ingelsi Cia Ltda: Inteligencia de datos en Ecuador. https://www.ingelsi.com.ec/about-us/

Instituto de Ingeniería del Conocimiento (IIC). (s. f.). Analítica descriptiva. Recuperado el 12 de diciembre de 2024, de https://www.iic.uam.es/big-data/analitica-descriptiva/#:~:text=La%20anal%C3%ADtica%20descriptiva%20consiste%20en,actual%20y%20pasado%20del%20negocio.

Johnson, M., & Gupta, R. (2020). Data-Driven Decision Making: Unlocking the Value of Big Data. Journal of Information Systems, 34(2), 123-135.

Khan, F. (2024, 23 julio). Best Data Mining Tools in 2024 | Astera. Astera. https://www.astera.com/es/type/blog/data-mining-tools/

Leffingwell, D. (2018). SAFe 4.5 Distilled: Applying the Scaled Agile Framework for Your Enterprise. Addison-Wesley.

Ley de Comercio Electrónico, Firmas y Mensajes de Datos. (2002). Registro Oficial Suplemento 557 de 17-abr-2002.

Ley de Protección de Datos Personales - Dirección Nacional de Registros Públicos. (2024b, abril 16). Dirección Nacional de Registros Públicos. https://www.registrospublicos.gob.ec/programas-servicios/servicios/proyecto-de-ley-de-proteccion-de-datos

Madaan, R., & Kumar, K. (2020). Prevalence of Visualization Techniques in Data Mining. En J Hemanth, M Bhatia, O Geman. (Eds), Data Visualization and Knowledge Engineering. Lecture Notes on Data Engineering and Communications Technologies (pp. 273-298). Springer. https://doi.org/10.1007/978-3-030-25797-2_12

Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2024). Chequeo Digital 2022 - 2023: El aprendizaje en la madurez digital en Ecuador. https://observatorioecuadordigital.mintel.gob.ec/wp-content/uploads/2024/09/CHEQUEO-DIGITAL.pdf

Ministerio de Telecomunicaciones y de la Sociedad de la Información. (n.d.). ¿Qué es FirmaEC? https://www.firmadigital.gob.ec/que-es-firmaec/

Ministerio de Telecomunicaciones y de la Sociedad de la Información. (n.d.). Ley de Comercio Electrónico, Firmas y Mensajes de Datos. https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2012/11/Ley-de-Comercio-Electronico-Firmas-y-Mensajes-de-Datos.pdf

Naeem, T. (2024, 3 septiembre). Understanding Structured, Semi-Structured, and Unstructured Data. Astera. https://www.astera.com/es/type/blog/structured-semi-structured-and-unstructured-data/

Nisbet, R., Elder, J., & Miner, G. (2019). Handbook of Statistical Analysis and Data Mining Applications (2nd ed.). Academic Press.

O. O. Olaniyi, A. Abalaka, and S. O. Olabanji, “Utilizing Big Data Analytics and Business Intelligence for Improved Decision-Making at Leading Fortune Company.” Sep. 14, 2023. Accessed: Jul. 09, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4571876

Oraculo. (2021, 25 febrero). La minería de datos debe cumplir con la Ley de Protección de Datos Personales - Dirección Nacional de Registros Públicos. Dirección Nacional de Registros Públicos. https://www.registrospublicos.gob.ec/la-mineria-de-datos-debe-cumplir-con-la-ley-de-proteccion-de-datos-personales/

Oraculo. (2021b, febrero 25). La minería de datos debe cumplir con la Ley de Protección de Datos Personales. Dirección Nacional de Registros Públicos. https://www.registrospublicos.gob.ec/la-mineria-de-datos-debe-cumplir-con-la-ley-de-proteccion-de-datos-personales/#:~:text=%E2%80%9CComo%20en%20cualquier%20lugar%20del,en%20la%20Ley%E2%80%9D%2C%20puntualiz%C3%B3.

Orellana Nirian, P., & López, J. F. (2020, enero 1). Cartera de clientes. Blog en Economipedia. Recuperado de https://economipedia.com/definiciones/cartera-de-clientes.html

Ortega, C. (2023, enero 23). Fuga de clientes: Qué es y métodos para evitar que los clientes se vayan. QuestionPro. https://www.questionpro.com/blog/es/fuga-de-clientes/

Ortega, C. (2023a, enero 23). Churn rate de clientes: Qué es, importancia y cómo evitarlo. QuestionPro. https://www.questionpro.com/blog/es/churn-rate-de-clientes/

Paz, M. (2022b, septiembre 22). Qué es analítica de datos. Importancia para las empresas. Cidei. https://cidei.net/que-es-analitica-de-datos/

Primicias. (2024, 25 de noviembre). Ecuador tiene dos formatos de firma electrónica: archivo y token. Recuperado de https://www.primicias.ec/noticias/tecnologia/ecuador-dos-formatos-firma-electronica/

PwC. (n.d.). Todo lo que debes conocer sobre la protección de datos personales. https://www.pwc.ec/es/entrevistas-de-temas-de-interes/todo-lo-que-debes-conocer-sobre-la-proteccion-de-datos-personales.html

R Core Team. (s. f.). R: A language and environment for statistical computing. Recuperado el 12 de diciembre de 2024, de https://www.r-project.org/

Registro Civil, Identificación y Cedulación. (n.d.). *La firma electrónica se empezó a utilizar en el Ecuador con fines tributarios*. Recuperado el 16 de diciembre de 2024, de https://www.registrospublicos.gob.ec/la-firma-electronica-se-empezo-a-utilizar-en-el-ecuador-con-fines-tributarios/

Rodríguez, D., & Rodríguez, D. (2024, 25 agosto). Minería de datos: Una guía esencial 2024. Mercately. https://blog.mercately.com/marketing/mineria-de-datos

RStudio. (n.d.). RStudio Desktop. Softonic. https://rstudio-desktop.softonic.com/

Sas. (2021). Analítica predictiva. Obtenido de https://www.sas.com/es_mx/insights/analytics/predictive-analytics.html#:~:text=La%20 anal%C3%ADtica%20predictiva%20es%20el,que%20suceder%C3%A1%20en%20el%20 futuro.

SAS. (s. f.). Minería de datos: Qué es y por qué es importante. https://www.sas.com/es_mx/insights/analytics/data-mining.html

SAS Institute Inc. (2021). SAS Enterprise Miner 14.3: User's Guide. SAS Institute.

Security Data. (n.d.). ¿Qué es la firma electrónica y para qué sirve? Recuperado el 25 de noviembre de 2024, de https://www.securitydata.net.ec/que-es-la-firma-electronica-para-que-sirve/

Security Data. (n.d.). ¿Qué es la firma electrónica y para qué sirve? Recuperado el 25 de noviembre de 2024, de https://www.securitydata.net.ec/que-es-la-firma-electronica-para-que-sirve/

Security Data. (s.f.). Nosotros: Security Data Ecuador. https://www.securitydata.net.ec/nosotros-security-data-ecuador/

Smith, J. (2021). Advanced Data Mining Techniques in Business Analysis. Journal of Business Analytics, 15(3), 45-60.

Sprinklr. (2024). Customer lifecycle. https://www.sprinklr.com/blog/customer-lifecycle/

Sprinklr. (2024). Customer lifetime value. https://www.sprinklr.com/cxm/customer-lifetime-value/

Superintendencia de Economía Popular y Solidaria. (2021). Ley orgánica de protección de datos personales. https://www.finanzaspopulares.gob.ec/wp-content/uploads/2021/07/ley_organica_de_proteccion_de_datos_personales.pdf

Sutherland, J., & Schwaber, K. (2020). The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game. Recuperado de https:

Torres Quezada, Y. S. (2020). [Título de la tesis] (Tesis de grado, Universidad Nacional de Loja). Repositorio Digital Universidad Nacional de Loja. https://dspace.unl.edu.ec/jspui/browse?type=author&value=Torres+Quezada%2C+Yulissa+Stefania

Vela López, M. (2022). Implementación de un modelo de minería de datos para predecir la deserción de los clientes en una empresa de telecomunicaciones (Tesis de maestría, Universidad de Santo Tomás). https://tesis.usat.edu.pe/bitstream/20.500.12423/5361/8/TL_VelaLopezMirko.pdf

Vera, A. (2023). La IA predictiva frente a la IA generativa en OpenText DevOps Cloud. Obtenido de https://discoverthenew.ituser.es/gestion-deapps/2023/09/la-ia-predictiva-frente-a-la-iagenerativa-en-opentext-devops-clou

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A. L., Tsiros, M., & Schlesinger, L. A. (2021). "Customer experience creation: Determinants, dynamics and management strategies". Journal of Retailing, 97(4), 457-473.

