PREDICCIÓN DE DESERCIÓN DE CLIENTES EN PLANES DE PREVISIÓN EXEQUIAL UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

CARLOS FELIPE CORTÉS CATAÑO

CARLOS LUIS MORA CAÑAS

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

DANIEL ENRIQUE GONZÁLEZ GÓME
Director

DIEGO LUIS LIÑARES OSPINA
Jurado

JULIAN GIL GONZALES
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Se firma en la Ciudad de Santiago de Cali a los quince días de mes de febrero del dos mil veinticuatro (15, 02, 2024).

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 15 de febrero de 2024

Autores: Carlos Felipe Cortés Cataño y Carlos Luis Mora Cañas

Título del Trabajo de Grado: "Predicción de deserción de clientes en planes de previsión exequial utilizando técnicas de aprendizaje automático"

Director: Daniel Enrique González Gómez

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que los estudiantes indicados arriba han implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Daniel Enrique González Gómez
Director del proyecto

Santiago de Cali, 15 de febrero de 2024

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado "Predicción de deserción de clientes en planes de previsión exequial utilizando técnicas de aprendizaje automático, el cual será realizado por el estudiante Carlos Felipe Cortés Cataño con código 8972739 y por el estudiante Carlos Luis Mora Cañas con código 8972788,bajo la dirección del profesor Daniel Enrique González Gómez.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Firma
Carlos Felipe Cortés Cataño

C.C. 1053849221 de Manizales

Firma
Carlos Luis Mora Caña

C.C. 1090302363 de Cúcuta

Firma
Daniel Enrique González Gómez

C.C. 16669372 de Cali

La presente Política de Privacidad es aplicable tanto a PROMOTORA LA AURORA S.A.S., en calidad de responsable del Tratamiento y a sus empleados directos e indirectos, como a todas aquellas terceras personas naturales o jurídicas a quienes Transmita o Transfiera Datos Personales de los Titulares que comprenden los Grupos de Interés del responsable del Tratamiento, cuando estos realicen algún Tratamiento sobre los mismos.

Las disposiciones de la presente Política de Privacidad son de obligatorio cumplimiento por parte de todos los empleados de Promotora La Aurora S.A.S., contratistas y terceros que obren en su nombre, quienes las observarán y respetarán en cumplimiento de sus funciones. En los casos en que no exista vínculo laboral, se entiende incluida esta obligación para que los que tengan acceso a los datos, se obliguen a cumplirlas. Su incumplimiento originará sanciones de tipo laboral o de responsabilidad contractual o extracontractual según el caso, lo anterior sin perjuicio del deber de responder patrimonialmente por los daños y perjuicios que cause a los titulares de los datos o a Promotora La Aurora S.A.S. por su incumplimiento o el indebido tratamiento de datos personales.

Promotora la Aurora autoriza a Carlos Felipe Cortes y Carlos Mora Cañas, para hace uso adecuado de los datos que suministren directamente desde la Aurora, para fines educativos.

AUTORIZACIÓN PARA TRATAMIENTO DE DATOS PERSONALES

Por medio de la presente, autorizo(amos) de manera previa, expresa e inequívoca a PROMOTORA LA AURORA S.A.S., en calidad de Responsable del Tratamiento de Datos Personales para que directamente o a través de un tercero recolecte, almacene, circule, utilice, publique y en general para que trate mis (nuestros) Datos Personales, para las finalidades generales de todos los Grupos de Interés y las específicas para Clientes y sus Colaboradores, contenidas en la Política de Privacidad y Protección de Datos Personales, la cual declaro(amos) conocer y entender, y que forma parte integral de la presente autorización y está siempre a disposición para su consulta en la página web www.funeraleslaaurora.com, y específicamente para el cumplimiento de normas legales y/o contractuales y las siguientes: i) Efectuar las gestiones pertinentes para el cumplimiento del objeto del presente contrato y cumplir con las obligaciones emanadas del mismo; ii) Transmitir y transferir a terceros nacionales o internacionales con los cuales tenga relación contractual los Datos Personales suministrados; iii) Envío de comunicaciones relacionadas con las finalidades contenidas en la Política de Privacidad, el objeto social de PROMOTORA LA AURORA S.A.S., o aliados estratégicos, publicidad, marketing, promociones, eventos, comercialización y promoción de productos y/o servicios, actualizaciones de contenido en el sitio web, alianzas y beneficios, campañas de actualización de datos, a través de los datos de contacto profesionales, empresariales y/o personales de los Titulares; iv) Recolección, tratamiento, procesamiento, operación, verificación, consultas, reportes, solicitudes y otras actuaciones relacionadas con la información positiva o negativa del comportamiento crediticio, financiero, comercial y de servicios, así como a obtener de cualquier fuente y/o reportar a la Central de información contratada y/o a cualquier entidad nacional o internacional y/o central de riesgos que maneje o administre bancos o bases de datos, toda la información referente a mi (nuestro) comportamiento frente al cumplimiento de obligaciones y relaciones que adquiera o adquiridas con dichos sectores, independientemente de la naturaleza del contrato que los origine. Así mismo, declaro (amos) que soy (somos) mayor (es) de edad y Titular de los datos suministrados y que los mismos son exactos, veraces y

completos. Manifiesto (amos) que fueron informados los derechos de conocer, actualizar, rectificar y solicitar que se supriman datos personales en los casos que proceda conforme a las normas vigentes, o de revocar la autorización para alguna(s) de las finalidades contenidas en la Política de Privacidad y Protección de Datos Personales, salvo en los casos que tenga un deber legal o contractual de permanecer en las bases de datos. Declaro (amos) que fue informada la facultad para autorizar el tratamiento de datos sensibles, entendidos estos como aquellos que afectan la intimidad del Titular o cuyo uso indebido pueda generar su discriminación, tales como los datos biométricos, datos de salud, datos personales de Niños, Niñas y/o Adolescentes, entre otros; y conociendo tales derechos, autorizo (amos) el tratamiento de datos sensibles propios, y de los beneficiarios de los servicios con el fin de que sean prestados los servicios por parte del Titular y sea ejecutado el contrato en los términos en los que se suscribe. Finalmente, declaro (amos) conocer que, en caso de requerir información adicional, podrá contactarse el Responsable del Tratamiento a través del correo electrónico autorizaciondatos@laaurora.co o directamente en las instalaciones ubicadas en Manizales en la Calle 50 No. 24 - 34.

Para efectos de cumplimiento entre ambas partes a la política de privacidad, se firma el documento a los seis (_6_) días del mes de diciembre de 2023 en la ciudad de Manizales

Juanita Mejía Ramírez.

Project Manager
La Aurora Funerales y Capillas

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: “Predicción de deserción de clientes en planes de previsión exequial utilizando técnicas de aprendizaje automático”

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Ciencia de Datos
4. ESTUDIANTE (S): Carlos Felipe Cortés Cataño - Carlos Luis Mora Cañas
5. CORREO ELECTRÓNICO: cfcortesc@javerianacali.edu.co - mora2406@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Transversal 11 C 60 B 14 en Manizales, 3046568997 - Carrera 71 D 07 en Bogotá, 3012517377
7. DIRECTOR: Daniel Enrique González Gómez
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: dgonzalez@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica): No aplica
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Promotora La Aurora S.A.S.
12. OTROS GRUPOS O EMPRESAS: No aplica
13. PALABRAS CLAVE (al menos 5): Deserción de clientes, Estrategias de retención, Planes de previsión exequial, Aprendizaje automático, Modelado predictivo.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Objetivos de desarrollo sostenible
15. FECHA DE INICIO (Desarrollo del proyecto): 3/11/2022
16. RESUMEN (máximo 400 palabras). El presente proyecto evaluó varias técnicas de clasificación para identificar los clientes propensos a presentar deserción en contratos de previsión exequial en una compañía funeraria para después de comparar varias técnicas, seleccionar la técnica de aprendizaje automático “XGBoost”. La retención de clientes es esencial para la competitividad, cobertura y rentabilidad de esta empresa, y mediante la aplicación de este modelo, se logra un “recall” equivalente al 89%, permitiendo la identificación de 578 contratos propensos a desertar. Esto proporciona a la funeraria una buena alternativa para implementar estrategias más precisas y dirigidas a retener sus clientes, contribuyendo así a sus objetivos de crecimiento y éxito a largo plazo.

PREDICCIÓN DE DESERCIÓN DE CLIENTES EN PLANES DE PREVISIÓN EXEQUIAL UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Estudiantes
Carlos Felipe Cortés Cataño
Código 897278
Carlos Luis Mora Cañas
Código 8972739

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Daniel Enrique González Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, FEBRERO 15 DE 2024

CONTENIDO

TABLA DE IMÁGENES

INTRODUCCIÓN

La creciente evolución tecnológica ha generado mercados cada vez más volátiles, lo que ha llevado a las empresas a enfocar sus esfuerzos en la búsqueda continua de estrategias que añadan valor a sus clientes y estimulen el crecimiento del negocio. En este contexto, el modelado predictivo ha revelado un enorme potencial para tomar decisiones fundamentadas en la comprensión de preferencias y el análisis de comportamiento, factores que influyen en el uso o abandono de servicios.

La retención de clientes es un factor que impacta el resultado económico de las empresas, según un estudio de Harvard Business Review; un aumento del 5% en la retención de clientes pudo aumentar las utilidades de la empresa en un 25% a 95% [1]. Para establecer campañas de retención de clientes precisas y dirigidas, se requiere analizar grandes volúmenes de datos para identificar las variables que inciden positiva y negativamente en el uso de un servicio, detectarlos con anticipación y tomar medidas para retenerlos.

En este trabajo, se estudiaron las variables que inciden en la deserción de programas de previsión exequial, variables que se emplearon en técnicas de aprendizaje automático para desarrollar un modelo predictivo que proporcionara alerta temprana sobre clientes con mayor riesgo de deserción. De esta manera, una compañía podría orientar sus políticas y estrategias de retención de clientes fundamentadas en datos.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Retener clientes es un desafío para las empresas funerarias, debido a que un aumento del 5% en la retención de clientes puede aumentar sus utilidades entre un 25% a 95% [1] y que estima que un 70% de los ingresos de las empresas provienen de clientes existentes [2], ingresos que para el caso de contratos prepagados de previsión exequial se conservan entre 5 a 20 años antes de su utilización. [3]

Una empresa del sector funerario de Manizales considera que peligra la sostenibilidad de la compañía a largo plazo por la disminución en los recaudos provenientes de contratos de previsión exequial durante el último año. Este fenómeno puede ser originado a una combinación de factores que incluyen insatisfacción en el servicio al cliente, cambios en gustos, competencia con otras empresas funerarias o cambios en el entorno económico.

La empresa reconoce que es más costoso adquirir clientes nuevos que retenerlos, por tal razón, su objetivo es identificar los clientes propensos a desertar para encuestarlos, ofrecerles programas de fidelización de clientes, cambiar los servicios ofrecidos y así favorecer su ciclo de vida. Sin embargo, el gran volumen de datos que genera la compañía funeraria, como los registros de los programas, las transacciones, y los comentarios de los clientes dificultan su identificación.

Por otro lado, esta empresa no cuenta con un sistema robusto para realizar análisis y modelamiento de datos, por lo cual esta iniciativa les ayudara a optimizar procedimientos y costos operacionales.

1.2. FORMULACIÓN DEL PROBLEMA

Basado en la información disponible, ¿Cuál técnica de aprendizaje automático permitiría maximizar su capacidad predictiva para mitigar el fenómeno de la deserción en planes de previsión exequial?

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Identificar la mejor técnica de aprendizaje automático que permita perfilar los clientes propensos a presentar deserción en planes de previsión exequial.

2.2. OBJETIVOS ESPECÍFICOS

- Construir un conjunto de datos, recopilando información pertinente sobre clientes y sus contratos de previsión exequial.
- Determinar las líneas de negocio, las técnicas de inteligencia artificial a utilizar y las variables de mayor incidencia mediante el entendimiento del conjunto de datos.
- Seleccionar y preparar las variables identificadas para el entrenamiento y prueba de modelos de aprendizaje automático.
- Entrenar técnicas de aprendizaje automático utilizando el conjunto de entrenamiento y evaluando el rendimiento de los modelos mediante las diferentes métricas.
- Realizar una comparación detallada de diversas técnicas de aprendizaje automático para seleccionar el modelo más adecuado.
- Implementar una función que transforme los datos y use el modelo elegido en una máquina virtual proporcionada por la empresa.

3. MARCO TEORICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

Se presenta la definición de los planes de previsión exequial, su importancia para la empresa funeraria y como la deserción de clientes podría afectar la sostenibilidad de esta. Así mismo, se exponen las técnicas de aprendizaje automático más usadas para la identificación de clientes propensos a presentar deserción y sus métricas de evaluación que serán el insumo de comparación entre técnicas.

3.1.1 PLANES DE PREVISIÓN EXEQUIAL

La empresa funeraria ha logrado establecer un negocio rentable [4] en respuesta a la creciente demanda generada por la percepción moderna de la muerte. Ofrece servicios que manejan ceremonias mortuorias, traslados y manipulación de cuerpos, así como la obtención y preparación de ataúdes, cremaciones y trámites civiles. Estos servicios están diseñados para hacer frente a los imprevistos que surgen tras el fallecimiento de un individuo. Sin embargo, aunque proporcionan soluciones necesarias, también plantean un desafío económico para los familiares y allegados.

Este dilema ha propiciado el surgimiento de diversos instrumentos financieros, tales como contratos de previsión exequial, planes pre-exequiales, pre-necesidad o seguros de deceso. Estos planes se encargan de cubrir las responsabilidades asociadas al fallecimiento, estableciendo condiciones específicas en un contrato. En España, según los datos proporcionados por ICEA, este sector ha experimentado un crecimiento constante, llegando a representar el 0.32% del Producto Interno Bruto (PIB). En el año 2019, el 59.60% de las defunciones contaban con cobertura de un seguro de decesos, según el mismo informe. Además, se destacó que para el año 2020, el 45.64% de la población disponía de un seguro de decesos [5].

