Modelo de clasificación de riesgo para prevenir la deserción de los clientes que tienen activa una póliza de seguro en el segmento de Auto Individual en una compañía de seguros.

Camilo Andrés Barbosa Hernández
Ingeniero en Telemática

Trabajo de Grado
FACULTAD DE INGENIERÍA
MAESTRÍA EN ANALÍTICA APLICADA

Director:
Miguel Ángel Uribe Laverde
PhD.

Universidad de
La Sabana

PÁGINA DE ACEPTACIÓN

_____

Director Trabajo de Grado:
Miguel Ángel Uribe Laverde
PhD.

_____

Jurado Nro 1.

_____

Jurado Nro 2.

Bogotá D.C., junio de 2024

TABLA DE CONTENIDO

TABLA DE FIGURAS

LISTA DE TABLAS

1. RESUMEN

Una empresa en su portafolio de productos ofrece pólizas de seguros y una de sus líneas de negocio está orientada a ofrecer pólizas de autos individuales. El objetivo de este trabajo es entrenar un modelo supervisado de clasificación de clientes que permita identificar y predecir quiénes podrían continuar y quiénes no renovarían sus pólizas de seguro.

Para ello, se recopilaron y analizaron datos históricos, incluyendo información del cliente en el momento de la activación y fin de la vigencia de la póliza, información del vehículo asegurado, valores de prima y valores asegurados en el momento de la renovación y fin de la vigencia, Y con estos datos se entrenaron varios modelos de aprendizaje automático supervisado dónde el mejor modelo entrenado fue Gradient Boosting obteniendo un AUC de 0.71. Con esto, se generó un score dónde se clasificaron los clientes en cuatro niveles de riesgo: Muy alto, alto, medio y bajo según la probabilidad de no renovación.

Con base en esta clasificación, se recomienda a la empresa implementar campañas personalizadas de retención enfocadas en clientes clasificados de muy alto y alto riesgo, así como mejorar los incentivos y la comunicación con los clientes de riesgo medio y bajo.

En el futuro, la compañía podrá mejorar significativamente la retención de su base instalada de clientes, entendiendo mejor su comportamiento y maximizando la retención. Este proceso permitirá una estabilidad y crecimiento sostenido de la unidad de negocio de seguros de autos individuales.

Palabras claves: Pólizas de seguro, modelo supervisado de clasificación, score, clasificación.

2. ABSTRACT

The company, as part of its product portfolio, offers insurance policies, with one of its business lines focused on providing individual car insurance policies. The objective of this work is to train a supervised classification model for customers that allows identifying and predicting who might continue and who would not renew their insurance policies.

Historical data was collected and analyzed, including customer information at the time of policy activation and expiration, insured vehicle information, premium values, and insured values at the time of renewal and policy expiration. Using this data, several supervised machine learning models were trained, with Gradient Boosting emerging as the best model, achieving an AUC of 0.71. A scoring system was developed to classify customers into four risk levels: very high, high, medium, and low based on the probability of non-renewal.