Worsley, S. (julio de 2024). ¿Qué es R? Introducción a la potencia del cálculo estadístico. Obtenido de DataCamp: https://www.datacamp.com/es/blog/all-about-r

Xia, R., & Li, J. (2022). A review of knowledge discovery in databases: Issues and challenges. IEEE Access, 10, 6000-6011.

Zendesk. (2023, febrero 14). Gestión de cartera de clientes: 6 tips para un buen manejo. https://www.zendesk.com.mx/blog/gestion-de-cartera-de-clientes/

8. ANEXOS

8.1. Carta de autorización

SECURITYDATA
TU IDENTIDAD DIGITAL, EN UNA FIRMA.

CONFIDENCIAL

ANEXO No. 5

Samborondón, jueves 12 de diciembre del 2024

Magíster

Erika Ascencio

Universidad Tecnológica ECOTEC

De mis consideraciones:

A través del presente, autorizo al señorita **JIMÉNEZ GALÁN NANCY ANGELICA,** con cédula de ciudadanía n° 0943709626 y código estudiantil 2018291675, respectivamente, estudiantes de la Unidad Académica **PROYECTO INTEGRADOR**, de la carrera TECNOLOGÍAS DE LA INFORMACIÓN de la Universidad Ecotec para que pueda hacer uso del nombre de nuestra empresa como **caso de estudio** para su proyecto de titulación llamado DESARROLLO DE UN MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA ANTICIPAR EL CRECIMIENTO DE LA CARTERA DE CLIENTES EN UNA EMPRESA PROVEEDORA DE SERVICIOS DE FIRMA ELECTRÓNICA.

El proyecto tiene como objetivo desarrollar una base de datos sintética mediante la observación de características del negocio de firmas electrónicas, con el fin de modelar y desarrollar su trabajo de titulación, garantizando que los datos utilizados no vulneraran la confidencialidad ni la integridad de la información original de la empresa ni serán utilizados datos sensibles. Asimismo, autorizamos la divulgación y publicación de los resultados de su investigación en los repositorios que la Universidad Ecotec tenga destinados para este fin, con la condición de que los datos sean representaciones sintéticas o accesibles públicamente.

Atentamente,

CRISTIAN ADRIAN PAZMINO CORTEZ

Firmado digitalmente por CRISTIAN ADRIAN PAZMINO CORTEZ
Fecha: 2024.12.16 16:52:34 -05'00'

Cristian Pazmiño

Security Data

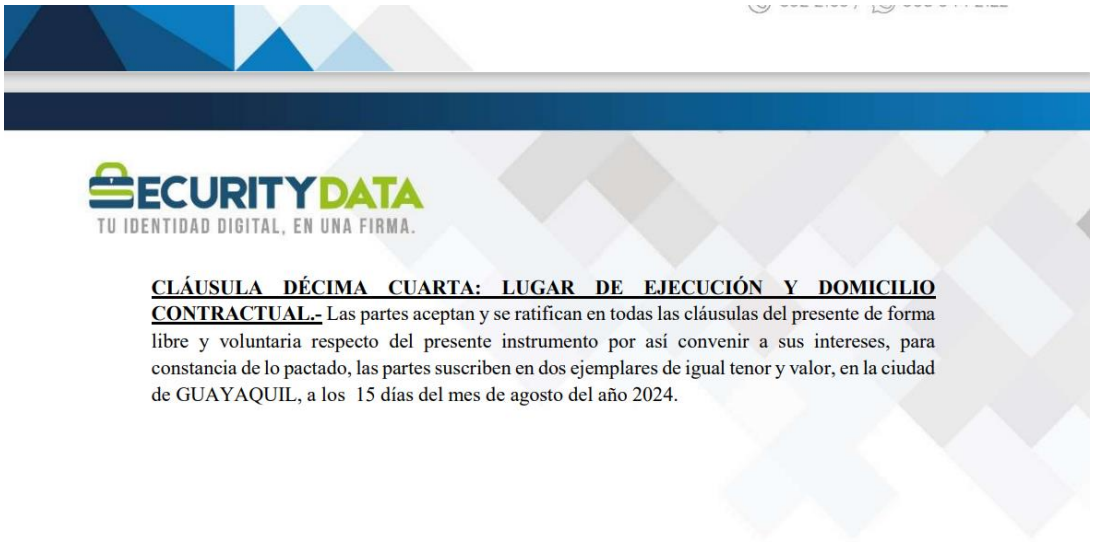
Supervisor Técnico

0969034182

8.2. Cláusulas de confidencialidad

Imagen #35

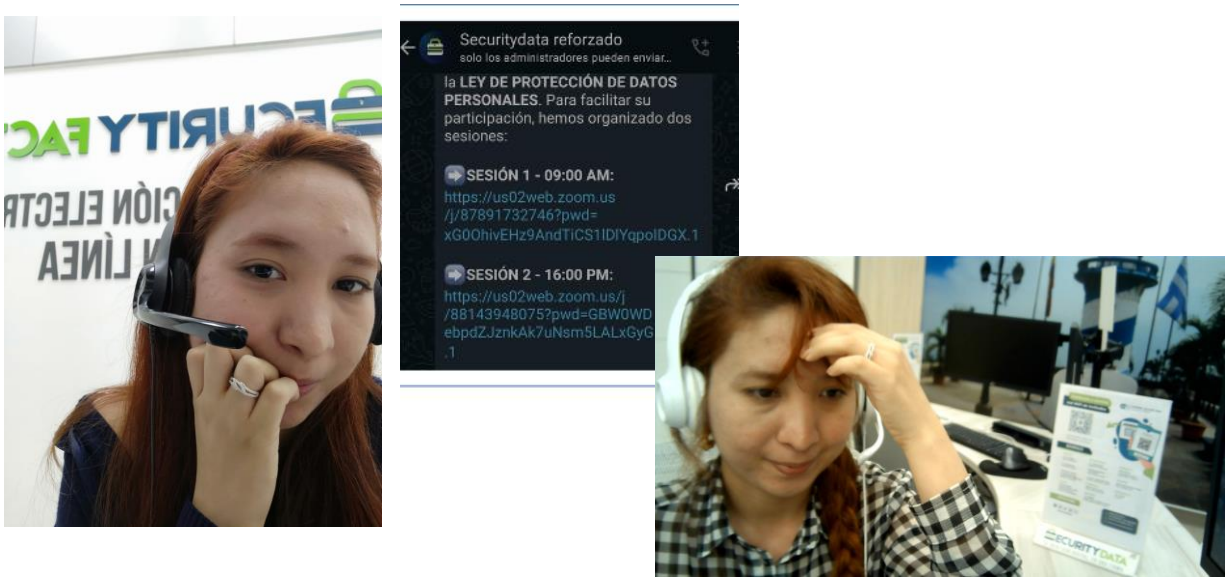
cláusula de confidencialidad



Queda claro que para la elaboración del caso de estudio se respetó todas las cláusulas que maneja la empresa.