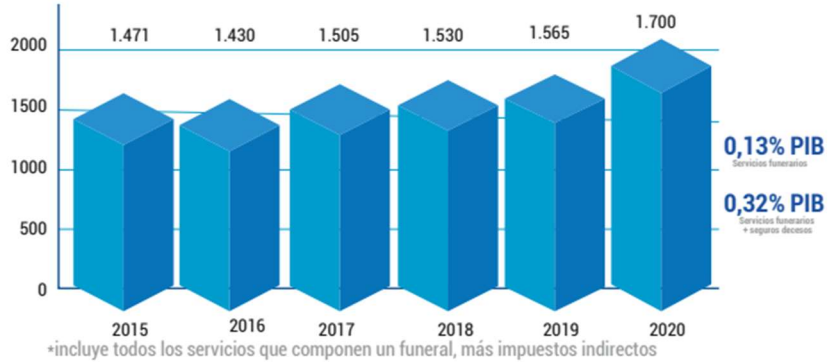


Ilustración 1. Volumen facturación servicios funerarios en España en Millones de euros [4]

La naturaleza intrínseca de estos contratos comparte elementos esenciales con las compañías aseguradoras, como los pagos recurrentes, el interés asegurable, el riesgo asegurable, la prima o precio del seguro y la obligación condicional con el asegurador. Se distinguen por ofrecerlas entidades cooperativas o sociedades comerciales, lo que implica para el tomador la obligación de realizar pagos oportunos hasta el momento su uso y para el asegurador, indemnizar a los beneficiarios en especie, nunca en efectivo [5]. Así, como estos contratos aseguran para la empresa funeraria un ingreso recurrente de 5 a 20 años, donde se ubica el promedio de utilización. [3]

3.1.2 DESERCIÓN DE CLIENTES

La deserción es una problemática que presentan empresas de diferentes sectores como son las telecomunicaciones, aseguradores o la banca [4] [5] [6]. Su definición se enmarca en la tendencia de los clientes a abandonar los productos o servicios de una marca. Sin embargo, en estos sectores se define como la cancelación o abandono de servicios por motivos voluntarios e involuntarios los cuales incluyen calidad de llamada inaceptable, plan de precios de la competencia más favorable, información errónea dada por ventas, expectativas del cliente no satisfechas, problema de facturación, mudanza, cambio en el negocio, etc. [7]. Para el empresario funerario, sus esfuerzos están orientados a conseguir mayores proporciones de mercado y dar reconocimiento a su marca, aspectos que se ve afectado por la cancelación de los servicios fúnebres, el incumplimiento de los contratos y la reducción de la tasa de mortalidad y natalidad. [3].

El porcentaje de desertores durante un periodo en particular se denomina tasa de abandono de clientes y su fórmula está conformada por la división del número de clientes perdidos durante el intervalo de tiempo por el número de clientes adquiridos y multiplicar es número por el 100 por ciento [8]. Cualquier esfuerzo que favorezca este indicador podría afectar positivamente los ingresos de las compañías en el corto y largo plazo. Según Feinberg, un incremento del 5% de retención de clientes, hace que la empresa aumente sus ingresos de 25% a un 85% [9]. Sin embargo, se deben identificar a los clientes propensos a desertar para aplicar estrategias para retenerlos [7]. Así mismo, estas industrias reconocen que es más costoso adquirir nuevos clientes que retener antiguos, [10] e identificarlos permite aplicar estrategias para retenerlos. Según la consultora Bain & Company, el costo de adquirir un nuevo cliente es entre 5 y 7 veces mayor que el costo de retener un cliente existente. Esto se debe a que los costos de adquisición incluyen marketing, ventas y servicio al cliente, mientras que los costos de retención se centran en el servicio al cliente. [11]

3.1.3 SELECCIÓN DE CARACTERISTICAS

La selección de características es un proceso importante en el aprendizaje automático, ya que permite reducir la dimensionalidad de los datos y mejorar el rendimiento del modelo. En este trabajo, se utilizarán dos métodos estadísticos para la selección de características: la prueba t-

Student y la prueba chi-cuadrado. Estos métodos resaltan para la selección de características porque son insensibles a la presencia de valores atípicos, eficientes para analizar gran cantidad de datos y proporcionan interpretación de la relación entre las características y la variable objetivo. [12], la prueba t-Student es un método estadístico paramétrico que se utiliza para comparar las medias de dos grupos. En el contexto de la selección de características, se puede utilizar para comparar la media de una característica entre dos grupos, uno con la etiqueta deseada y otro con la etiqueta no deseada. Si la diferencia entre las medias es significativa, entonces la característica es relevante para la clasificación [13]

La prueba chi-cuadrado es un método estadístico no paramétrico que se utiliza para comparar dos distribuciones de probabilidad. En el contexto de la selección de características, se puede utilizar para comparar la distribución de una característica entre dos grupos, uno con la etiqueta deseada y otro con la etiqueta no deseada. Si la diferencia entre las distribuciones es significativa, entonces la característica es relevante para la clasificación [13].

Adicionalmente, el uso del análisis de componentes principales (PCA) es una oportunidad para reducir la dimensionalidad de un conjunto de datos. A diferencia de los métodos no favorece la interpretación, sin embargo, es útil para trabajar con conjuntos de datos de gran tamaño, reduciendo las características, pero sin perder mucha información. El PCA encuentra un conjunto de nuevas variables, llamadas componentes principales, que son combinaciones lineales de las variables originales. Estas nuevas variables están ordenadas de modo que la primera componente principal captura la mayor parte de la varianza del conjunto de datos, la segunda componente principal captura la mayor parte de la varianza restante, y así sucesivamente [14].

3.1.4 APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es el subcampo de la inteligencia artificial que se ocupa del diseño y desarrollo de algoritmos que permiten a las computadoras simular aprendizaje mediante un conjunto de datos. Los algoritmos pueden identificar patrones que representan sucesos o eventos, como aprendizaje se refiere a los algoritmos.

Estos algoritmos se pueden clasificar en aprendizaje supervisado, semi supervisado y no supervisado. El aprendizaje supervisado se basa en la detección de patrones sobre los datos que permiten clasificar los datos con una etiqueta previamente definida. El algoritmo semi supervisado es relativamente similar al anterior pero también permite crear nuevas clasificaciones a datos sin etiquetar [15]. Finalmente, el no supervisado se encarga de agrupar los datos según categorías definidas propiamente por el algoritmo, este último es altamente utilizado para generar conocimiento adicional de los datos.

Dentro de los algoritmos supervisados encontramos la capacidad de clasificar de manera automática nuevos eventos a partir del conocimiento de eventos pasados. Para que esto sea posible se debe contar con información suficiente para identificar las características que permiten que este se ubique en una etiqueta u otra [16].

3.1.5 TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA CLASIFICACIÓN

Diversas compañías modelan los usuarios que van a desertar mediante aprendizaje automático, permitiendo identificar y cuantificar el impacto de los clientes que serán dados de baja. Los resultados permiten plantear escenarios y acciones comerciales de retención [17]. Existen diferentes algoritmos que han presentado soluciones para diferentes industrias en su identificación, pero se presentan a continuación aquellos óptimos en la predicción de deserciones en la industria bancaría, de telecomunicaciones y aseguradoras según la literatura, y la regresión logística, un modelo base que favorece la interpretación de los resultados.

3.1.5.1 REGRESIÓN LOGISTICA

La regresión logística es un modelo estadístico que se utiliza para predecir una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de una o más variables independientes o predictoras. La regresión logística se basa en la función de probabilidad logística, que es una función matemática que se utiliza para modelar la probabilidad de que una variable categórica tome un determinado valor [13]. La función de probabilidad logística se expresa como la siguiente formula:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-wz}}$$

Donde $P(y = 1| x)$ representa la probabilidad de que la variable categórica tome el valor 1, considerando el vector de características x. El vector de parámetros w se obtiene mediante la estimación a partir de los datos de entrenamiento. x representa el vector de características asociado a un dato de entrada.

3.1.5.2 ARBOLES DE DECISIÓN – RANDOM FOREST

Los árboles de decisión son modelos de aprendizaje supervisado que se utilizan tanto para tareas de clasificación como de regresión. Su estructura jerárquica y su fácil interpretación los convierten en una herramienta poderosa para comprender las relaciones entre variables y realizar predicciones precisas, así mismo, son fundamentales para la construcción de modelos más robustos dentro del aprendizaje automático donde se combinan múltiples modelos débiles para crear un modelo fuerte. [19]

En un árbol de decisión, cada nodo interno representa una pregunta sobre una característica del conjunto de datos, y las ramas que emanan de ese nodo representan las posibles respuestas a esa pregunta. Las hojas del árbol representan las clases o valores de destino que se predicen.

3.1.5.3 CLASIFICADOR DE AUMENTO DE GRADIENTE - GRADIENT BOOSTING CLASSIFIER

El aumento de gradiente fue propuesto por primera vez por Friedman [10] en 2001. La Ilustración 2 ejemplifica como se construye una serie de árboles donde cada otro árbol intenta corregir los errores de su árbol predecesor mediante el método de descenso del gradiente y es iterativo hasta que logra la estimación optima de la variable objetivo. Este algoritmo no elimina o modifica árboles predecesores, pero si un gran número de árboles son añadidos al tiempo. [20]

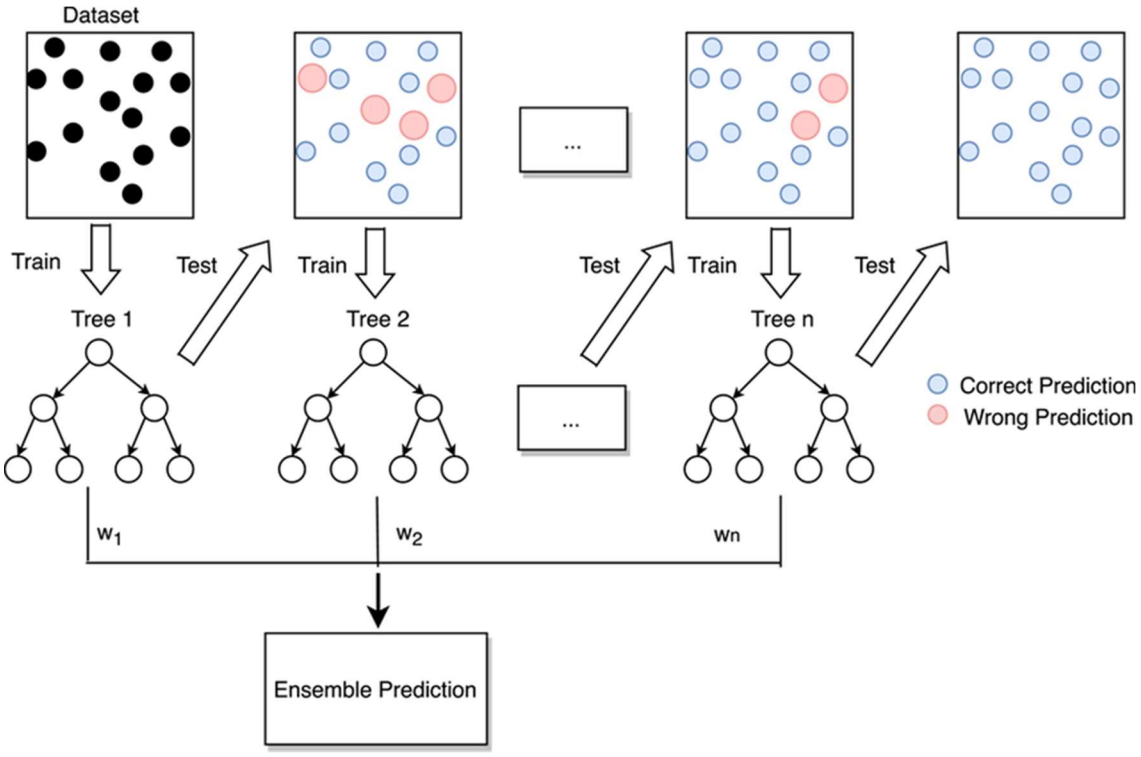


Ilustración 2. Gradient Boosting Classifier [19]

3.1.5.4 AUMENTO DEL GRADIENTE EXTREMO – XGBOOST

En 2016, Chen y Guestrin [19] modificaron el concepto de descenso del gradiente introducido por Friedman para crear árboles de decisión diseñados para aumentar la velocidad y el rendimiento del algoritmo al que llamaron “XGBoost”, usando el máximo poder computacional. A diferencia del aumento del gradiente utiliza regularización para evitar sobre ajuste, cada clasificador maneja conjunto de datos disperso en el que los valores faltantes se manejan correctamente y admite una

poda de árboles o eliminar ramas innecesarias para ser más eficiente.

XGBoost hace referencia al aumento del gradiente extremo, el cual es un método de aprendizaje supervisado que combina árboles para proporcionar un modelo de aprendizaje automático más generalizable. Debido a su buena optimización de caché, el clasificador XGboost produce buenos resultados en el modelo de predicción, pero requiere más tiempo de entrenamiento para el proceso de iteración. XGBoost es un algoritmo potente y versátil, [19]pero tiene una gran cantidad de hiperparámetros que pueden ser difíciles de ajustar, a continuación, se presentan los parámetros más importantes.

Learning rate: El learning rate es un parámetro que controla la cantidad de ajuste que se realiza en cada iteración. Un learning rate más alto conducirá a un ajuste más agresivo, mientras que un learning rate más bajo conducirá a un ajuste más conservador. El valor predeterminado es 0.1.

Subsample: El subsample es un parámetro que controla la proporción de datos que se utilizan para entrenar cada árbol. Un subsample más bajo conducirá a un modelo más robusto a los sesgos de los datos, mientras que un subsample más alto conducirá a un modelo más preciso. El valor predeterminado es 1.0.

Colsample_bytree: El colsample_bytree es un parámetro que controla la proporción de características que se utilizan para entrenar cada árbol. Un colsample_bytree más bajo conducirá a un modelo más robusto a los sesgos de los datos, mientras que un colsample_bytree más alto conducirá a un modelo más preciso. El valor predeterminado es 1.0.