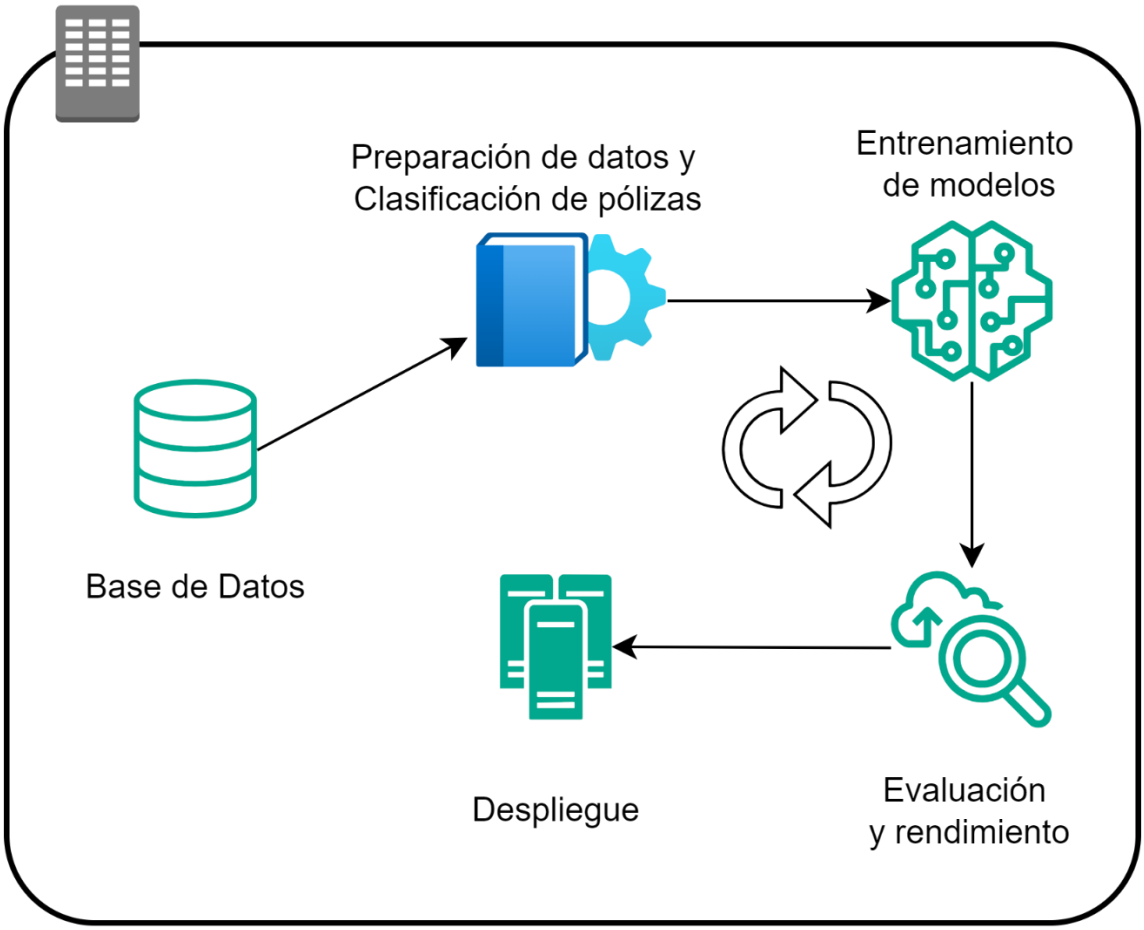
Based on this classification, the company is recommended to implement personalized retention campaigns targeting customers classified as very high and high risk, as well as improve incentives and communication with medium and low-risk customers.

In the future, the company will be able to significantly improve the retention of its installed customer base by better understanding their behavior and maximizing retention. This process will enable the stability and sustained growth of the individual car insurance business unit.

Keywords: Insurance policies, supervised classification model, score, classification.

3. RESUMEN GRÁFICO

Figura 1. Resumen gráfico



4. INTRODUCCIÓN

La retención de clientes es un desafío crítico para las empresas de seguros, la capacidad de identificar y predecir la probabilidad de que los clientes renueven o abandonen sus pólizas puede ayudar en reducir la tasa de abandono. El presente proyecto se basó en entrenar un modelo supervisado de clasificación para predecir la renovación de pólizas de seguro de autos individuales utilizando la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).

Se analizaron las variables y características más relevantes para entrenar un modelo de aprendizaje automático dónde se recopilaron datos históricos de los clientes que incluían información en el momento de activación de la póliza y fin de la vigencia de la póliza, información del vehículo asegurado, valores de prima, valores asegurados, información geográfica, score crediticio, entre otros.

Se plantearon preguntas de investigación que guiaron y dieron la ruta para el desarrollo del proyecto. Se investigaron técnicas de análisis de datos y de modelos de aprendizaje automático existentes que pudieran ser aplicados con el fin de predecir la probabilidad de abandono de un cliente. El modelo que mejor resultado obtuvo en el entrenamiento fue Gradient Boosting con un AUC de 0.71.

Finalmente, se exploraron los métodos y métricas más adecuados para validar y evaluar el rendimiento del modelo. El uso de la métrica AUC (Área Bajo la Curva) permitió una evaluación precisa de la capacidad para distinguir entre clientes que renovarán la póliza y aquellos que no.

5. PREGUNTA DE INVESTIGACIÓN

Para este proyecto fue necesario plantear una serie de preguntas que fueron la base para el desarrollo del proyecto.