8.3. Fase de comprensión del negocio

imagen #36
Comprensión del negocio



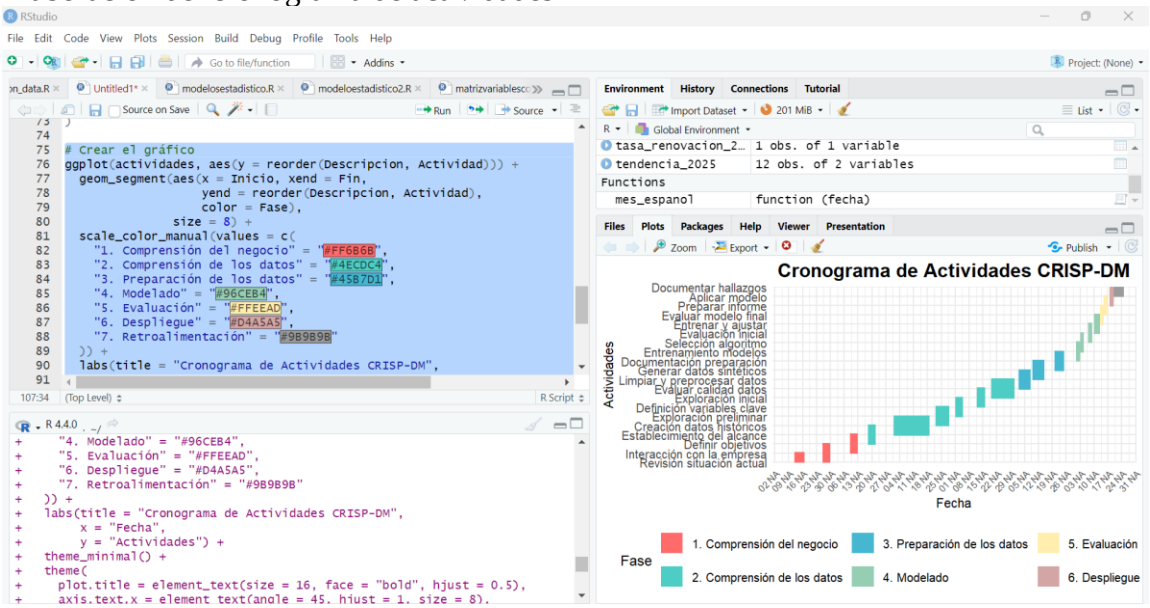
fuente: elaboración propia

Paso importante para la creación de los datos sintéticos y entender cómo es la ley de protección de datos en las empresas

8.4. Elaboración de actividades usando CRISP-DM en Rstudio

Imagen 37

Elaboración del cronograma de actividades



fuente: Elaboración propia

Tabla 13

Cronograma de actividades

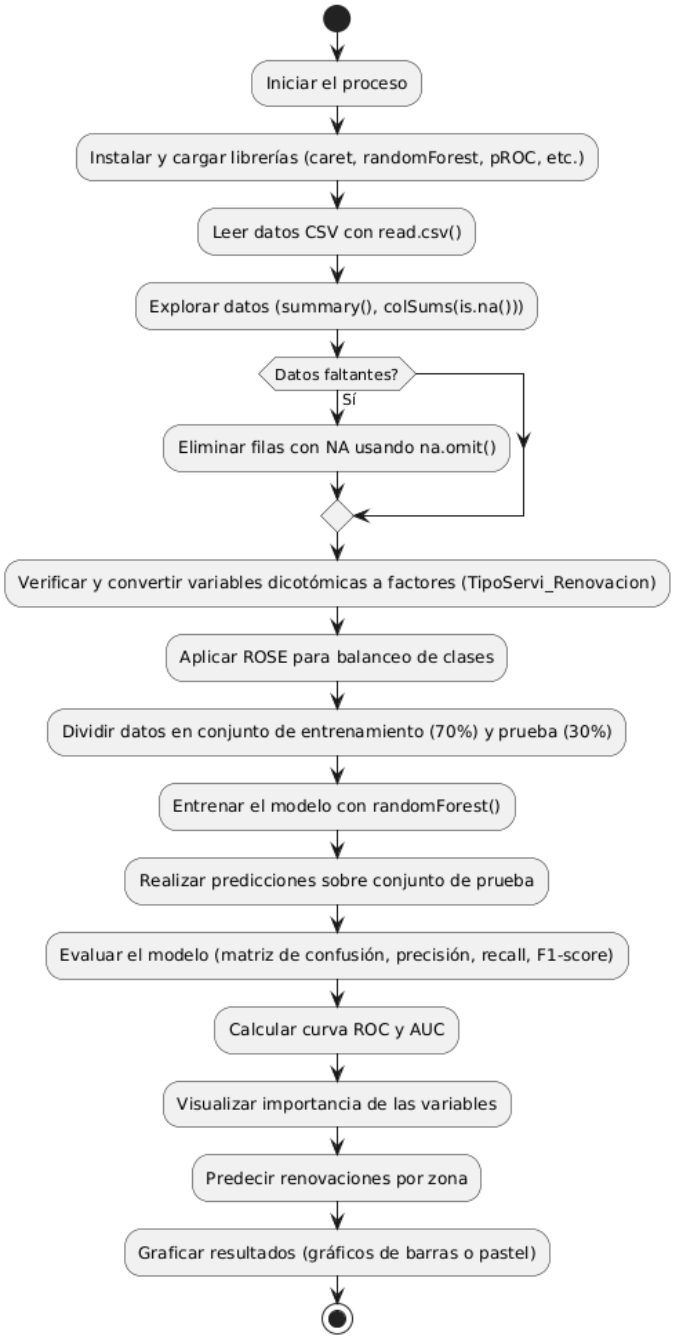
Fase de CRISP-DM	Actividad/Descripción
Fase 1 . Comprensión del negocio	- Revisión de la situación actual de la empresa y análisis del mercado. mayo del 12 al 17 - Interacción con la empresa para entender el negocio para el caso de estudio.26 al 30 de mayo - Definir los objetivos del proyecto y el problema a resolver. 9 al 13 de junio
fase 2. Comprensión de los datos	- Establecimiento del alcance del proyecto y los recursos necesarios. 18 de junio al 22 de junio - Creación de datos históricos sintéticos. 1 de julio al 19 de julio - Exploración preliminar de los datos para identificar calidad y consistencia. 22 de julio al 29 de julio - Definición de variables clave para el análisis predictivo. 1 de agosto al 5 de agosto - Realizar una exploración inicial de los datos para identificar patrones y tendencias.12 de agosto al 16 de agosto - Evaluar la calidad de los datos y detectar problemas. 19 de agosto al 31 de agosto
Fase 3. Preparación de los datos	- Limpiar y preprocesar los datos (anonimización, eliminación de duplicados, etc.). del 2 al 8 de septiembre - Generar datos sintéticos adicionales para el conjunto de datos con la creación de nuevas variables derivadas y realizar transformaciones necesarias. del 9 al 15 de septiembre - Documentación del proceso de preparación de datos. del 20 al 25 de septiembre
Fase 4. Modelado	- Entrenamiento de modelos utilizando el conjunto de datos de entrenamiento. del 1 al 3 de octubre -Seleccionar el algoritmo mediante evaluación de la literatura del 3 de octubre al 5 de octubre - Evaluación inicial del modelo para determinar su rendimiento elección random forest 7 al 9 de octubre - Entrenar el modelo y ajustar parámetros.10

	al 13 de octubre
Fase 5. Evaluación	- Evaluar el modelo utilizando métricas como precisión, recall, F1-score y AUC.stakeholders para determinar el modelo más adecuado.del 13 de octubre al 15 de octubre -Preparar un informe de evaluación del modelo y su desempeño. del 15 de octubre al 17 de octubre
Fase 6.Despliegue	-Aplicar el modelo en un escenario de renovación y no renovación aplicar en un escenario favorecedor del 18 al 20 de octubre
7.retroalimentación	- documentar los hallazgos de los resultados , realizar las recomendaciones a base de los resultados de igual manera las conclusiones del 20 de octubre al 25 de octubre

fuente: Elaboración propia

8.4.1. Diagrama de flujo del Modelado random forest

Imagen 38
Diagrama de flujo del Desarrollo del modelo en plantuml



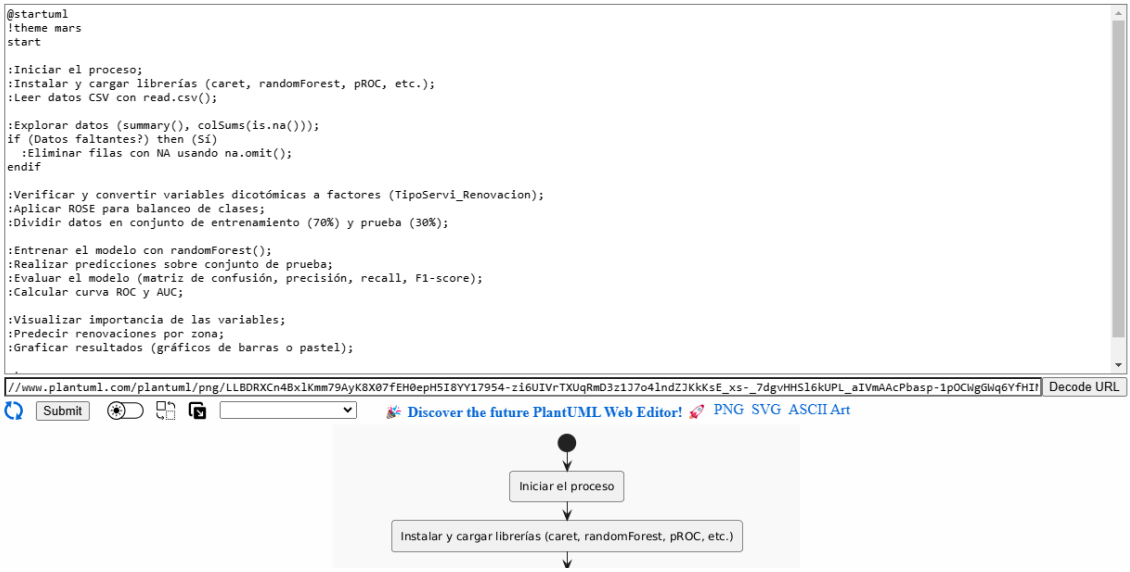
fuente: Elaboración propia

//www.plantuml.com/plantuml/png/LLBDRXCn4BxlKmm79AyK8X07fEH0epH5I8YY17954-zi6UIVrTXUqRmD3z1J7o4lndZJKkKsE_xs-_7dgvHHSl6kUPL_aIVmAAcPbasp-1pOCWgGWq6YfHINTQd7JhVtO57kT71y49BJ0oOmkg8y0y7GHNyJXP9-