N_estimators: El n_estimators es un parámetro que controla el número de árboles que se construyen. Un n_estimators más alto conducirá a un modelo más preciso, pero también puede conducir a un sobreajuste. El valor predeterminado es 100.

3.1.5.5 CLASIFICADOR DE IMPULSO ADAPTATIVO – ADABOOST CLASSIFIER

Yoav Freund y Robert Schapire [20] propusieron este clasificador en el año 1995. La Ilustración 3 muestra la combinación de clasificadores débiles que dan baja precisión para obtener un clasificador fuerte y preciso, para ello debe entrenarse iterativamente en varios ejemplos de entrenamiento pesados para ajustarlos, minimizando el error de entrenamiento. A diferencia de las dos técnicas vistas anteriormente este clasificador es sensible a ruido o datos no lineales y tiene menor velocidad de entrenamiento y predicción que el XGboost.

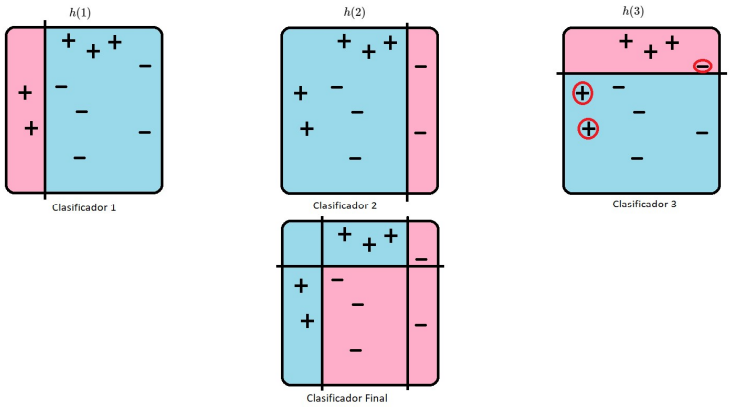


Ilustración 3. Adaboost Classifier [Autor, 2023]

3.1.5.6 PERCEPTRÓN MULTICAPA – MLP

El perceptrón multicapa (MLP) es un algoritmo de aprendizaje supervisado que aprende una función lineal o no lineal

El perceptrón multicapa (MLP) es un algoritmo de aprendizaje supervisado que aprende de acuerdo con la siguiente función no lineal:

$$f(x): \sigma(wx + b)$$

Donde σ es la función de activación, como lo es una función sigmoidea, RM (Rectified Linear Unit) o RO (Rectified Output). w es el vector de pesos que conecta las neuronas de la capa anterior con la neurona actual. b es el sesgo de la neurona. x es el vector de entradas a la neurona.

El MLP tiene varias capas de neuronas, cada una con un conjunto de pesos asociados. Las neuronas de una capa están conectadas a las neuronas de la siguiente capa, minimizando una función de costo. La función de costo mide la diferencia entre las predicciones del MLP y los valores reales del conjunto de datos de entrenamiento.

El MLP se utiliza en una amplia gama de aplicaciones, como clasificación, regresión y procesamiento de lenguaje natural. Se diferencia de la regresión logística porque entre la capa de entrada y la de salida, puede haber una o más capas no lineales, llamadas capas ocultas. En el caso del MLP, la función que relaciona las entradas con las salidas es una función no lineal. Esto significa que la función no puede expresarse como una combinación lineal de las entradas. La capa oculta del MLP permite que la función sea no lineal [21]. La Ilustración 1 muestra un MLP de una capa oculta con salida escalar.

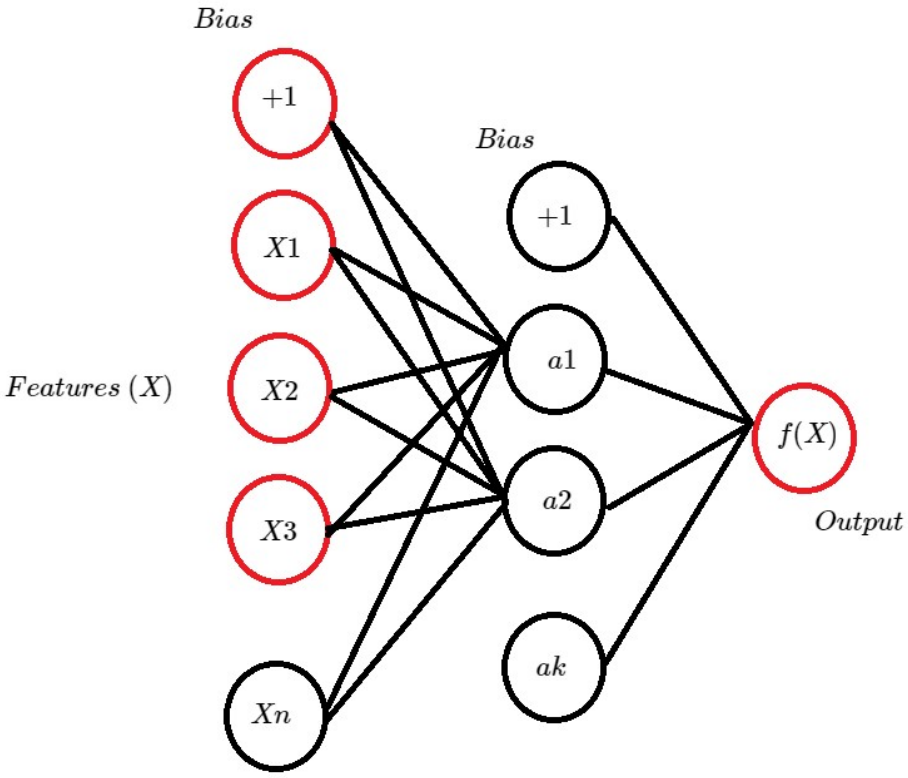


Ilustración 4.Perceptrón multicapa [Autor,2023]

La capa de entrada es la capa más a la izquierda de un perceptrón multicapa. Consta de neuronas, cada una representa una característica de entrada. Las neuronas de la capa oculta transforman los valores de la capa de entrada mediante una suma lineal ponderada, seguida de una función de activación no lineal, como la función tangente hiperbólica. La capa de salida recibe los valores de la última capa oculta y los transforma en valores de salida. [22]

3.1.6 EVALUACIÓN

Un modelo de predicción debe generalizar cualquier conjunto de datos nunca visto, para lo que dividir el conjunto de datos original en datos de entrenamiento y datos de prueba para que el conjunto. Sobre cada técnica se usará el conjunto de prueba para evaluar el desempeño del modelo determinado mediante la capacidad que tiene cada una de ellas para representar los datos y detectar la deserción.

Los resultados obtenidos se enmarcan en métricas de evaluación las cuales permiten cuantificar el desempeño del modelo y la confianza en la clasificación realizada por el mismo. Las métricas de

evaluación son herramientas que se utilizan para medir el rendimiento de un modelo de clasificación. Estas métricas ayudan a los científicos de datos a evaluar la precisión y la eficacia de sus modelos. [23]

3.1.6.1. EXACTITUD

La exactitud es la métrica de evaluación más simple. Mide la proporción de predicciones correctas del total de predicciones. La exactitud se calcula de la siguiente manera:

$$\boldsymbol{Exactitud} = \frac{\boldsymbol{TP+TB}}{(\boldsymbol{TP+FP+TN+FN})}$$

Donde $\boldsymbol{TP}$: Verdaderos positivos, $\boldsymbol{FP}$: Falsos positivos, $\boldsymbol{TN}$: Verdaderos negativo y $\boldsymbol{FN}$: Falsos negativos

3.1.6.2. PRECISIÓN

La precisión mide la proporción de abandonos correctamente detectados entre todos los predichos. La *precisión* se calcula de la siguiente manera:

$$\boldsymbol{Precisión} = \frac{TP}{(TP+FP)}$$

3.1.6.3. RECALL

El recall mide la proporción de abandonos correctos detectados contra el total de abandonos reales. El *recall* se calcula de la siguiente manera:

$$\boldsymbol{Recall} = \frac{TP}{(TP+FN)}$$

3.1.6.4. F1-SCORE

El F1-Score mide la proporción de abandonos correctos detectados contra el total de abandonos reales. Es definido como la media armónica entre precisión y recall. El F1-Score se calcula de la siguiente manera:

$$\boldsymbol{F1-Score} = \frac{2\ x\ \mathrm{Pr}ecsion\ x\ Recall}{(precisión+recall)}$$

3.1.6.5. CURVA ROC

La curva *ROC* mide la capacidad para discriminar casos positivos de casos negativos. Tiene un área bajo la curva (*AUC*) entre 1 y 0, una curva *ROC* con *AUC* igual 1 es un clasificador perfecto.

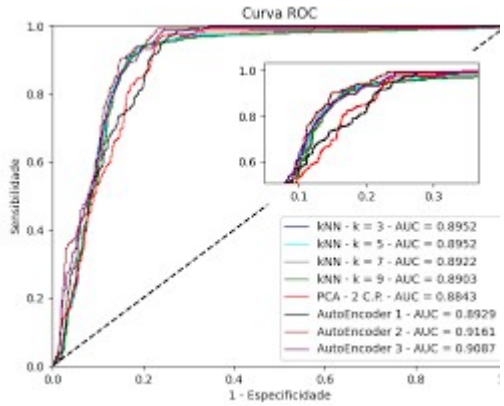


Ilustración 5. ROC curve evaluation of edge detector performance [25]

3.1.7 OPTIMIZACIÓN

La optimización del algoritmo es el proceso de encontrar los parámetros del algoritmo que dan el mejor rendimiento en un conjunto de datos de prueba. La búsqueda cuadrada y la validación cruzada son dos técnicas comunes que se utilizan para la optimización del algoritmo. La búsqueda cuadrada y la validación cruzada se pueden combinar para mejorar la eficiencia de la optimización del algoritmo [12]. La búsqueda cuadrada puede encontrar un rango de parámetros prometedores, y la validación cruzada puede usarse para evaluar los parámetros dentro de ese rango.

La búsqueda cuadrada es un método de optimización que utiliza un algoritmo de búsqueda para encontrar la combinación de parámetros que maximiza una función objetivo. En el contexto de la optimización del algoritmo, la función objetivo suele ser una medida del rendimiento del modelo en un conjunto de datos de prueba. La búsqueda cuadrada puede ser un método eficaz para encontrar los parámetros óptimos del algoritmo. Sin embargo, puede ser un método costoso, ya que requiere evaluar el modelo para cada combinación de parámetros [13].

La validación cruzada es un método de evaluación que se utiliza para estimar el rendimiento de un modelo en un conjunto de datos de prueba. La validación cruzada funciona dividiendo el conjunto de datos en k subconjuntos. El modelo se entrena en $k-1$ subconjuntos y se evalúa en el subconjunto restante. Este proceso se repite k veces, y el rendimiento del modelo se calcula como el promedio de los k resultados. La validación cruzada es un método eficaz para estimar el rendimiento del modelo en un conjunto de datos de prueba. Sin embargo, puede ser un método costoso, ya que requiere entrenar el modelo *k* veces [13].

3.2 ANTECEDENTES

Dentro de la revisión sistemática de la literatura se seleccionaron referentes que abordó la

problemática de la deserción de clientes utilizando técnicas de aprendizaje automático en sectores donde las reglas de negocio incluyen servicios prepagados o pagos recurrentes mensuales, como son los seguros, telecomunicaciones, bancos o suscripciones en ámbitos nacionales e internacionales.

Titulo - Autores	Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático – Jesús David Falla Arango	Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones – Andrés Felipe Echeverri	Modelo de predicción de deserción de clientes de tarjetas de crédito – Brenda Denisse Cobeña Terán
Técnicas usadas	Gradient Boosting Classifier, MLP Classifier, Random Forest Classifier, XGB Classifier, Logistic Regression, GaussianNB, KNeighbors Classifier y Linear Discriminant Analysis.	Decission Tree Classifier, Random Forest Classifier y XGboost.	Decissión Tree Clasifier y Adabag Adaboost.
Hitos	Los árboles de decisión superan a todos los modelos de regresión logística examinados. Es importante definir las fuentes y periodo de observación para tener imagen completa de la interacción del cliente. Cuanto más cualitativo sea el conjunto de datos, más precisos son los pronósticos. Cuanto más cualitativo sea el conjunto de datos, más precisos son los pronósticos. Las variables más influyentes fueron intención de cancelación 3 meses antes de la baja, e interacción voz de cliente negativa. XG Classifier fue uno de los mejores algoritmos con AUC de 0.78, Precisión de 0.74, Recall de 0.67 y F1-Score de 0.025, así como Gradient Boosting Classifier con un AUC de 0.78, Precisión de 0.73, Recall de 0.68 y F1 Score de 0.7061. [10]|	El modelo CRISP-DM es altamente recomendable. Se realizo imputación a los datos nulos. Definió alcance para el abandono al mes siguiente en Se selecciono XGBoost por su alta adaptabilidad, robustez y capacidad de trabajar con datos desbalanceados usando los hiperparametros Scale_pos_weight: 30, N_estimators:300 y Max-depth:2 y obteniendo las métricas: Recall de 0.66, Precisión de 0.05 y AUC de 0.95. [14]	Determina variables explicativas usando estimadores de Kaplan – Meier y Regresión de Cox pero obtuvieron errores del 45.19%. Se propone usar arboles de decisión y el algoritmo de Adaboost, los árboles obtuvieron error promedio de 6.30% y el Adaboost tuvo erro de 3.77%. [15] Las variables más importantes fueron antigüedad y monto promedio de consumo diferido. Se recomienda la implementación de reportes periódicos de monitoreo.