[3TlRzGmestch15-813hCCS5wjmSR0mXXDxTfD7hUB3P_1XS5dq4c5Q-sxaqxKx3R5PyCftb0qxRj4XhkmLnFo1vTnf0fVMn1umGmkzD3sm0iDeuz1wNisM6QDB-keAHg4mBEe-TSnIbqtAk5FGdtRCyfOnX9CWkCA8m7HmawjZ6V7hr24g0AsropWldEGzoHZFnpIo6EQ3c66derk8bkUxlRm822S427mLA4Zi2gAKeAk-QHEtvEJg6A_ouXJoWAMIYWPweBy-7jwrRj3LBeW63Uw-TqUsVKr9gF7RbOIIuQCTNHbj3n4KK9IEDKeneWHUtoGlNClrJY4Lsvu3KUi_1nIX13NnAV7iEi4bf-cdL4vsPmy-vDiYfUbTVeR77AP8kC2FeuDCRgnxgwttCgJxxO3r6qJyjO5HocbmAKvPlQLXq1-N_Hwbz_WMCCrUydmNE56hcuZ5rzItTo-gjR7PNoW2BA6hMGbCdLwqiv3iqLXQvuzmy0](#)

Imagen 41

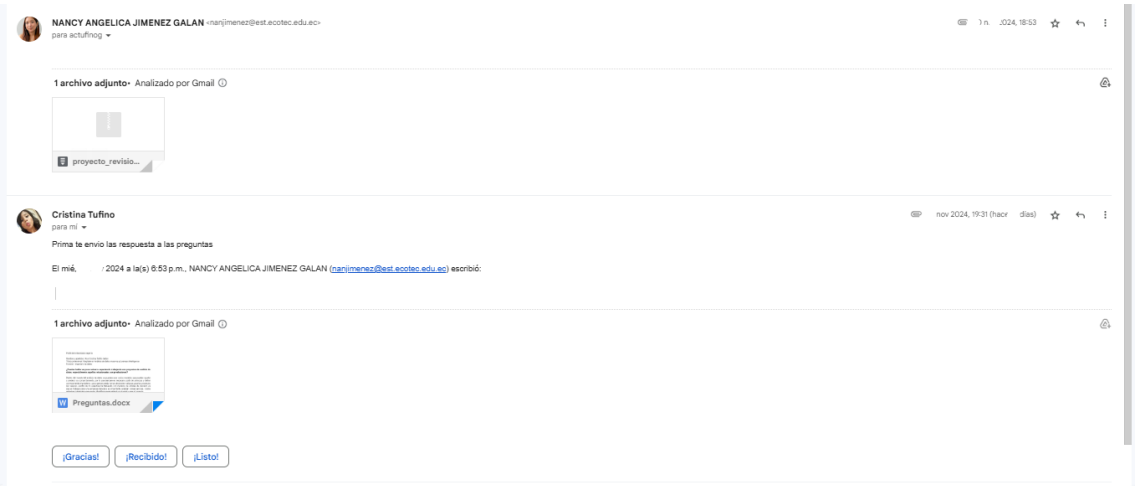
Código Desarrollo del modelo en plantuml



fuente: Elaboración propia

8.5. Entrevista con profesional en Datos masivos y evaluación del modelo para el caso de estudio

Imagen 39
Envío de preguntas para entrevista



fuente: Elaboración propia

Nombre y apellidos: Ana Cristina Tufiño Galan
Título profesional: Magíster en Análisis de datos masivos y business Intelligence
Función: Ingeniero de Datos en Diners Club del Ecuador
Perfil Profesional : https://www.linkedin.com/in/ana-cristina-tufi%C3%B1o-gal%C3%A1n-3a4502182/

Entrevista
fecha y hora: 1 de Octubre del 2024, 12h00.
Forma: Vía Google Meet Tiempo de entrevista: 60 minutos

Entrevistador: Nancy Angelica Jimenez Galan

¿Puedes hablar un poco sobre tu experiencia trabajando con proyectos de análisis de datos, especialmente aquellos relacionados con predicciones?

El entrevistado explica que en el análisis de datos es crucial partir de premisas claras y definir el problema de manera precisa para tomar decisiones valiosas para el negocio. Mencionó su experiencia trabajando con modelos de árboles de decisión, especialmente en el sector bancario, donde es esencial analizar consecuencias y costos, optimizar estrategias y detectar oportunidades relevantes para el negocio.

¿Cómo describirías tu enfoque al abordar un nuevo proyecto de análisis de datos desde la fase inicial hasta la entrega de resultados?

para esta pregunta el entrevistado contestó

Antes de iniciar un nuevo proyecto siempre se debe tener claro el problema a resolver, considerando las técnicas y su enfoque metodológico de tal manera que se determine claramente la información y el tipo de datos a usar dentro del análisis.

Se debe realizar una recolección de datos y de ser el caso la limpieza de los mismos de tal manera que que toda la información se unifique en una sola fuente y su uso sea lo más manipulable posible, considerando siempre los datos sensibles de una empresa, es decir cuáles se podría anonimizar ya que trabajar con información privada de una persona es puede involucrar problemas legales

Realizar una análisis exploratorio de los datos es decir encontrar patrones, correlaciones, distribución de variables con el objetivo de entender el comportamiento de los datos para poder aplicar el mejor modelo para poder predecir.

Escoger el mejor modelo dependiendo del problema a resolver , esta parte es fundamental ya que se entrenará al modelo ya que aquí se identificará la presión de los datos.

Ingreso de nuevos datos al modelo, después de realizar el modelo, se debe trabajar al 100 % para entrenar el modelo ya que si uno de los alcances del proyecto es consumir data en tiempo real se puede realizar sin ningún problema.

¿Qué herramientas y tecnologías utilizas con más frecuencia en tu trabajo diario (por ejemplo, Python, R, SQL, etc.) y por qué?

El entrevistado mencionó que utiliza herramientas como Python y SQL, ya que son fundamentales para realizar análisis de datos de manera eficiente y efectiva. Con 7 años de experiencia en el mercado, considera que estos lenguajes son clave para llevar a cabo proyectos de análisis complejos.

En su experiencia, ¿Qué tipos de algoritmos considera más convenientes para desarrollar un modelo predictivo para anticipar la cartera de clientes de una empresa?

Según el entrevistado, el uso de regresión logística y árboles de decisión es adecuado, especialmente en el sector bancario, donde se analiza el comportamiento histórico de los clientes, como el consumo de tarjetas de crédito y la identificación de patrones relevantes.

Preprocesamiento de Datos

¿Cómo manejas los valores atípicos en los conjuntos de datos? ¿Utilizas alguna técnica específica para detectarlos y corregirlos?

El entrevistado explicó que se pueden utilizar técnicas como el uso de cuartiles y desviaciones estándar para detectar y corregir valores atípicos. Dependiendo del objetivo del análisis, se puede optar por eliminar o reemplazar estos valores para evitar que afecten la interpretación de los resultados.

¿Qué importancia le das a la normalización o estandarización de los datos antes de entrenarlos un modelo ?

El entrevistado resaltó que la normalización y estandarización son cruciales para el preprocesamiento de datos, ya que impactan directamente en la precisión, velocidad y estabilidad del modelo.

¿Qué técnicas de minería de datos pueden verse afectadas por la anonimización de datos?