Tabla 1. Referentes bibliográficos nacionales [Autor-2023]

Titulo - Autores	Predicting Customer Churn Using Recurrent Neural Networks – Jesper Ljungehed	Customer churn prediction for an insurance company – Huigevoort, C.W.J.M.
Técnicas usadas	Combinación de K-means y Recurrent Neural Network.	Combinación de K-means y Recurrent Neural Network.
Hitos	El algoritmo K-means podría usarse para realizar un análisis más completo de los predictores ya que obtuvo grupo prometedores, lo que concluye que había patrones que debían distinguirse. Los resultados de la evaluación muestran que K-means es aplicable para investigar aún más la salida de la predicción de abandono sin la necesidad de un algoritmo de extracción de reglas, lo que resulta en un tiempo de ejecución más corto y una implementación más simple. [22] El análisis del algoritmo K-means realizado en esta tesis indicó que la RNN era capaz de identificar tendencias, pero no pudo predecir desviaciones de la tendencia. Vale la pena investigar el potencial de entrenar un predictor de RNN por grupo, en lugar de entrenar un modelo único para todos. Dada una solución de agrupamiento significativa de K clusters, podría valer la pena investigar el rendimiento de promediar la salida de K diferentes RNNs deformadas por separado en cada cluster.	Los datos se encuentran con desbalance para lo cual se propuso analizar proporciones de balaceo 80:20, 70:30, 60:333 y 50:50.Para redes neuronales el mejor resultado lo obtuvieron con balaceo 70:30, mientras para la regresión logística fue 50:50.Las redes neuronales tienen mejor desempeño para incrementar los beneficios, sin embargo, la regresión logística es importante para el departamento de marketing porque permite ver los pesos de cada variable, lo cual era más importante para este problema y se selecciona la regresión logística. [21] Las variables más significativas fueron número de quejas y de contactos. No se tuvo relación con la ubicación del cliente.

Tabla 2. Referentes bibliográficos internacionales [Autor-2023]

En los estudios revisados los algoritmos de aprendizaje automático efectivos para la predicción de deserción son XGBoost, árboles de decisión y regresión logística. La recomendación general es utilizar XGBoost para la predicción de churn. Sin embargo, es importante considerar las características específicas de la empresa y del cliente al elegir un algoritmo, por lo tanto, los estudios proponen algoritmos como son AdaBoost y XG Classifier.

Por otro lado, es importante considerar las siguientes variables al construir un modelo de predicción de deserción: Intento de cancelación, Interacción negativa con el servicio al cliente, Antigüedad del cliente, Consumo diferido, Número de quejas y Número de contactos.

4. CONJUNTO DE DATOS CON INFORMACIÓN RELEVANTE DE CONTRATOS

La compañía facilita dos conjuntos de datos para el estudio de retención de clientes:

1. Conjunto de datos de programas:

Información: Datos relacionados con los contratos de previsión exequial de la compañía.

Tamaño: 71.492 contratos.

Atributos: 20, incluyendo variables como tipo de programa, cantidad de personas y mascotas, y valor de la cuota.

2. Conjunto de datos de beneficiarios:

Información: Datos relacionados con las personas beneficiarias de los contratos de previsión exequial.

Tamaño: 310.521 beneficiarios.

Atributos: 14, incluyendo variables como edad y profesión.

La preparación de datos es un proceso fundamental en el análisis de datos. Consistió en corregir los errores, eliminar o imputar los datos faltantes y estandarizar los formatos de los datos y los nombres de las variables de las dos fuentes de datos, estas fueron unidas por llave primería nombre del titular. Dentro de las transformación más relevantes se imputa el valor de la cuota en los registros vacíos teniendo en cuenta el número de mascotas y personas inscritas, se realiza un cálculo de la duración de programa donde para los programas activos se tiene en cuenta la fecha actual, se calculan la cantidad de descuentos obtenidos, se realiza un análisis de texto a las observaciones realizadas por clientes durante llamadas telefónicas usando el algoritmo de código abierto SentimentIntensityAnalyzer, se obtiene la cantidad de gestiones de recaudo exitosas, se calcula la latitud, longitud y nivel socioeconómico usando el algoritmo de pago ArcGis y se extrae la cantidad de personas vinculadas a un plan de acuerdo a su parentesco y a su profesión. Finalmente, para los atributos categóricos, se agrupan las clases si estas no representan el 1% del total de los datos en una variable denominada “otros”.

Tras estas transformaciones se obtiene un único conjunto de datos con 27.412 registros y 50 atributos, atributos que serán objeto de estudio en la selección de variables. Con estos datos finales, se presenta a la entidad el análisis descriptivo de los datos entregados dentro de características importantes se mencionan:

- La mayoría de los clientes están en una edad entre los 40 a 60 años, con un nivel socio económico bajo entre estrato 2 y 3.
- Tienen entre 4 a 8 personas inscritas como beneficiaros del plan exequial.

- El promedio del valor de las cuotas ronda los $20000 y son pocos los valores que se alejan del promedio.
- La moda de cantidad de estados activos son 3, donde los días que han estado activo rondan 214 días.
- Los planes en su mayoría presentan gestiones de recaudo entre 4 a 8 veces por contrato.
- El motivo principal de deserción ha sido relacionado el incumplimiento de pago o dificultad para ubicar al usuario principal.

Dicha información se presenta a la entidad con el objetivo de compartir zonas de valor comercial para impulsar campañas dirigidas, ya que se observaron zonas sin clientes en estratos medios-altos.

5. DETERMINAR LINEAS DE NEGOCIO, TÉCNICAS Y VARIABLES DE MAYOR INCIDENCIA

El conjunto de datos relacionado con los programas presenta un variable llamada tipo de programa, esta variable hace referencia a las líneas de negocio de la compañía, planes empresariales o familiares. Tras una conversación con los expertos de la compañía, informan que los planes familiares tienen un comportamiento distinto a los planes empresariales. Múltiples planes empresariales pueden verse afectados por las decisiones de una sola empresa, mientras que los planes familiares se ven afectados solo por el titular. La Ilustración 5 evidencia la distribución de registros por líneas de negocios y al limitar el alcance solo a planes familiares ocasiona una reducción significativa del conjunto de datos superior a los 30 mil registros.

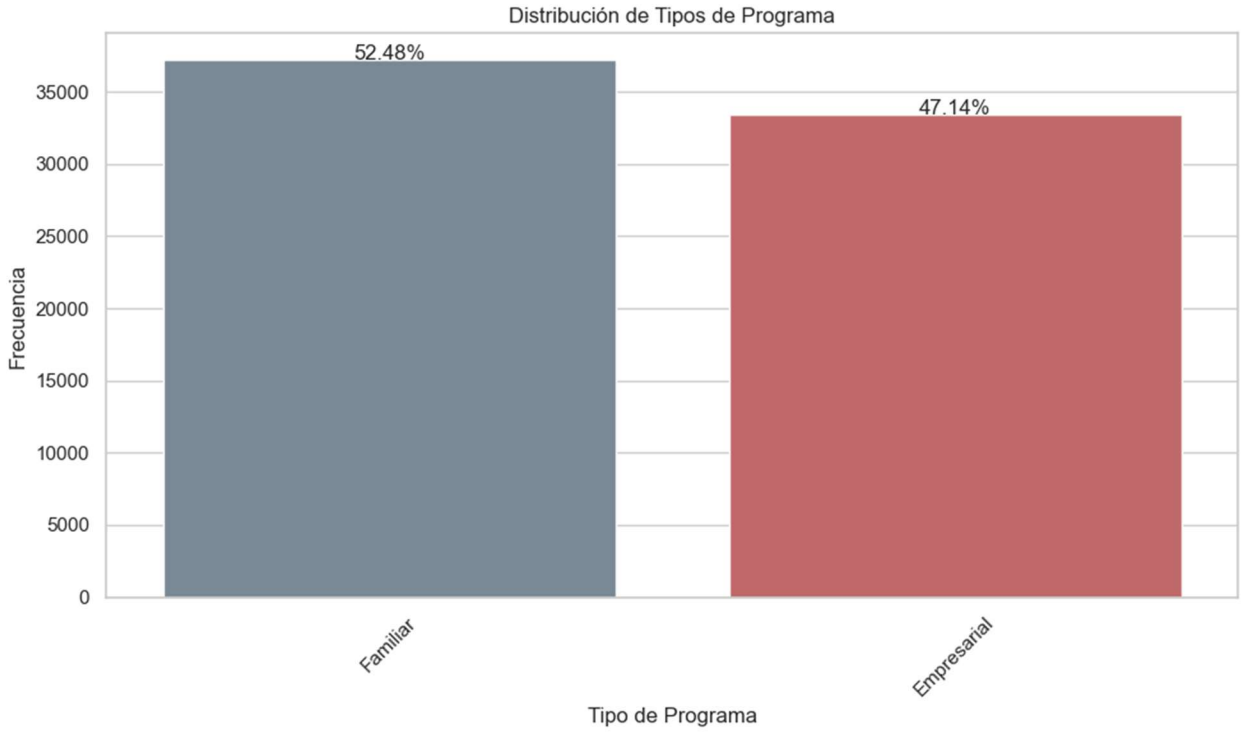


Ilustración 6. Distribución líneas de negocio [Autor,2023]

Adicionalmente, en conjunto con el director del proyecto y tras una revisión bibliográfica se determina que se aplicaran 3 técnicas de aprendizaje automático y 1 de aprendizaje profundo, Clasificador De Aumento De Gradiente, Aumento Del Gradiente Extremo y Clasificador De Impulso Adaptativo y Perceptrón Multicapa.

5.1. DETERMINACION DE VARIABLES DE MAYOR INCIDENCIA

Para determinar las variables que tienen mayor incidencia o capacidad de aprendizaje, se utilizaron 2 test: el chi-cuadrado para variables categóricas y pruebas t-Student para variables numéricas. En el caso de las variables categóricas, se agruparon algunos valores en la variable "otros" debido al alto volumen de valores únicos con poca frecuencia o inferior al 1% de la cantidad de los datos. Adicionalmente a las variables numéricas se les realiza un análisis de correlación para evitar multicolinealidad.

5.1.1. VARIABLES NUMÉRICAS

La prueba t-Student evidenció que las variables numéricas, excepto los cambios en la cuota presentan diferencias significativas en las medias entre usuarios activos e inactivos. Esta disparidad sugiere la existencia de factores que influyen en la actividad del usuario. La razón de la presente disparidad varía de acuerdo con la característica y se recomienda realizar un análisis más profundo para comprenderla. No obstante, para el presente estudio, la existencia de esta diferencia es crucial, ya que permite al modelo aprender sobre ella y, por lo tanto, mejorar su capacidad para predecir la actividad del usuario. En otras palabras, la diferencia en las medias proporciona información valiosa al modelo sobre las características que distinguen a los usuarios activos de los inactivos. A continuación, se presentan algunas diferencias observadas:

Se espera que los titulares con mayor edad (promedio de 7 años) sean más propensos a tener el programa activo. El valor de la última cuota evidencia que cuanto mayor es el valor, más propenso es a que el plan esté activo. Sin embargo, una gran cantidad de factores afecta esta variable, como la fecha de pago del plan o la cantidad de servicios, por lo que no se tendrá en cuenta en el modelo. Por otro lado, el día de pago ideal acordado parece tener una incidencia de 1 día en la rapidez con que se paga el plan. La duración del plan también tiene una incidencia, donde después de 2000 días, se podría ser más propenso a tener el plan activo. La cantidad de facturas generadas también evidencia una mayor cantidad de planes activos, como se observa en la Ilustración 6, pero para evitar información redundante, solo se tendrá en cuenta la duración.

En el caso de los descuentos otorgados, se evidencia que, tanto en porcentaje como en valor, afectan a la actividad o inactividad de un tomador. En especial, se tendrá en cuenta el porcentaje, ya que los valores relativos pueden compararse a lo largo del tiempo. En el caso de las gestiones, se evidencia que cuantas más gestiones se realicen, sean exitosas o no, es más propenso a estar

inactivo. Finalmente, el promedio de las edades de los inscritos parece ser menor para los beneficiarios de programas activos.

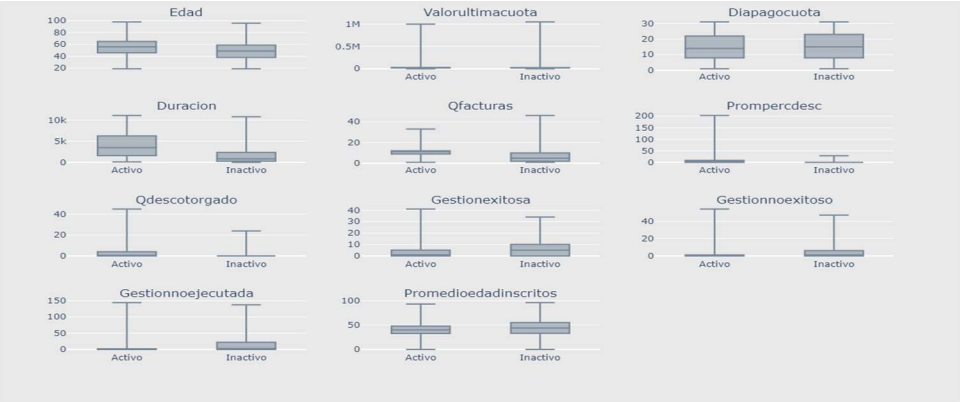


Ilustración 7. Variables Numéricas [Autor,2023]

Una vez identificado las variables más significativas, se realizó una matriz de corrección para evitar multicolinealidad por encima de 0.7; pero no se evidencia valor que supere umbral de multicolinealidad dicho anteriormente, como se observa en la Ilustración 7:

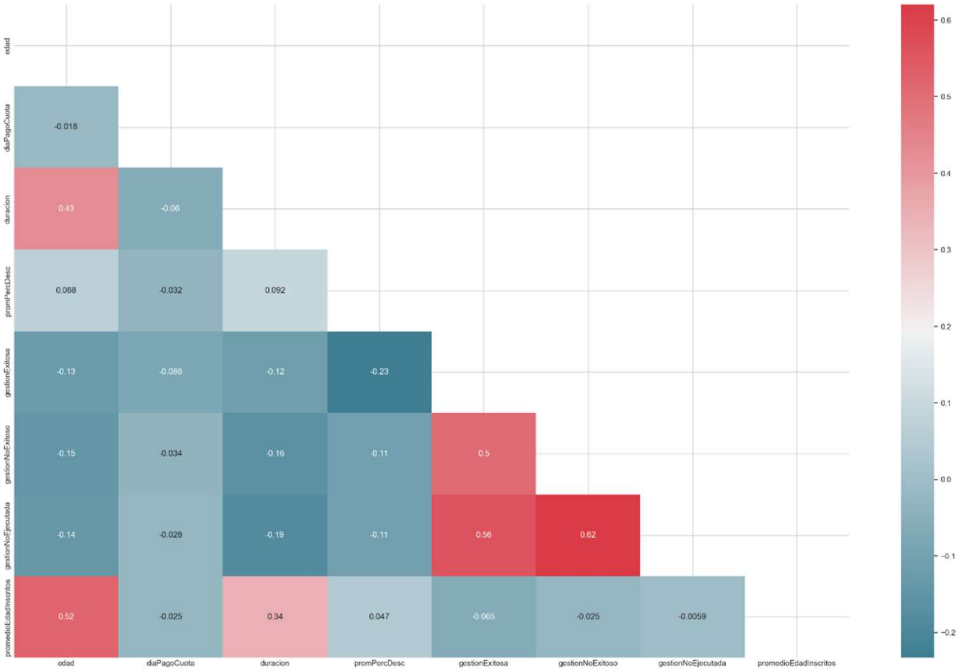


Ilustración 8. Matriz de Correlación Variables Numéricas [Autor,2023]

5.1.2. VARIABLES CATEGÓRICAS

Tras realizar la prueba chi-cuadrado, observamos que la cantidad de personas y mascotas representa una diferencia significativa. En concreto, los casos de 0 mascotas, 0 personas y 5 personas presentan mayor cantidad de inactivos, frente a las otras categorías. En cuanto a las profesiones que tiene el tomador como sus beneficiarios, se observa que hay profesiones que son menos propensas a desertar, como son pensionados, agricultores y comerciantes. El nombre del plan permite ver mayor cantidad de actividad, como Remanso o tener familiares afiliados y etiquetados, permite evidenciar mayor cantidad de planes activos en contraste con quienes no tienen afiliados familiares, como se observa en la Ilustración 8. Finalmente, se realizó una exploración de las palabras más significativas entre los grupos de sentimiento positivo o negativo, pero no tenía sentido su clasificación.

A continuación, se presentan un ejemplo de variables categóricas con un valor p (valor de probabilidad) dada por la prueba chi-cuadrado inferior a 0.05, en estas gráficas se presentan de color rojo los usuarios activos y en gris los usuarios inactivos.

Ilustración 9. Variables Categóricas [Autor,2023]

Las variables que se incluirán en el estudio de las diferentes técnicas de aprendizaje automático son las siguientes: cantidad de personas y mascotas inscritas, localidad, nombre del plan, agricultores inscritos, comerciantes inscritos, docentes inscritos, empleados inscritos, estudiantes inscritos, personas dedicadas al hogar inscritos, personas dedicadas a oficios varios inscritas, pensionados inscritos, inscripción o no de la esposa, esposo, hermanos, hijas y la madre, el día que paga la cuota, el promedio de descuento, el promedio de gestiones de recaudo exitosas o no ejecutadas y el promedio de edad.

6. MODELAMIENTO DESERTORES

El presente proyecto usará este tipo de técnicas que permitan realizar una clasificación binaria para determinar si un cliente tiene propensión a la deserción a partir del conocimiento de las características de aquellos datos históricos presentados por clientes que hayan cancelado sus planes de previsión exequial de manera voluntaria o involuntaria. Una característica es una propiedad medible del fenómeno observado, y se define la observación como cada evento independiente que fue registrado. Para este caso cada observación representa a un cliente que haya o no presentada deserción y las características es cada una de las variables que caracterizan al cliente y su comportamiento.

A continuación, se presenta un enfoque sistemático para determinar el mejor modelo que permita identificar posibles desertores de planes de previsión exequial como se observa en la Ilustración 9. Para ello, se siguieron seis etapas fundamentales: Una preparación de datos de entrenamiento y prueba, creación de modelo inicial, balanceo de clases, entrenamiento y evaluación, selección y optimización de hiperpárametros e implementación en la empresa.

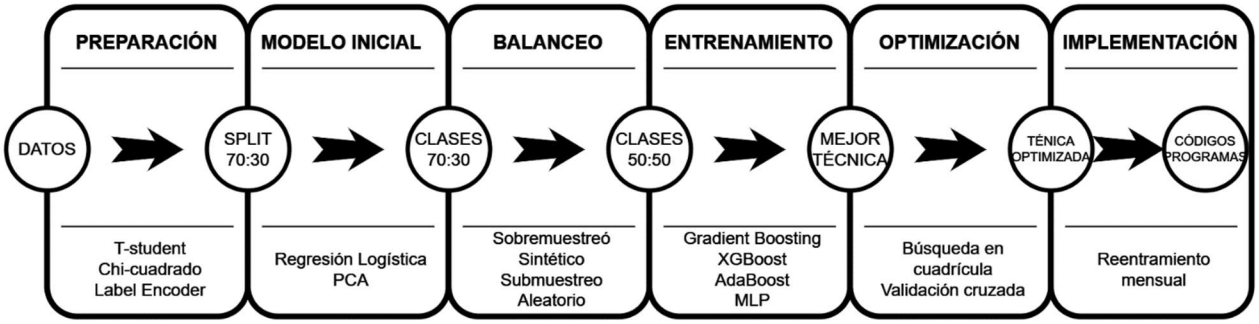


Ilustración 10. Etapas modelamiento [Autor2023]

6.1. PREPARACIÓN DE DATOS DE ENTRENAMIENTO Y PRUEBAS

Se reemplaza la clase activo por 0 y la clase inactivo por 1, para poder ser usada la variable estado en los modelos, así mismo, se usa el LabelEncoder para codificar variables categóricas en números enteros. Este método es útil para muchos algoritmos de aprendizaje automático, que requieren que las variables sean numéricas, adicionalmente favorece la interpretación de los datos, mejora la precisión de algunos modelos y reduce la dimensionalidad de los datos.

El LabelEncoder codifica las variables categóricas de la siguiente manera: Asigna un número entero a cada categoría y las categorías se ordenan alfabéticamente. Finalmente, se dividen los datos en 70% datos de entrenamiento y 30% datos de pruebas como se evidencia en la siguiente Ilustración 10.

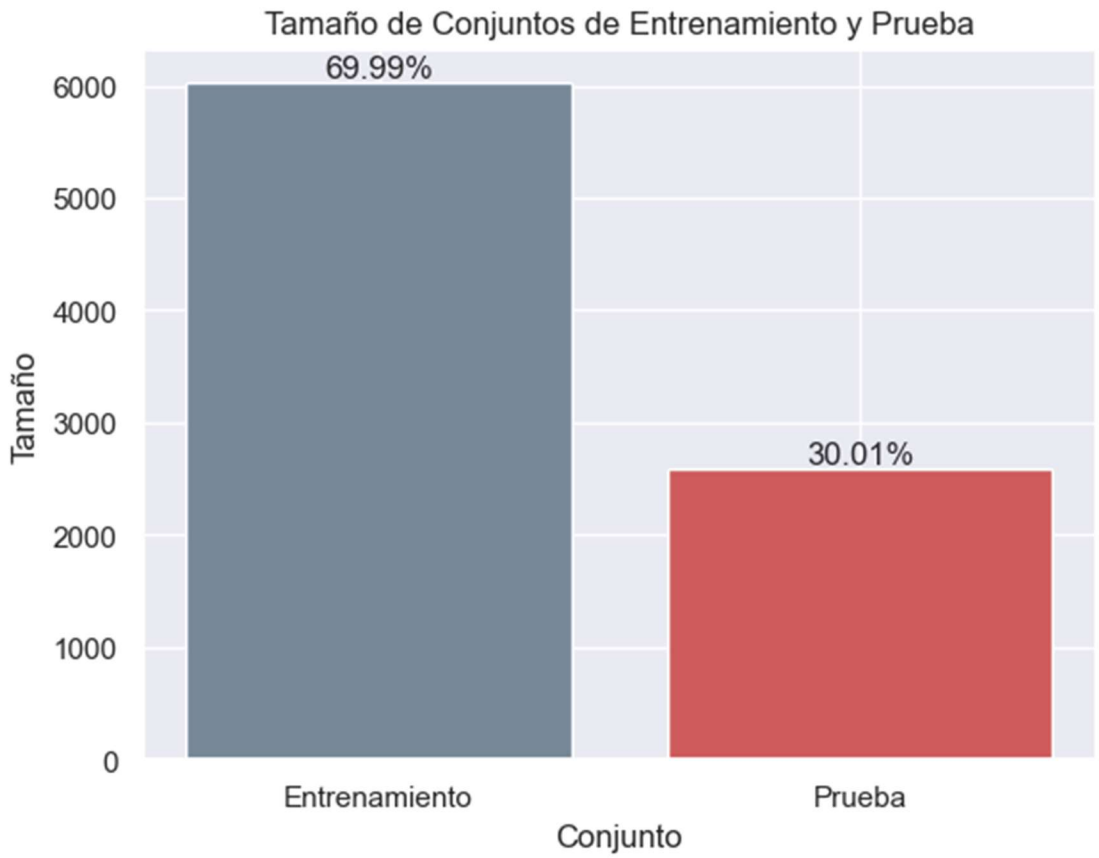
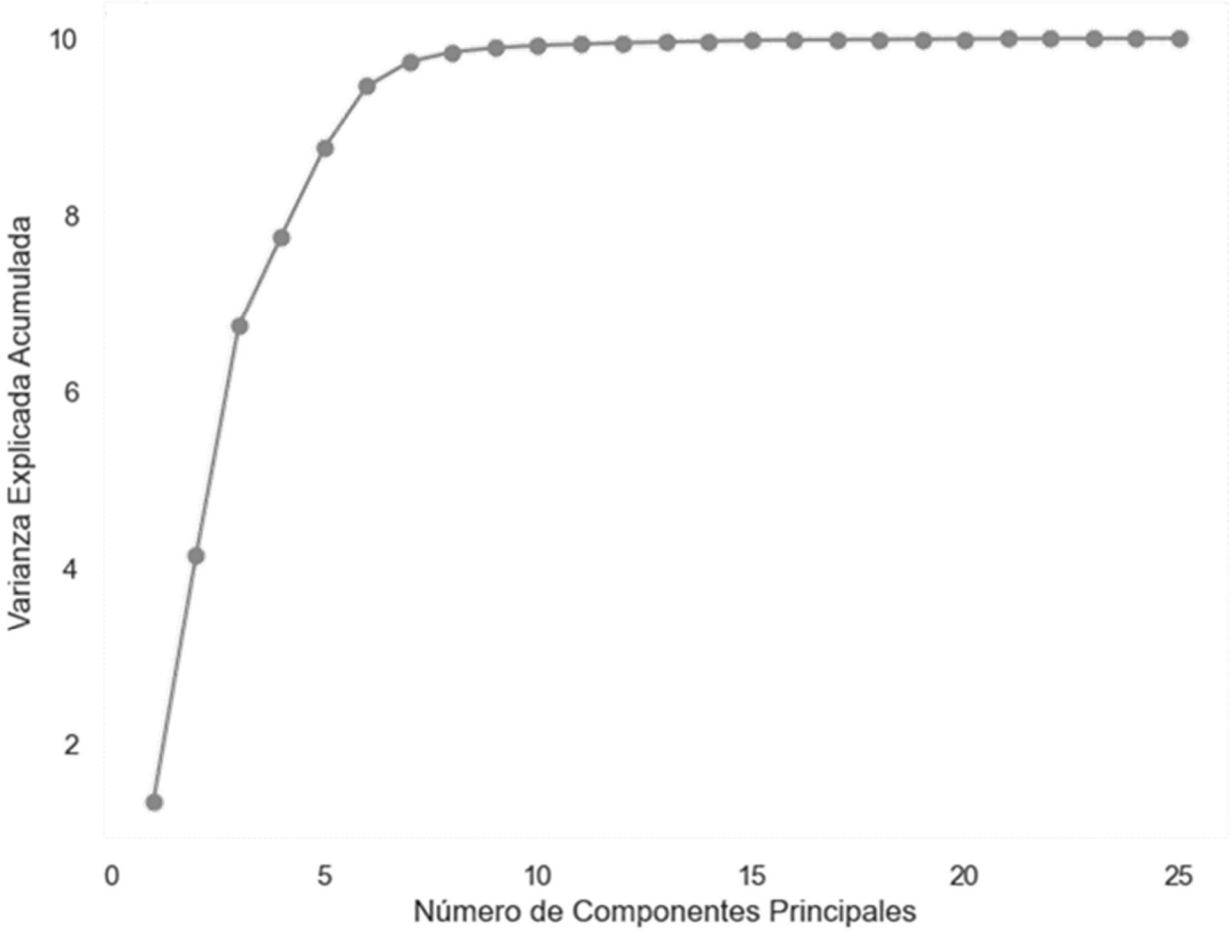


Ilustración 11. Test y Prueba [Autor, 2023]

6.2. ANÁLISIS PCA Y CREACIÓN DEL MODELO INICIAL

En esta etapa, se crea un modelo inicial de regresión logística. En primer lugar, se valida que las variables significativas tienen un impacto positivo en recall, en comparación con un modelo que incluye todas las variables, esto se valida tras la evaluación de dos modelos diferentes de regresión logística donde en uno se incluyeron todas las variables y en el otro solo las variables significativas. Tras la evaluación se observó un aumento del recall de 0.38 a 0.61. Posteriormente, se opta por una técnica adicional de selección de características para comparar la selección de características anterior con un análisis de componentes principales (PCA) con 6 componentes, determinados mediante la técnica del codo, para reducir la dimensionalidad de los datos y eliminar posibles correlaciones redundantes.