Aunque el entrevistado mencionó no haber trabajado con todas las técnicas de minería de datos, indicó que la anonimización de datos podría reducir la capacidad de procesamiento y dificultar la creación de predicciones precisas, así como la identificación de relaciones significativas.

¿Es posible utilizar datos anonimizados para construir modelos predictivos con un alto nivel de precisión?

El entrevistado afirmó que sí, la anonimización de datos no impide la construcción de modelos precisos. Si se realiza de manera adecuada, puede proteger la privacidad sin comprometer la precisión de las predicciones.

¿Podría compartir ejemplos de proyectos exitosos donde se haya utilizado la anonimización de datos?

Un ejemplo proporcionado fue el uso de datos anonimizados por empresas como Netflix o Spotify, que utilizan la información de los usuarios para hacer recomendaciones personalizadas sobre películas o música.

Modelado y Algoritmos

¿Cómo decides qué modelo usar para un problema específico, ejemplo a la predicción de renovaciones y no renovaciones?

El entrevistado explicó que, para problemas de predicción, primero se define claramente el problema y las soluciones posibles. En el caso de renovaciones y no renovaciones, utilizaría árboles de decisión debido a su capacidad para manejar múltiples posibilidades y generar soluciones basadas en premisas.

¿Cómo evalúas el rendimiento de un modelo? ¿Qué métricas de evaluación consideras más relevantes en este tipo de problemas

El entrevistado detalló que, dependiendo de la naturaleza del problema, las métricas adecuadas son cruciales. Para problemas de clasificación, utilizaría el F1 Score, mientras que para problemas de regresión se inclinaría por el MSE. Es importante elegir las métricas correctas según los objetivos y el tipo de modelo.

¿Has usado Random Forest frente a otros modelos , cuando crees que es preciso utilizarlos en qué escenarios?

El entrevistado mencionó que ha utilizado Random Forest, que es un algoritmo basado en árboles de decisión, y es adecuado para clasificación y regresión debido a su capacidad para manejar una gran variedad de datos.

¿Cómo interpretas los resultados de un modelo? ¿Qué tan fácil es para ti extraer insights útiles en los proyectos?

Con 7 años de experiencia, el entrevistado destacó que le resulta fácil levantar requerimientos y entregar respuestas precisas a los clientes. Aunque cada caso es único,

su experiencia como analista de datos le permite interpretar y extraer insights útiles de manera eficiente.

¿Utilizas alguna técnica para interpretar modelos complejos?

El entrevistado mencionó que emplea técnicas como la "Feature Importance", que permite evaluar la relevancia de cada variable en la predicción del modelo. Esta técnica ayuda a identificar las características clave que influyen en las decisiones del modelo.

Resultados y Aplicación en el Negocio

¿Cómo evalúas la relevancia de los resultados obtenidos del modelo para el negocio? ¿Cómo te aseguras de que las predicciones sean útiles y aplicables?

El entrevistado subrayó la importancia de evaluar la precisión técnica y la capacidad de los resultados para mejorar los objetivos clave del negocio. Para garantizar que las predicciones sean útiles, se combinan métodos técnicos, pruebas con datos reales y una estrecha colaboración con las partes interesadas.

¿Has trabajado en proyectos similares donde los resultados del análisis llevaron a cambios importantes en la estrategia de la empresa?

El entrevistado afirmó que sí, los resultados de sus proyectos han influido en la toma de decisiones clave dentro de la empresa, destacando el valor que aporta tanto al equipo de trabajo como a los clientes.

¿Cómo te mantienes actualizado con las últimas tendencias y herramientas en el campo del análisis de datos?

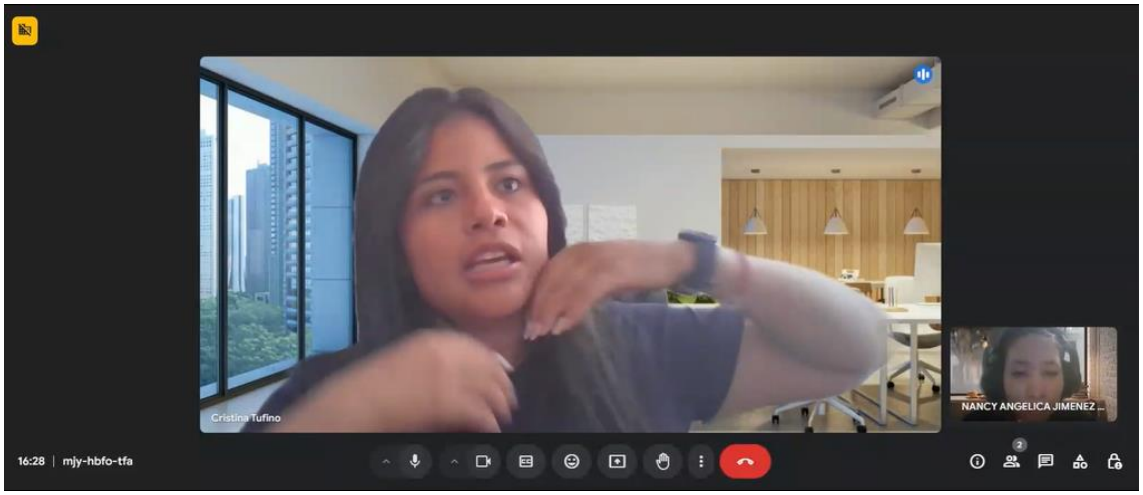
El entrevistado se mantiene actualizado mediante cursos, publicaciones y el seguimiento de nuevas tecnologías en el campo del análisis de datos.

¿Qué consejo le darías a alguien que está comenzando en el campo del análisis de datos y quiere trabajar en proyectos de predicción y modelado?

El entrevistado aconsejó que, aunque el camino es largo, siempre hay que estar aprendiendo, ya que la tecnología cambia constantemente. El análisis de datos trae constantemente innovaciones y oportunidades para aportar valor al mundo

Imagen 40

Imagen de la reunión respecto a los resultados e indicaciones



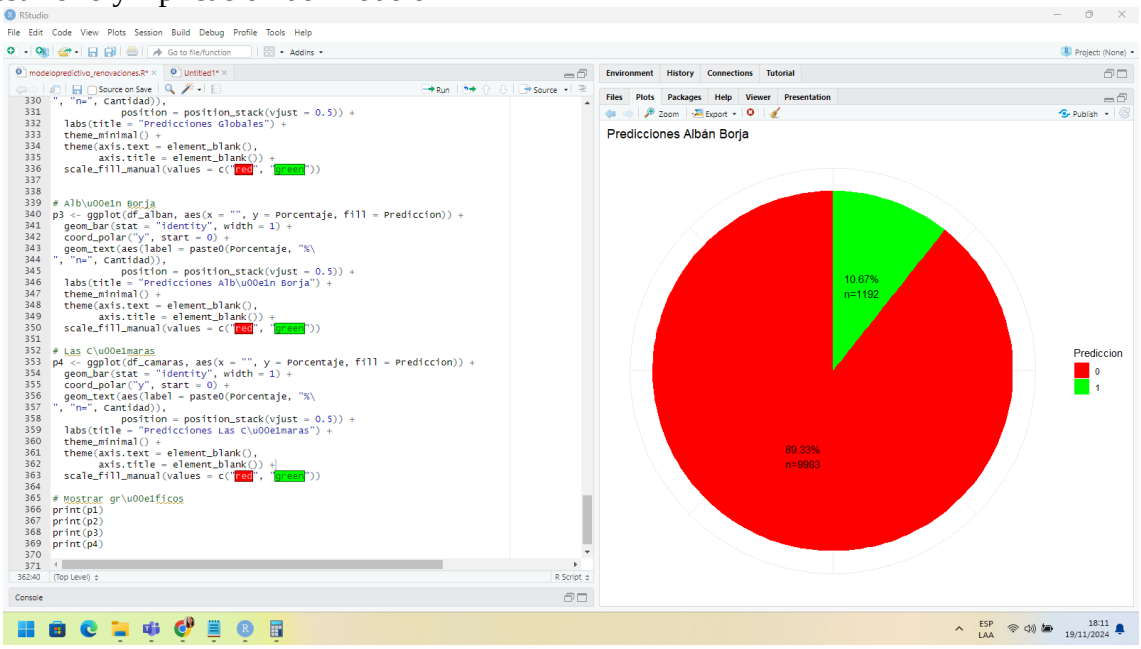
fuente: Elaboración propia

8.6. Procedimientos Importantes Realizados en el caso de estudio

8.6.1. Desarrollo del modelo en RStudio código comentado

Imagen 43

Desarrollo y Aplicación del modelo



fuente: Elaboración propia

```
# Cargar las librerías necesarias

installed.packages()

install.packages("caret")
```

```
install.packages("tidyr")
install.packages("viridis")
install.packages("smotefamily")
install.packages("reshape2")
install.packages("ROSE")
installed.packages()

library(randomForest)
library(caret)
library(pROC)
library(dplyr)
library(smotefamily)
library(ggplot2)
library(ROSE)
library(tidyr)
library(viridis)
library(reshape2)

# Cargar el archivo CSV con los datos históricos
datos_procesados <- read.csv("C:\\Users\\NANCY\\Documents\\datos historicos 2023-2024\\clientes_preparados_modelo.csv.csv")

# comprension de los datos
resumen <- summary(datos_procesados)
resumen

# Verificar si hay valores faltantes en los datos
print("Verificando valores faltantes en los datos:")
print(colSums(is.na(datos_procesados)))

# Eliminar las filas con valores faltantes
datos_procesados <- na.omit(datos_procesados)
```

```
# Ver las primeras filas de los datos cargados
head(datos_procesados)

# Asegurarse de que la variable objetivo sea factor
datos_procesados$TipoServi_Renovacion <- as.factor(datos_procesados$TipoServi_Renovacion)

# Verificar y codificar variables dicotómicas
# Asegurarse de que las variables dicotómicas estén en formato binario (0, 1)

# Revisar las variables dicotómicas
variables_dicotomicas <- c("TipoServi_Renovacion", "TipoServi_nuvfirma",
                           "TipoServi_otros", "estado_ambiente",
                           "zona_comercial", "zona_empresarial")

# Verificar que todas las variables dicotómicas sean 0 o 1
for (var in variables_dicotomicas) {
  unique_values <- unique(datos_procesados[[var]])
  print(paste("Valores únicos para", var, ":", paste(unique_values, collapse = ", ")))
}

# Asegurar que las variables estén en formato factor para modelado
for (var in variables_dicotomicas) {
  datos_procesados[[var]] <- as.factor(datos_procesados[[var]])
}

# Confirmar la conversión
str(datos_procesados[variables_dicotomicas])

# Calcular las frecuencias absolutas de la variable de respuesta
# Asumiendo que la variable de respuesta es 'TipoServi_Renovacion'
frecuencias_respuesta <- table(datos_procesados$TipoServi_Renovacion)
# Mostrar las frecuencias absolutas
```

```
print("Frecuencias absolutas de la variable de respuesta:")
print(frecuencias_respuesta)

# Preparar los datos
# Asegurarse de que la variable respuesta sea factor
datos_procesados$TipoServi_Renovacion <- as.factor(datos_procesados$TipoServi_Renovacion)

# Seleccionar variables para el balanceo
variables_modelo <- c("TipoServi_Renovacion", "dias_desde_compra", "Dias_servicio",
                      "TipoServi_nuvfirma", "TipoServi_otros", "estado_ambiente",
                      "zona_comercial", "zona_empresarial")

datos_balance <- datos_procesados[, variables_modelo]

# Aplicar ROSE para balancear las clases
datos_balanceados <- ROSE(TipoServi_Renovacion ~ ., data = datos_balance, seed = 123)$data

# Mostrar las frecuencias antes y despu\u00e9s del balanceo
print("Frecuencias antes del balanceo:")
print(table(datos_balance$TipoServi_Renovacion))

print("\
Frecuencias despu\u00e9s del balanceo:")
print(table(datos_balanceados$TipoServi_Renovacion))

# Visualizar la distribuci\u00f3n antes y despu\u00e9s del balanceo
par(mfrow=c(1,2))

# Gr\u00e1fico antes del balanceo
barplot(table(datos_balance$TipoServi_Renovacion),
        main="Distribuci\u00f3n Original",
        col=c("lightblue", "lightgreen"),
```

```
    names.arg=c("No Renovaci\u00f3n", "Renovaci\u00f3n"),
    ylim=c(0, max(table(datos_balance$TipoServi_Renovacion))))

# Gr\u00e1fico despu\u00e9s del balanceo
barplot(table(datos_balanceados$TipoServi_Renovacion),
    main="Distribuci\u00f3n Balanceada",
    col=c("lightblue", "lightgreen"),
    names.arg=c("No Renovaci\u00f3n", "Renovaci\u00f3n"),
    ylim=c(0, max(table(datos_balance$TipoServi_Renovacion))))

# Verificar la distribuci\u00f3n de algunas variables num\u00e9ricas antes y despu\u00e9s del balanceo
print("\
Resumen estad\u00edstico de d\u00edas_desde_compra antes del balanceo:")
print(summary(datos_balance$dias_desde_compra))

print("\
Resumen estad\u00edstico de d\u00edas_desde_compra despu\u00e9s del balanceo:")
print(summary(datos_balanceados$dias_desde_compra))

# Filtrar valores negativos en 'días_desde_compra' después del balanceo
datos_balanceados <- datos_balanceados[datos_balanceados$dias_desde_compra >= 0, ]

# Recalcular el resumen estadístico después de eliminar valores negativos
print("\
#Resumen estadístico de días_desde_compra después de eliminar valores negativos:")
#print(summary(datos_balanceados$dias_desde_compra))

# Verificar las frecuencias de la variable de respuesta después de la limpieza
print("\
Frecuencias después de eliminar valores negativos:")
print(table(datos_balanceados$TipoServi_Renovacion))
```

```
## verificando

# Análisis de variables cualitativas
# Seleccionar variables cualitativas
variables_cualitativas <- c("TipoServi_nuvfirma", "TipoServi_otros", "estado_ambiente",
                            "zona_comercial", "zona_empresarial")

# Calcular las frecuencias absolutas para cada variable cualitativa
frecuencias_cualitativas <- lapply(variables_cualitativas, function(var) {
  table(datos_procesados[[var]])
})

# Mostrar las frecuencias absolutas
names(frecuencias_cualitativas) <- variables_cualitativas
frecuencias_cualitativas

# Visualizar las frecuencias absolutas para cada variable cualitativa
par(mfrow=c(3,2))
for (var in variables_cualitativas) {
  barplot(table(datos_procesados[[var]]),
          main=paste("Frecuencias de", var),
          col=c("lightblue", "lightgreen"),
          names.arg=c("0", "1"),
          ylim=c(0, max(table(datos_procesados[[var]]))))
}



# Análisis de variables cuantitativas
# Seleccionar variables cuantitativas
variables_cuantitativas <- c("dias_desde_compra", "Dias_servicio", "año_caducidad",
                             "mes_caducidad", "dia_caducidad", "diasemana_compra")
```

```
# Calcular resumen estadístico para cada variable cuantitativa
resumen_cuantitativas <- lapply(variables_cuantitativas, function(var) {
  summary(datos_procesados[[var]])
})

# Mostrar el resumen estadístico
names(resumen_cuantitativas) <- variables_cuantitativas
resumen_cuantitativas

# Visualizar las distribuciones de las variables cuantitativas
par(mfrow=c(3,2))
for (var in variables_cuantitativas) {
  hist(datos_procesados[[var]],
       main=paste("Distribución de", var),
       xlab=var,
       col="lightblue",
       breaks=30)
}

# Calcular estadísticas resumidas para la variable 'dias_desde_compra'
resumen_dias_desde_compra <- summary(datos_procesados$dias_desde_compra)

# Mostrar el resumen estadístico
resumen_dias_desde_compra

#predictivo con Random forest

# Asegurarse de que la variable objetivo sea factor
datos_procesados$TipoServi_Renovacion <- as.factor(datos_procesados$TipoServi_Renovacion)

# Dividir los datos en conjuntos de entrenamiento y prueba
set.seed(123)
```

```
trainIndex <- createDataPartition(datos_procesados$TipoServi_Renovacion, p = 0.7,
                                  list = FALSE,
                                  times = 1)
datos_train <- datos_procesados[trainIndex,]
datos_test <- datos_procesados[-trainIndex,]

# Verificar las dimensiones de los conjuntos de entrenamiento y prueba
dim(datos_train)
dim(datos_test)

# Crear y entrenar el modelo Random Forest
set.seed(123)
modelo_rf <- randomForest(TipoServi_Renovacion ~ .,
                          data = datos_train,
                          ntree = 500,
                          importance = TRUE)
# Realizar predicciones en el conjunto de prueba
predicciones <- predict(modelo_rf, datos_test)
prob_predicciones <- predict(modelo_rf, datos_test, type = "prob")

# Calcular matriz de confusi\u00f3n
conf_matrix <- confusionMatrix(predicciones, datos_test$TipoServi_Renovacion)

# Imprimir resultados
print("Matriz de Confusi\u00f3n y M\u00e9tricas:")
print(conf_matrix)

# Precisión, Recall, F1-score
precision <- posPredValue(predicciones, datos_test$TipoServi_Renovacion)
recall <- sensitivity(predicciones, datos_test$TipoServi_Renovacion)
f1_score <- (2 * precision * recall) / (precision + recall)
```

```
cat("\nPrecisión:", precision, "\n")
cat("Recall:", recall, "\n")
cat("F1-score:", f1_score, "\n")

# Calcular curva ROC y AUC
roc_obj <- roc(datos_test$TipoServi_Renovacion, prob_predicciones[,2])

# Visualizar la curva ROC
plot(roc_obj, main="Curva ROC", col="blue")
abline(a=0, b=1, lty=2, col="red")

print("\
\u00c1rea bajo la curva ROC (AUC):")
print(auc(roc_obj))

# Visualizar la importancia de las variables
varImpPlot(modelo_rf, main="Importancia de Variables")
### predecir renovacion y no renovacion
# Realizar predicciones globales con el modelo entrenado
predicciones_globales <- predict(modelo_rf, datos_procesados)
# Crear dataframe con predicciones y zonas
resultados_df <- data.frame(
  prediccion = predicciones_globales,
  zona_comercial = datos_procesados$zona_comercial,
  zona_empresarial = datos_procesados$zona_empresarial
)
# Predicciones Globales
pred_globales <- table(predicciones_globales)
porc_globales <- prop.table(pred_globales) * 100
# Predicciones para Alb\u00e1n Borja (Zona Comercial 1)
alban_borja <- subset(resultados_df, zona_comercial == "1")
pred_alban <- table(alban_borja$prediccion)
```

```
porc_alban <- prop.table(pred_alban) * 100
#  Predicciones para Las C\u00e1maras (Zona Empresarial 1)
las_camaras <- subset(resultados_df, zona_empresarial == "1")
pred_camaras <- table(las_camaras$prediccion)
porc_camaras <- prop.table(pred_camaras) * 100
# Crear dataframes para visualizaci\u00f3n
df_global <- data.frame(
  Zona = "Global",
  Prediccion = names(pred_globales),
  Cantidad = as.numeric(pred_globales),
  Porcentaje = round(as.numeric(porc_globales), 2)
)

df_alban <- data.frame(
  Zona = "Alb\u00e1n Borja",
  Prediccion = names(pred_alban),
  Cantidad = as.numeric(pred_alban),
  Porcentaje = round(as.numeric(porc_alban), 2)
)
df_camaras <- data.frame(
  Zona = "Las C\u00e1maras",
  Prediccion = names(pred_camaras),
  Cantidad = as.numeric(pred_camaras),
  Porcentaje = round(as.numeric(porc_camaras), 2)
)

# Combinar todos los resultados
df_combined <- rbind(df_global, df_alban, df_camaras)

# Imprimir resultados
print("Predicciones Globales:")
print(df_global)
```

```
print("Predicciones Alb\u00e1n Borja:")
print(df_alban)
print("Predicciones Las C\u00e1maras:")
print(df_camaras)

# 2. Gr\u00e1ficos de pastel para cada zona
# Global
p2 <- ggplot(df_global, aes(x = "", y = Porcentaje, fill = Prediccion)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(Porcentaje, "%\
", "n=", Cantidad)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Predicciones Globales") +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank()) +
  scale_fill_manual(values = c("red", "green"))

# Alb\u00e1n Borja
p3 <- ggplot(df_alban, aes(x = "", y = Porcentaje, fill = Prediccion)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(Porcentaje, "%\
", "n=", Cantidad)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Predicciones Alb\u00e1n Borja") +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank()) +
  scale_fill_manual(values = c("red", "green"))
```

```
# Las C\u00e1maras
p4 <- ggplot(df_camaras, aes(x = "", y = Porcentaje, fill = Prediccion)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(Porcentaje, "%\
", "n=", Cantidad)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Predicciones Las C\u00e1maras") +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank()) +
  scale_fill_manual(values = c("red", "green"))

pred_summary <- as.data.frame(table(predicciones))
colnames(pred_summary) <- c("Categoria", "Cantidad")




# Crear el gráfico
p <- ggplot(pred_summary, aes(x = Categoria, y = Cantidad, fill = Categoria)) +
  geom_bar(stat = "identity") +
  labs(title = "Proyección de Renovaciones para 2025",
       x = "Categoría (0 = No Renovación, 1 = Renovación)",
       y = "Cantidad de Clientes") +
  theme_minimal() +
  scale_fill_manual(values = c("#FF9999", "#99CCFF"))

# Mostrar gr\u00e1ficos
print(p2)
print(p3)
print(p4)
```