Sin embargo, el PCA no mejora el recall respecto al modelo inicial con las variables significativas antes lo disminuye a 0.34. Además, dificulta la interpretación de los resultados. Por ello, se decide no continuar con el análisis y se continúa usando el modelo inicial con las variables significativas como observaremos a continuación en la Ilustración 11



7

Ilustración 12. Análisis componentes principales [Autor,2023]

6.3. BALANCEO DE CLASES

En muchos conjuntos de datos, una clase mayoritaria puede estar representada de forma desproporcionada en comparación con otra clase minoritaria. En este caso, se observa una relación de 70:30. Para abordar este desequilibrio, se utilizan dos técnicas de balanceo de clases: el submuestreo aleatorio y el sobremuestreo sintético. Ambas técnicas igualan la distribución de las clases en una relación de 50:50 como se observa en la Ilustración 12, lo que ayuda a evitar sesgos en el modelo.

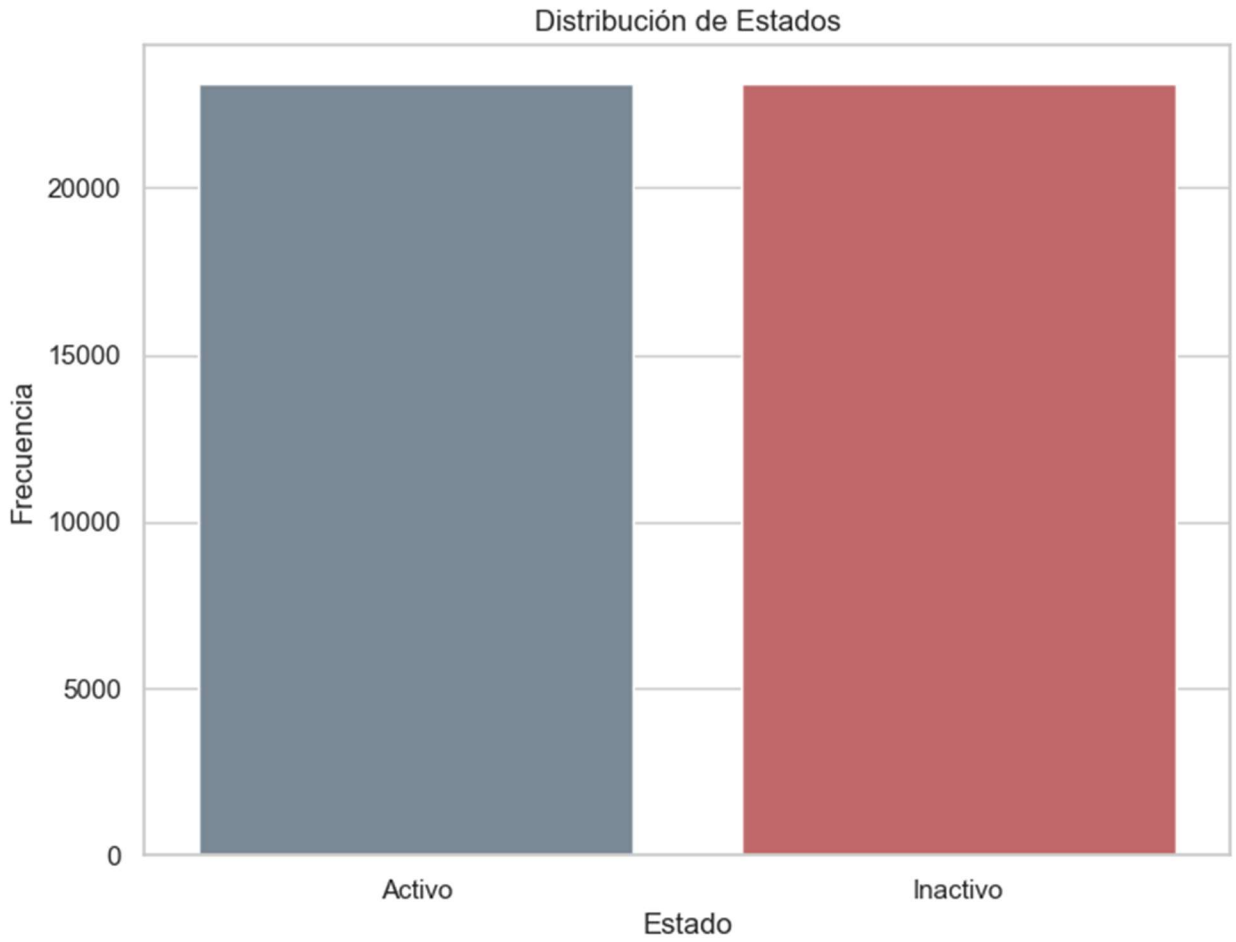


Ilustración 13. Balanceo [Autor, 2023]

Sin embargo, se observa que el sobremuestreo sintético podría introducir sesgos en los datos y puede hacer que el modelo se sobreajuste. Por ello, se decide continuar usando el submuestreo aleatorio, que es una técnica más robusta y que generaliza mejor a los datos nuevos.

6.4. ENTRENAR LAS TÉCNICAS, OBTENER MÉTRICAS DE EVALUACIÓN Y SELECCIONAR EL MODELO OPTIMO

En esta etapa, se entrenan y evalúan cuatro modelos diferentes de aprendizaje automático: XGBoost, AdaBoost, Perceptrón Multicapa y Gradient Boosting Classifier. Cada modelo se ajusta al conjunto de datos de entrenamiento y se evalúa su rendimiento mediante la métrica de “recall”, como se puede observar a continuación en la Ilustración 13.

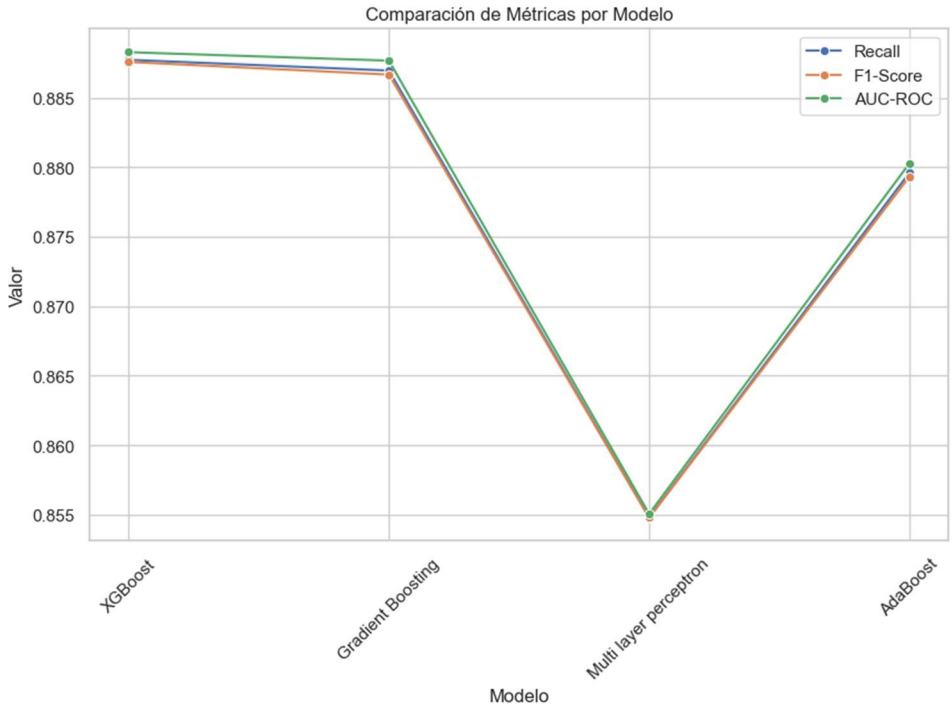


Ilustración 14.Métricas Modelos [Autor, 2023]

Los resultados de las métricas de evaluación evidencian alto rendimiento para clasificación de activos en nuestro modelo inicial de regresión logística. Esto significa que los cuatro modelos tuvieron un buen rendimiento al identificar a los clientes que seguirán siendo clientes, Sin embargo, el objetivo es identificar aquellos que no serán clientes, es allí, donde distinguir la métrica de recall para activos e inactivos toma relevancia. Podemos evidenciar en nuestro modelo inicial un recall para clasificar activos equivalentes a 0.98 mientras que para inactivos ejemplifica una reducción significativa equivalente a 0.61. A partir de allí, los esfuerzos del proyecto se enfocaron en aumentar esta métrica. A continuación, se presentan los resultados obtenidos:

Modelo	Precisión Activos	Precisión Inactivos	Recall Activos	Recall Inactivos	F1-score Activos	F1-score Inactivos
Regresión logística	0.93	0.84	0.98	0.61	0.95	0.71
PCA	0.88	0.73	0.98	0.32	0.44	0.32
Regresión logística sobremuestreo	0.89	0.90	0.90	0.89	0.89	0.89
Regresión logística submuestreo	0.80	0.85	0.86	0.79	0.83	0.82
XGBoost submuestreo	0.85	0.93	0.93	0.85	0.89	0.88
GBC submuestreo	0.85	0.94	0.94	0.83	0.89	0.88

MLP submuestreo	0.84	0.87	0.86	0.84	0.86	0.85
AdaBoost submuestreo	0.84	0.93	0.93	0.83	0.88	0.87

Tabla 3.Resultados métricos de evaluación [Autor, 2023]

La Ilustración evidencia los resultados presentados por la matriz de confusión del modelo con mayor métrica de recall que es XGBoost:

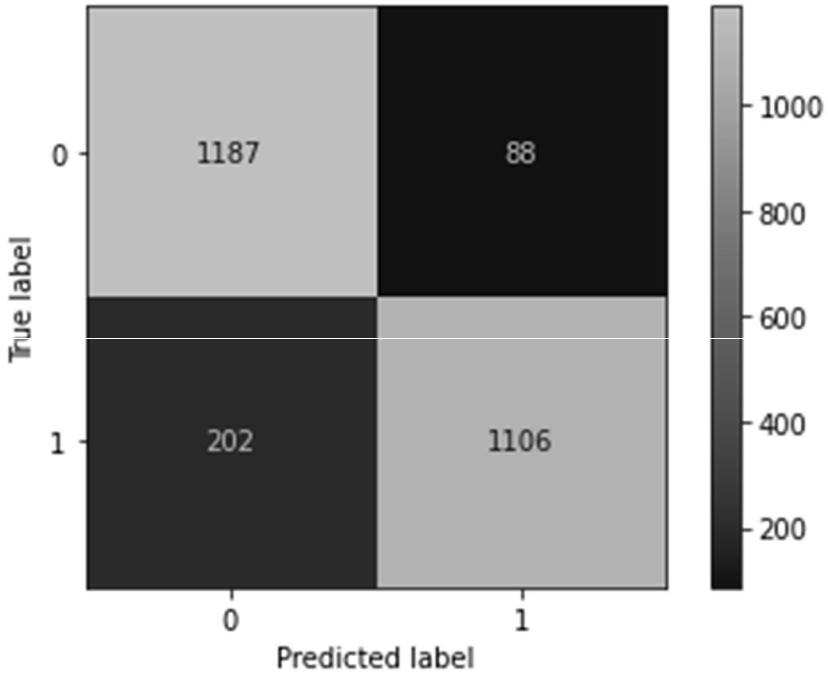


Ilustración 15. Matriz confusión mejor modelo [Autor, 2023]

6.5. SELECCIÓN DE HIPERPARÁMETROS PARA EL MEJOR MODELO

En esta última etapa, se realiza una búsqueda exhaustiva de hiperparámetros para el modelo XGBoost con el fin de encontrar la combinación óptima que maximice el rendimiento de la "recall". Se utiliza una validación cruzada con 5 iteraciones para evaluar cada combinación de parámetros.

Para la búsqueda de hiperparámetros, se utilizan técnicas como la búsqueda en cuadrícula. En este caso, se exploraron las combinaciones para los siguientes parámetros:

Learning Rate: 0.1, 0.01, 0.001

Number of Estimators: 100, 200, 300

Max Depth: 3, 5, 7, 9

Subsample: 0.8, 0.9, 1.0

Colsample by Tree: 0.8, 0.9, 1.0

Los resultados obtenidos evidencian que la combinación ideal es Colsample Bytree 0.9, Learning Rate: 0.1, Profundidad de los árboles: 7, Número de árboles: 300 y Subsample: 0.8.

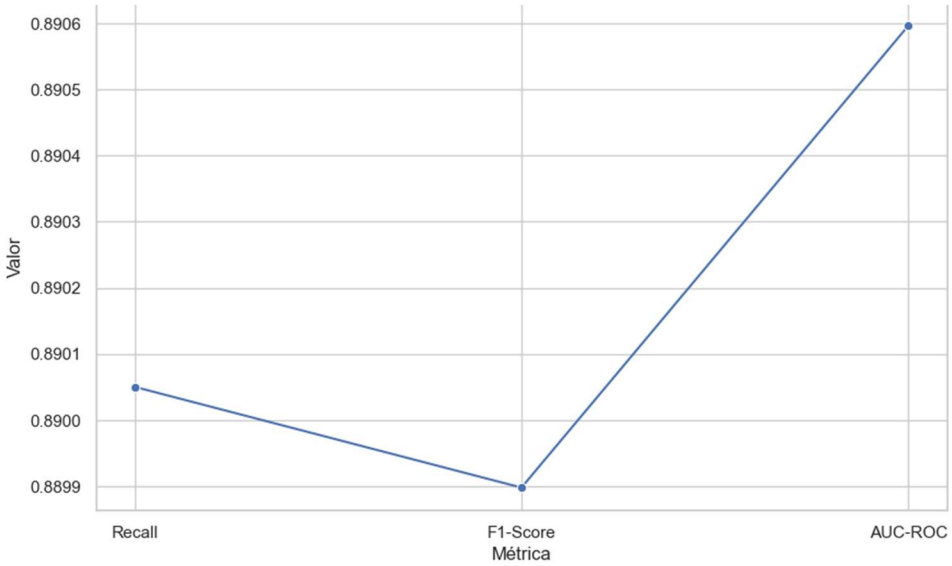


Ilustración 16.Métricas Mejor Modelo [Autor,2023]

7. IMPLEMENTACIÓN EN LA EMPRESA

Se han presentado las cinco etapas clave para determinar el mejor modelo óptimo para detectar posibles desertores de planes de previsión exequial. Cada etapa desempeño un papel fundamental en el proceso de selección del modelo más efectivo. En este capítulo, se detalla la implementación exitosa de la función de transformación de datos y el entrenamiento del modelo XGBoost optimizado en una computadora local. Esta etapa marca un hito significativo en el desarrollo de la tesis, ya que representa la aplicación práctica de los conocimientos teóricos y la preparación de los datos.