print(p)

Mostrar las cantidades y porcentajes

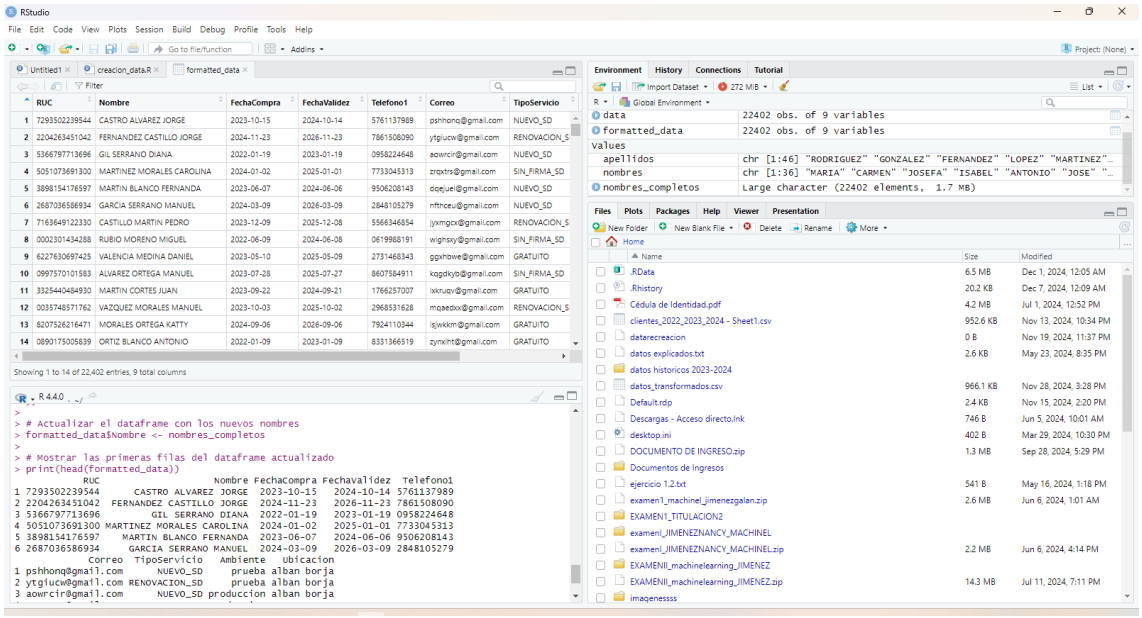
pred_summary$Porcentaje <- pred_summary$Cantidad / sum(pred_summary$Cantidad) * 100

print(pred_summary)

8.6.2. Codigo Creación de Data documentado el código de los pasos realizados

Imagen 44

Creación de Data



fuente: Elaboración propia

Cargar librerías necesarias

library(dplyr)

library(stringi)