En este caso, XGBoost está presentando una métrica de "recall" alta, equivalente al 89%. El modelo puede predecir correctamente el 89 % de los clientes que desertarán. En la tabla 1 se presentan los resultados de la métrica de "recall" en cada entrenamiento:

Modelo	Balanceo	Recall Inactivos
Regresión logística	No	0.61
PCA	No	0.32
Regresión logística	Sobremuestreo 50:50	0.89
Regresión logística	Submuestreo 50:50	0.79
XGBoost	Submuestreo 50:50	0.85
GBC	Submuestreo 50:50	0.83
MLP	Submuestreo 50:50	0.84
AdaBoost	Submuestreo 50:50	0.83
XGBoost optimizado hiperpárametros	Submuestreo 50:50	0.89

Tabla 4.Evaluación de los modelos [Autor, 2023]

Antes de la implementación, se configuró el entorno de desarrollo en la computadora local. Se aseguró la instalación de todas las bibliotecas, sistema operativo y dependencias necesarias, incluyendo XGBoost y otras herramientas de procesamiento de datos.

Se diseñó e implementó una función para la transformación de datos, entrenamiento y predicción. Esta función aborda aspectos específicos del conjunto de datos estudiados en etapas anteriores, como la unificación de bases de datos, selección de características, manipulación de valores nulos, la normalización de atributos y la codificación de variables categóricas. Se presta especial atención a la escalabilidad y eficiencia para manejar grandes conjuntos de datos.

Así mismo, la función reentrena el modelo usando los últimos datos provisionados y predice los tomadores propensos a desertar que se encuentren activos actualmente. Los resultados obtenidos en una primera iteración del modelo optimo arrojan 570 programas propensos a desertar. Dado esta cantidad de programas sobre los cuales la empresa debe intervenir, se decide reentrenar y ejecutar el modelo una vez al mes en una máquina provisionada por la empresa. El resultado es dos archivos de texto CSV con cada uno de los códigos de programa propensos a desertar y sus atributos.

8. CONCLUSIONES

Este proyecto representa un hito significativo para la empresa funeraria ya que no se contaba con este conocimiento y presentaba un vacío crucial en el conocimiento de los datos, los cual ayudo a proporcionar a la empresa una comprensión más profunda y detallada de sus clientes. Con lo cual se buscó establecer un precedente importante para la mejora continua y la innovación en el ámbito funerario.

En él se utilizaron técnicas de inteligencia artificial para predecir la deserción de un cliente del contrato de planes de previsión exequial en una empresa funeraria. Se determinó que la línea de negocio a trabajar es del segmento persona, se realizaron análisis estadísticos donde se identificaron 26 características con diferencias estadísticas significativa, adicional a esto el algoritmo desarrollado, basado en la técnica XGBoost, fue el mejor calificado bajo la métrica recall, llegando al de 0.89, lo cual significa que tiene la capacidad de identificar correctamente el 89% de los clientes propensos a desertar.

Los resultados del estudio son de gran relevancia para la empresa funeraria, dado que es su primer acercamiento a la aplicación de inteligencia artificial que les permiten identificar los clientes en riesgo de desertar y desarrollar estrategias para retenerlos. El algoritmo se implementará en una máquina provisionada por la empresa y se reentrenará mensualmente para actualizar sus resultados.

8.1. TRABAJOS FUTUROS

Se pueden considerar las siguientes acciones para mejorar el rendimiento del modelo.

Selección de variables

En la etapa de selección de variables, se usó PCA y las técnicas estadísticas t-Student, chi-cuadrado donde se consideraron las variables que tenían una diferencia significativa entre las medias de las clases. Sin embargo, se podría considerar utilizar otras técnicas de selección de variables, como el análisis de importancia de variables o LDA que puede proporcionar información adicional sobre la contribución de cada variable al modelo.

Experimentar con otros modelos de aprendizaje automático

En la etapa de selección de modelos, se consideraron cuatro modelos de aprendizaje automático. Se podría considerar experimentar con otras técnicas como Support Vector Machines o RNN, que pueden tener un mejor rendimiento en este problema.

Reducir la complejidad del modelo

El modelo XGBoost seleccionado tiene una profundidad de árbol de 7, lo que puede indicar que el modelo es complejo y puede estar sujeto a sobreajuste. Se podría considerar reducir la profundidad del árbol para mejorar la generalización del modelo ajustando el hiperpárametro “Depth” dentro de la definición de XGBoost.

Mejora clasificación de observaciones

Dentro del estudio se pretendía usar el sentimiento captado en las observaciones usando un algoritmo de código abierto, sin embargo, se observó poca relevancia en los comentarios usando nubes de palabra. Se propone mejorar las futuras observaciones realizadas en estos comentarios para en un futuro tener más asertividad en la determinación del sentimiento. Así se podría incluir esta variable dentro del modelo construido. Mediante los comentarios podremos obtener

Determinar nivel socioeconómico

Dentro del estudio se observó que el nivel socioeconómico tiene una diferencia significativa, sin embargo, más del 75% de los datos no presentan, para obtener los datos faltantes, se puede usar el servicio de pago ArcGis o similar. Esta variable no se incluyó en el estudio de las diferentes técnicas porque está fuera del alcance del proyecto.

El modelo XGBoost presentado en el estudio presenta "recall" equivalente al 89%. Sin embargo, se podrían considerar las recomendaciones anteriores para mejorar aún más el rendimiento del modelo.

Esta información adicional podría ser útil para la empresa, ya que le permitiría tomar decisiones informadas sobre cómo mejorar el modelo y, en última instancia, mejorar la retención de clientes.

Propuestas de retención de clientes

La retención de clientes en la industria funeraria no solo se trata de ofrecer servicios de alta calidad, sino también de brindar apoyo emocional a largo plazo y mostrar agradecimiento por la confianza continua de las familias en duelo. Cuando identificamos posibles desertores, es crucial actuar rápidamente. Aquí se proponen tres opciones clave para retener a esos clientes valiosos:

Personalización de Servicios: Ofrecer servicios personalizados que se ajusten a las necesidades y creencias específicas de la familia en duelo. Desde la elección de música hasta la decoración, adaptar cada detalle para brindar un servicio excepcional.

Programas de Apoyo a Largo Plazo: Desarrollar programas de apoyo emocional a largo plazo para las familias en duelo, incluyendo asesoramiento y eventos conmemorativos anuales. Estos gestos demuestran el compromiso continuo de la funeraria con el bienestar de la familia.

Descuentos y Beneficios para Clientes Frecuentes: Ofrecer descuentos y beneficios exclusivos a clientes que han utilizado los servicios de la funeraria por un largo tiempo. Esto fomenta la fidelidad y demuestra gratitud por su elección repetitiva.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] Harvard Business Review, «The Power of Customer Retention,» 2009. [En línea].

[2] Forrester, «The Customer Experience Revolution,» 2022. [En línea].

[3] C. P. Vélez Zapata, «Modelo de riesgo crediticio para la empresa funeraria,» *Revista Ciencias Estratégicas,* pp. 33-47, 06 2009.

[4] Panasef, «Radiografía del sector funerario,» *Panasef,* 2021.

[5] J.-H. Ahna, S. Hana y Y.-S. Lee, «Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry,» *Telecommunications Policy,,* pp. 552-568, 2006.

[6] I. Bose y X. Chen, «Hybrid models using unsupervised clustering for prediction of customer churn,» *Journal of Organizational Computing & Electronic Commerce,* pp. 133-151, 2009.

[7] A. Khan, S. Jamwal y M. Sepehri, «Applying Data Mining to Customer Churn Prediction in an Internet Service Provider,» pp. 8-14, 2010.

[8] N. Guangli, Z. Lingling y S. Yong, «The Analysis on the Customers Churn of Charge Email Based on Data Mining Take One Internet Company for Example,» *Institute of Electrical and Electronics Engineers,* pp. 843-847, 2011.

[9] J. D. Falla Arango, «Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático.,» Bogotá, 2021.

[10] J. H. Friedman, , «Greedy function approximation: A gradient boosting machine,» *The Annals of Statistics 29,* p. 1189–1232., 2001.

[11] C. Bong-Horng, T. Ming-Shian y S. Cheng, «Toward a hybrid data mining model for customer retention,» *Knowledge-Based Systems,* pp. 703-718, 2007.

[12] B. &. Company, «The Customer Loyalty Imperative,» 2017.

[13] I. H. F. E. &. H. M. A. Witten, «Data mining: Practical machine learning tools and techniques,» vol. 3er ed., nº New York, NY: Morgan Kaufmann, 2011.

[14] T. T. R. &. F. J. H. Hastie, «The elements of statistical learning (2nd ed.). New York, NY: Springer.,» 2009.

[15] I. T. Jolliffe, «Principal component analysis (2nd ed.),» *Springer,* 2002.

[16] J. M. &. G. A. Calero, «Aprendizaje automático: conceptos básicos y aplicaciones,» *McGraw-Hill,* vol. Madrid, 2022.

[17] A. F. Echeverri Giraldo, «Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones,» Medellin, 2019.

[18] B. D. Cobeñan Terán, «Modelo de predicción de deserción de clientes de tarjetas de crédito,» Guayaquil, 2016.

[19] Scikit-learn, «Introducción a los Árboles de Decisión,» [En línea]. Available: https://scikit-learn.org/stable/modules/tree.html . [Último acceso: 09 12 2023].

[20] L. W. A. M. Z. M. X. S. Q. Y. &. G. J. Zang T., «Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning,» *Advancing Earth and Space Science,* 2021.

[21] T. &. G. C. Chen, «XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.,» *ACM,* pp. 785-795, 2016.

[22] Y. Freund y . R. E. Schapire, «A desicion-theoretic [sic] generalization of on-line learning and an application to boosting,» Springer Berlin Heidelberg, Berlin, 1995.

[23] A. Ng, «Fundamentos de aprendizaje automático,» *Coursera,* 2022.

[24] B. Y. &. C. A. Goodfellow I., «Deep Learning,» *MA: MIT Press,* p. 15, 2016.

[25] A. Y. N. a. E. S. C. Michael I. Jordan, «Machine Learning for Healthcare: Fundamentals and Applications,» *MIT Press,* 2015.

[26] P. A. G. J. R. J. A. K. Jain, «ROC curve evaluation of edge detector performance,» Vols. %1 de %219, Número:** 4, Abril 1997, pp. 401-410.

[27] R. Feinberg y M. Trotter, « Immaculate deception: the unintended negative effects of the CRM revolution: maybe we would be better off without customer relations management,» *Defying the limits,* pp. 26-31, 2001.

[28] C. Huigevoort, «Customer churn prediction for an insurance company,» 2015.

[29] J. Ljungehed, «Predicting Customer Churn Using Recurrent Neural Networks,» DEGREE PROJECT COMPUTER SCIENCE AND ENGINEERING,SECOND CYCLE, 30 CREDITS, STOCKHOLM, 2017.

[30] M. A. Santamaría Novoa, «Los productos de previsión exequial y los seguros exequiales: ¿son productos diferentes?, ¿qué implicaciones prácticas acarrea su ofrecimiento simultáneo en el mercado?,» UNIVERSIDAD EXTERNADO DE COLOMBIA, Bogotá, 2018.

[31] Amazon Web Services, «¿Qué es la informática en la nube?,» 2022.

[32] T. Chen y C. Guestrin, «XGBoost,» *InProceedings of the 22nd ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining,* 2016.

[33] Bain & Company, «The Customer Loyalty Imperative,» 2019. [En línea].

[34] L. Martínez, «La industria funeraria en México: un negocio que nunca muere,» *Revisa Nexos,* pp. 48-53, 2022.

[35] L. F. J. H. O. R. A. &. S. C. J. Breiman, «Classification and regression trees. Belmont, CA:

Wadsworth International Group,» 1984.

[36] T. T. R. &. F. J. H. Hastie, «The elements of statistical learning (2nd ed.),» Vols. %1 de %2New York, NY: Springer., 2009.

[37] FUNOS, «Informe de tendencias del sector funerario,» 2021.