Crear listas de nombres y apellidos reales

nombres <- c("MARIA", "CARMEN", "JOSEFA", "ISABEL", "ANTONIO", "JOSE", "MANUEL", "FRANCISCO",

"DAVID", "JUAN", "JAVIER", "DANIEL", "CARLOS", "JORGE", "LUIS", "ALBERTO",

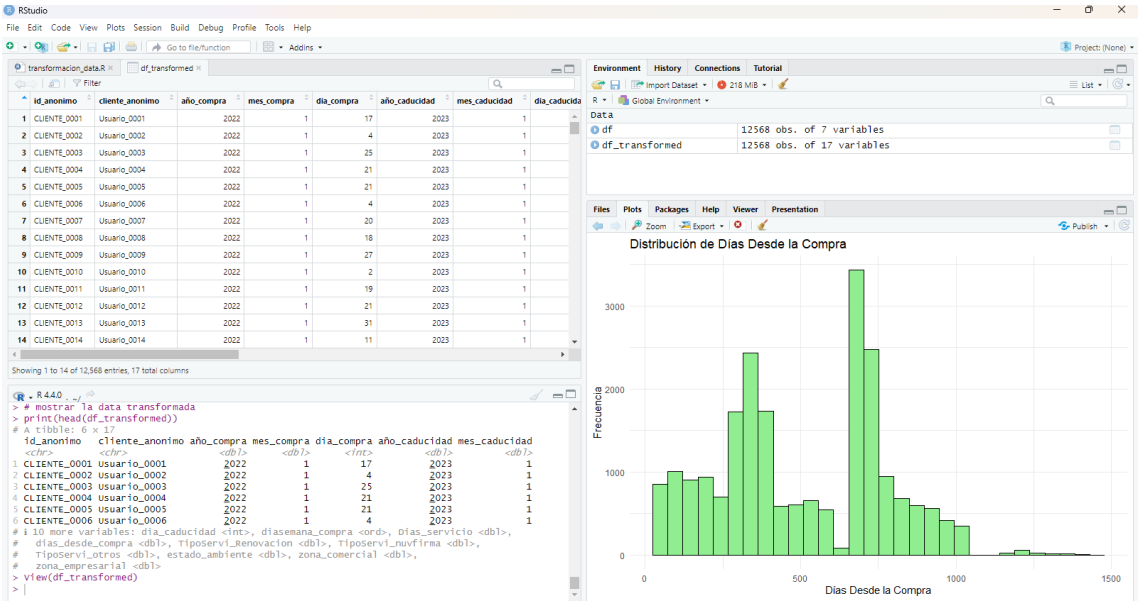
"MIGUEL", "RAFAEL", "PEDRO", "ANGEL", "ALEJANDRO", "FERNANDO", "PABLO",

```
          "MARINA", "KATTY", "ALEXANDRA", "DIANA", "CAROLINA", "ANDREA",
"PATRICIA",
          "MONICA", "KARLA", "ELIZABETH", "GABRIELA", "FERNANDA",
"VICTORIA")
apellidos <- c("RODRIGUEZ", "GONZALEZ", "FERNANDEZ", "LOPEZ", "MARTINEZ",
"SANCHEZ",
           "PEREZ", "GARCIA", "MARTIN", "JIMENEZ", "RUIZ", "HERNANDEZ",
"DIAZ",
           "MORENO", "ALVAREZ", "ROMERO", "ALONSO", "GUTIERREZ",
"NAVARRO",
           "TORRES", "DOMINGUEZ", "RAMOS", "VAZQUEZ", "RAMIREZ", "GIL",
"SERRANO",
           "MORALES", "MOLINA", "BLANCO", "SUAREZ", "CASTRO", "ORTEGA",
"DELGADO",
           "ORTIZ", "MARIN", "RUBIO", "SANZ", "NUÑEZ", "MEDINA", "IGLESIAS",
           "CORTES", "CASTILLO", "VALENCIA", "ORDONEZ", "GUERRERO",
"CAMPOS")
# Generar 22402 registros
data <- data.frame(
  RUC = stri_rand_strings(22402, 13, pattern = "[0-9]"),
  Nombre = replicate(22402, paste(stri_rand_strings(1, 7, pattern = "[A-Z]"),
stri_rand_strings(1, 8, pattern = "[A-Z]"), stri_rand_strings(1, 10, pattern = "[A-Z]"))),
  FechaCompra = sample(seq(as.Date('2022-01-01'), as.Date('2024-12-31'), by="day"),
22402, replace = TRUE),
  Telefono1 = stri_rand_strings(22402, 10, pattern = "[0-9]"),
  Correo = replicate(22402, paste0(stri_rand_strings(1, 7, pattern = "[a-z]"), "@gmail.com")),
  TipoServicio = sample(c("NUEVO_SD", "SIN_FIRMA_SD", "RENOVACION_SD",
"GRATUITO"), 22402, replace = TRUE),
  Ambiente = sample(c("produccion", "prueba"), 22402, replace = TRUE),
  Ubicacion = sample(c("alban borja", "camaras"), 22402, replace = TRUE)
)
# Generar FechaValidez con duración de 1 o 2 años desde FechaCompra
data <- data %>%
  mutate(FechaValidez = ifelse(runif(22402) > 0.5, FechaCompra + 365, FechaCompra +
730))
```

```
# Convertir las fechas a formato "YYYY-MM-DD" y reorganizar las columnas
formatted_data <- data %>%
  mutate(
    FechaCompra = as.character(as.Date(FechaCompra, origin = "1970-01-01")),
    FechaValidez = as.character(as.Date(FechaValidez, origin = "1970-01-01"))
  ) %>%
  select(RUC, Nombre, FechaCompra, FechaValidez, Telefono1, Correo, TipoServicio, Ambiente, Ubicacion)
# Generar nombres completos aleatorios sin duplicación
set.seed(123)
nombres_completos <- replicate(22402, paste(
  sample(apellidos, 1),  # Primer apellido
  sample(apellidos, 1),  # Segundo apellido
  sample(nombres, 1),    # Nombre
  collapse = " "
))
# Actualizar el dataframe con los nuevos nombres
formatted_data$Nombre <- nombres_completos
# Mostrar las primeras filas del dataframe actualizado
print(head(formatted_data))
resumen<-summary(formatted_data)
resumen
```

8.6.3. Código transformación de la data con código comentado

Imagen 45
Transformación de la Data



fuente: Elaboración propia

```
library(dplyr)
library(lubridate)
library(readxl)


# file.choose()


# leer el archivo de excel
df <- read_excel("C:\\Users\\NANCY\\Downloads\\Cliente_Caducados_2022_2023_2024 (1).xlsx")


# Crear identificadores anonimizados
set.seed(123)  # For reproducibility
df_transformed <- df %>%
  mutate(
    # anonimizacion
    id_anonimo = paste0("CLIENTE_", sprintf("%04d", row_number())),
    cliente_anonimo = paste0("Usuario_", sprintf("%04d", row_number())),
```

```
  # componentes de fecha
  año_compra = year(as.Date(FechaCompra)),
  mes_compra = month(as.Date(FechaCompra)),
  dia_compra = day(as.Date(FechaCompra)),
  año_caducidad = year(as.Date(FechaValidez)),
  mes_caducidad = month(as.Date(FechaValidez)),
  dia_caducidad = day(as.Date(FechaValidez)),

  # dias de la semana
  diasemana_compra = wday(as.Date(FechaCompra), label = TRUE),

  # dias del servicio
  Dias_servicio = as.numeric(as.Date(FechaValidez) - as.Date(FechaCompra)),

  # dias desde la compra
  dias_desde_compra = as.numeric(Sys.Date() - as.Date(FechaCompra)),

  # indicadores del tipo de servicio
  TipoServi_Renovacion = ifelse(TipoServicio == "RENOVACION_SD", 1, 0),
  TipoServi_nuvfirma = ifelse(TipoServicio == "NUEVO_SD", 1, 0),
  TipoServi_otros = ifelse(TipoServicio != "NUEVO_SD" & TipoServicio !=
"RENOVACION_SD", 1, 0),

  # estado del entorno
  estado_ambiente = ifelse(Ambiente == "PRODUCCION", 1, 0),

  # indicadores de la zona
  zona_comercial = ifelse(zona == "alban borja", 1, 0),
  zona_empresarial = ifelse(zona == "camaras", 1, 0)
 ) %>%
 select(
```

```
  id_anonimo , cliente_anonimo, año_compra, mes_compra, dia_compra,
  año_caducidad, mes_caducidad, dia_caducidad, diasemana_compra,
  Dias_servicio, dias_desde_compra, TipoServi_Renovacion,
  TipoServi_nuvfirma, TipoServi_otros, estado_ambiente,
  zona_comercial, zona_empresarial
 )

# mostrar la data transformada
print(head(df_transformed))

# exportar el archivo
write.csv(df_transformed, "C:/Users/NANCY/Documents/clientes_preparados_modelo.csv",
row.names = FALSE)
```