10. ANEXOS

10.1. Anexo 1- Diccionario de datos

Nombre Atributo	Descripción
codigoPrograma	Código único asignado al programa de previsión exequial
edad	Edad del titular del plan
qPersonas	Cantidad de personas
qMascotas	Cantidad de mascotas
localidad	Ciudad donde se realizó la venta del programa
valorUltimaCuota	Valor actual de la cuota mensual
idTomador	Cedula del titular del plan
tomador	Nombre del pagador del programa, responsable del contrato
diaPagoCuota	El día en que se paga la cuota del plan durante un mes
inactivoadmin	Cantidad de veces que un titular se reporta como inactivo por errores administrativos
inactivocosto	Cantidad de veces que un titular se reporta como inactivo por costo del servicio
inactivofallecimiento	Cantidad de veces que un titular se reporta como inactivo por utilización del servicio
inactivoinactivo	Cantidad de veces que un titular se reporta como inactivo, pero se desconoce el motivo
inactivoincumplimiento	Cantidad de veces que un titular se reporta como inactivo por incumplimiento en pago
inactivoinfluencia	Cantidad de veces que un titular se reporta como inactivo por influencia de personas
inactivomala	Cantidad de veces que un titular se reporta como inactivo por mala atención
inactivorecuperado	Cantidad de veces que un titular se reporta como inactivo y fue recuperado
inactivoubicacion	Cantidad de veces que un titular se reporta como inactivo por cambios en ubicación
duración	Duración del plan

estado	Estado actual, activo o inactivo
qFacturas	Cantidad de facturas generadas
promPercDesc	Promedio del porcentaje de descuentos otorgados
qDescOtorgado	Cantidad de facturas que han recibido un descuento
cambioCuota	Diferencia entre el valor de la cuota inicial y la actual
gestionCambioCuotas	Cantidad de veces que se ha gestionado el cambio de la cuota
gestionExitosa	Cantidad de gestiones de recaudo exitosas
gestionNoExitoso	Cantidad de gestiones de recaudo no exitosas
gestionNoEjecutada	Cantidad de gestiones de recaudo programadas, pero no ejecutadas
sentimiento	Sentimiento obtenido de las observaciones realizadas por clientes en llamadas usando el algoritmo SentimentIntensityAnalyzer
promedioEdadInscritos	Promedio de la edad de las personas inscritas
nombrePlan	El nombre del plan
profesionTomador	Profesión titular
agricultor	Cantidad de agricultores inscritos al plan
comerciante	Cantidad de comerciantes inscritos al plan
docente	Cantidad de docentes inscritos al plan
empleado	Cantidad de empleados inscritos al plan
estudiante	Cantidad de estudiantes inscritos al plan
hogar	Cantidad de personas dedicadas al hogar inscritos al plan
independiente	Cantidad de independientes inscritos al plan
oficiosVarios	Cantidad de personas dedicas a oficios varios inscritos al plan
pensionados	Cantidad de pensionados inscritos al plan
esposa	Evidencia inscripción de la esposa
esposo	Evidencia inscripción del esposo
hermana	Cantidad de hermanas del titular inscritas al plan
hermano	Cantidad de hermanos del titular inscritos al plan
hija	Cantidad de hijas inscritas al plan
hijo	Cantidad de hijos inscritos al plan

madre	Evidencia inscripción de la madre
nivelSocioeconomico	Estrato socioeconómico del titular

Tabla 5. Diccionario de datos [Autor, 2023]

10.2. Anexo 2- Análisis plataformas inteligencia espacial

En este análisis exhaustivo de los servicios de geocodificación, se lleva a cabo una evaluación integral de la precisión en relación con la capacidad de respuesta de los servicios disponibles o de fácil acceso en el área de estudio al momento de la investigación. Para lograrlo, se desarrolla un código en Python haciendo uso de la biblioteca Pandas, permitiendo así realizar solicitudes a múltiples servicios de geocodificación.

Este código recopila los resultados de las coordenadas geográficas solicitadas y los almacena en un DataFrame. Los servicios evaluados incluyen nombres destacados en el campo de la geocodificación, como Google, Bing, Nominatim y Sitidata. La obtención de coordenadas precisas se convierte en el núcleo del análisis, proporcionando una visión detallada de la calidad y eficacia de cada servicio.

La geocodificación es el proceso de convertir direcciones o descripciones de ubicaciones en coordenadas geográficas (latitud y longitud). En este análisis, examinaremos cuatro geocodificadores: Nominatim, Google Geocoding API, Bing Maps API y SitiData Geocoding API. Analizaremos sus características, ventajas y desventajas para ayudar al negocio a decidir sobre cual servicio es más conveniente para la organización.

10.2.1. Metodología

La evaluación de la precisión de los geocodificadores es un paso crucial para determinar cuál de las opciones disponibles es la más adecuada para satisfacer nuestras necesidades. En este estudio, seguimos una metodología rigurosa que comprende los siguientes pasos:

Paso 1: Recopilación de Datos de Prueba y Selección de Puntos de Referencia

Para llevar a cabo una evaluación precisa, es fundamental contar con un conjunto de datos de prueba diversificado. En este casi utilizamos la base de ubicaciones de direcciones regionales del ICBF, que proveen una base de ubicaciones distribuida en todo el país lo que permite una variabilidad de opciones para el estudio. Adicional se establece una ubicación verificada a mano, a partir de la cual se van a hacer las validaciones con las salidas de los geocodificadores.

Paso 2: Geocodificación de las Direcciones de Prueba

Cada uno de los geocodificadores en evaluación se utiliza para convertir las direcciones del conjunto de datos de prueba en coordenadas geográficas. Se presta especial atención a la conIlustraciónción óptima de cada geocodificador utilizando Geopy, el cual permite una consolidación de solicitudes a múltiples servicios de geocodificación, asegurando así que se

obtengan los resultados más precisos posibles.

Paso 3: Comparación con los Puntos de Referencia

Las coordenadas geográficas generadas por los geocodificadores se comparan exhaustivamente con las coordenadas reales de los puntos de referencia. Para cuantificar la precisión de cada geocodificador, se calcula la distancia euclidiana entre las coordenadas proporcionadas y las coordenadas reales. Además, se emplean métricas de error, como el error absoluto medio (MAE) y el error cuadrático medio (MSE), para obtener una evaluación más detallada de las discrepancias.

Paso 4: Análisis de Resultados y Conclusiones

Los resultados de la comparación se someten a un análisis detenido. Esto implica determinar cuál de los geocodificadores ofrece la mayor precisión en función de las métricas de error utilizadas. Se pueden realizar análisis por regiones geográficas para identificar posibles diferencias en la precisión en áreas específicas

10.2.2. Servicios de Geocodificación

Nominatim:

Nominatim es un geocodificador de código abierto que utiliza datos de OpenStreetMap (OSM) para convertir direcciones y descripciones de ubicaciones en coordenadas geográficas. Es desarrollado y mantenido por la comunidad de OpenStreetMap.

Ventajas:

- Fuente abierta y gratuita.
- Datos de OpenStreetMap de alta calidad.
- Soporte internacional.
- API de acceso público sin necesidad de una clave de API.

Desventajas:

Limitaciones de uso para aplicaciones comerciales.

Rendimiento puede ser lento en comparación con servicios comerciales.

Actualizaciones de datos pueden no ser tan frecuentes.

Google Geocoding API:

Google Geocoding API es un servicio de geocodificación proporcionado por Google que utiliza su vasta base de datos geoespacial.

Ventajas:

- Datos globales de alta calidad y actualizados.
- Amplia gama de funcionalidades geoespaciales.

- Integración con otras API de Google.

Desventajas:

- Costos asociados para un uso intensivo.
- Requiere una clave de API.
- Políticas de uso estrictas.

Bing Maps API:

Bing Maps API es un servicio de geocodificación proporcionado por Microsoft como parte de su plataforma de mapas, Bing Maps.

Ventajas:

- Integración con otras soluciones de Microsoft.
- Datos globales y actualizados.
- Documentación detallada y soporte.

Desventajas:

- Costos variables según el uso.
- Requiere una clave de API.
- Menos popular que Google en algunas regiones.

SitiData Geocoding API:

Descripción: SitiData Geocoding API es un servicio de geocodificación centrado en América Latina que utiliza datos actualizados y de alta calidad en la región, desarrollado por Servinformación.

Ventajas:

- Enfoque en datos de América Latina.
- Datos actualizados y de alta calidad en la región.
- Diversas opciones de licencia.

Desventajas:

- Cobertura limitada fuera de América Latina.
- API de acceso público limitada en uso gratuito.
- Menos conocido que otros geocodificadores globales

Recopilación de Datos de Prueba y Selección de Puntos de Referencia

Para este análisis comparativo, buscamos conjuntos de datos con información diversa sobre ubicaciones, lo que permite una revisión manual sencilla para usar como punto de referencia. En este contexto, elegimos el conjunto de datos de las Direcciones Regionales del ICBF, que obtuvimos de datos.gov.co. Este conjunto de datos ofrece información sobre ubicaciones departamentales en Colombia, incluyendo el municipio correspondiente, y la dirección de la respectiva Regional del ICBF.

Geocodificación de las Direcciones de Prueba

Para llevar a cabo la geocodificación de las direcciones de prueba, se utilizó la biblioteca Python Geopy. Geopy es una herramienta ampliamente utilizada para realizar operaciones geoespaciales, incluida la conversión de direcciones en coordenadas geográficas.

Se alistó el dataset de tal manera que se estableció una variable "dirección", que se establecía a partir de la suma de las tres variables que ya habían de esta manera:

A partir de esa variable, se genera una función de geocodificación con Geopy donde establecemos 'dirección' como parámetro. La función se comunica directamente con los servicios de geocodificación nombrados anteriormente, a excepción de Sitidata, a los cuales se puede acceder mediante API KEYS de acceso previamente establecidas y obtenidas.

Comparación con los Puntos de Referencia

Se establece una función Haversine que permite evaluar la distancia entre dos puntos a partir del radio de la Tierra, cómo parámetro de la función establecemos las dos coordenadas de entrada, retornando una distancia en KMS, para cada una de la Direcciones y para cada uno de los servicios.

Análisis de Resultados y Conclusiones

Se puede ver a partir de la función de haversine, los resultados de cercanías en promedio de cada uno de los servicios usados para la comparación

Capacidad de Respuesta

A partir del análisis y de las respuestas dadas según las peticiones a las API's. se puede determinar qué servicios retornan un valor sin importar cercanía. Se observa que los servicios de Google y Bing siempre proveen una respuesta, seguidos de Sitidata y por último Nominatim

Precisión

Se evidencia que de todos los registros (En total 33), Sitidata es el proveedor que mejor resultado demuestra a la hora de geocodificar las direcciones dadas, con un promedio de 108 Metros de Lejanía con respecto al punto de referencia, le siegue Google con 521 Metros de promedio en lejanía, en tercer lugar está el servicio de Bing, que provee una distancia de 1,302 Kilómetros y por último se encuentra el servicio basado en OpenStreetMaps (Nominatim), con una distancia promedio de más de 11 kilómetros, sin contar la cantidad de muestras georreferenciadas.

En la gráfica mostrada, se evidencia que los servicios que ofrecen una mejor respuesta son Google y Bing, dando siempre un retorno. Sin embargo, Sitidata a pesar de ofrecer un índice de 78% de

Respuesta sobre las direcciones solicitadas, genera una precisión mucho mayor con respecto a los demás, ofreciendo información más fiable en comparación.
Se descarta cualquier uso posible de Nominatim debido a obtener la tasa de respuesta más basa sobre las solicitudes, además de dar una precisión que se aleja y no es viable en ningún sentido.

Dicho lo anterior, para el modelamiento se decide quitar las variables geográficas ya que actualmente no se cuenta con dicha información de forma regular.

10.3. Anexo 3 – Text Mining

Se realizo análisis de N gramas para ver la respectiva distribución y escoger los mejores casos, los cuales se le presentaron a la empresa y con esto tomaron decisiones para próximos comunicados.

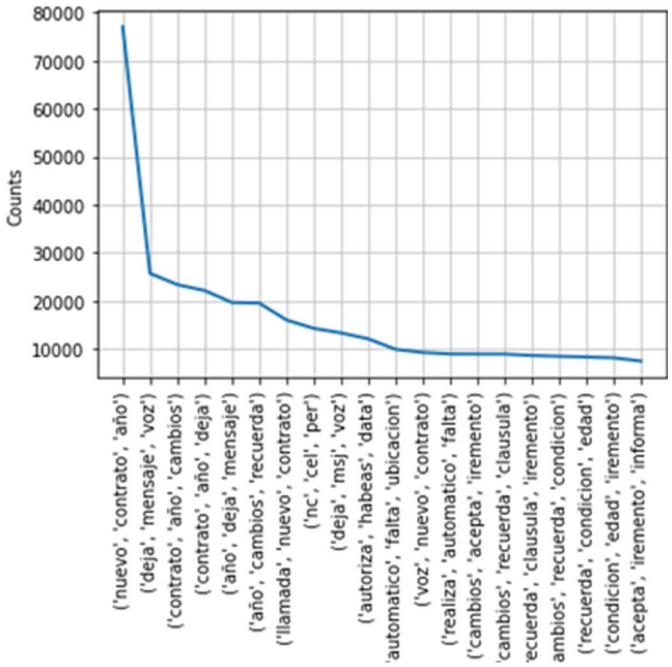


Ilustración 17. Análisis Trigramas [Autor, 2023]

A continuación, se observa el Power BI presentado a la empresa, donde se identificaba zonas donde tienen más usuarios y posibles sitios atacar, además de esto la respectiva nube de palabras, que se relaciona con la imagen anterior “Ilustración 17” y la respectiva distribución de sus servicios por estratos.

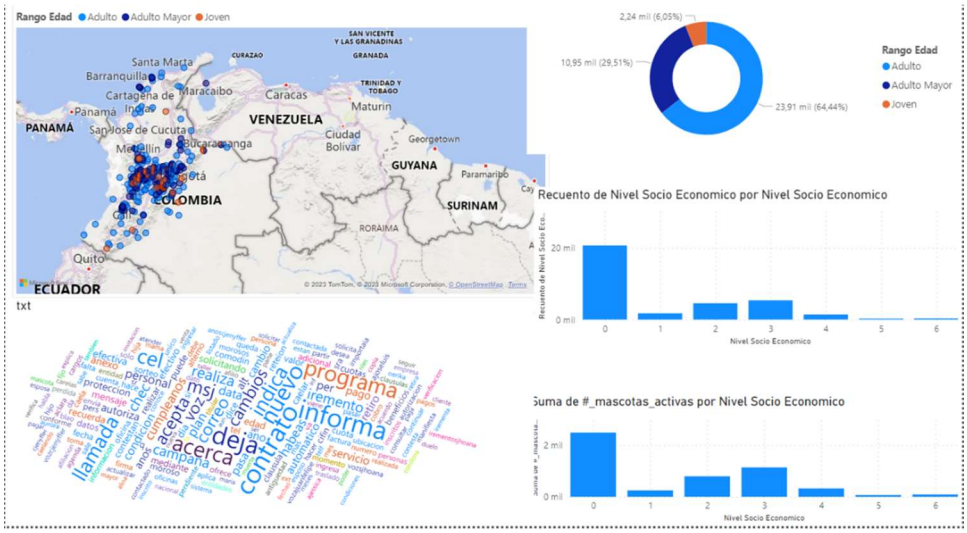


Ilustración 18. Análisis exploratorio de datos inicial [Autor, 2023]