- ¿Qué técnicas de análisis de datos y modelos de aprendizaje automático existen que se puedan utilizar para identificar patrones de deserción de clientes en el segmento de seguros de autos individuales y predecir la probabilidad de abandono?

- ¿Qué variables y características son más relevantes para entrenar un modelo de aprendizaje automático que pueda identificar los factores de riesgo que llevan a la deserción de los clientes?

- ¿Qué métodos y métricas son más apropiados para validar y evaluar el rendimiento de los modelos de pérdida de clientes en la compañía de seguros para el segmento de auto Individual?

6. MARCO CONCEPTUAL

6.1. Marco Teórico.

6.1.1. Análisis de datos.

El análisis de datos es utilizado para identificar patrones y tendencias del comportamiento de los productos y/o sus clientes. Las empresas, los gobiernos y las organizaciones recopilan datos y son una oportunidad para entender la situación actual y actuar de forma oportuna para situaciones futuras, por ello los datos son utilizados para prepararse hacia lo desconocido(Hush, 2020). Al aplicar técnicas estadísticas y de minería de datos, es posible explorar los conjuntos de datos para descubrir información relevante y comprender las causas de la pérdida de clientes.

6.1.2. Aprendizaje automático.

El aprendizaje automático, una rama de la inteligencia artificial que ofrece herramientas y técnicas para desarrollar modelos predictivos y analíticos a partir de los datos. Al utilizar algoritmos de aprendizaje es posible construir modelos que puedan predecir la probabilidad de ocurrencia de un conjunto de datos.

6.1.3. Modelo Predictivo.

El análisis predictivo utiliza datos históricos para predecir eventos futuros. Los datos históricos se utilizan para entrenar modelos matemáticos que capturan las tendencias importantes. Los modelos predictivos toman estos datos históricos para predecir lo que pasará o bien para sugerir acciones a implementar con el fin de obtener resultados óptimos(Mathworks, n.d.). En los últimos años este tipo de análisis ha tenido mucha acogida y gracias a los avances tecnológicos que permiten almacenar y procesar grandes volúmenes de datos se ha logrado generar modelos de aprendizaje matemáticos para identificar diferentes patrones de comportamiento.

6.1.4. Machine Learning.

Machine Learning es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y técnicas que permiten a las computadoras aprender y mejorar automáticamente a partir de la experiencia sin ser programadas explícitamente para ello. Las técnicas de machine learning son necesarias para mejorar la precisión de los modelos predictivos. Dependiendo de la naturaleza del problema que se está atendiendo, existen diferentes enfoques basados en el tipo y volumen de los datos y las principales técnicas de aprendizaje son supervisado, no supervisado, por refuerzo y deep learning. (IBM, n.d.)

6.1.5. Modelos de Riesgo.

Los modelos de riesgo son herramientas analíticas utilizadas para evaluar la probabilidad y el impacto de diferentes eventos de riesgo. En la industria de seguros, los modelos de riesgo son fundamentales en la evaluación y cuantificación de los riesgos asociados a las pólizas para clasificar a los clientes con conducta histórica que ocasionaría en la no renovación de su póliza.

6.1.6. Algoritmos supervisados de clasificación.

Los algoritmos supervisados forman parte de una de las ramas de Machine Learning. Su principal objetivo es entrenar modelos utilizando un conjunto de datos que poseen características específicas, dónde, la clase que se desea predecir es conocida. Durante el proceso de entrenamiento, el algoritmo adquiere conocimiento sobre las características de los datos para luego, con una nueva muestra de datos que comparte las mismas características pero no son conocidas por el modelo, el modelo entrenado debe estar en la capacidad de predecir la clase correspondiente.

Estos algoritmos se entrenan utilizando una variable dependiente (Y), que representa la variable a predecir, junto con un conjunto de datos con una o más

variables independientes (X), que capturan las características fundamentales de los datos.

6.1.6.1. Regresión Lineal.

Es un modelo estadístico que busca comprender la relación entre una variable dependiente y una o más variables independientes al ajustar una línea recta a los datos. El objetivo es minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos, lo que permite predecir valores continuos basados en la relación lineal entre las variables.

6.1.6.2. Random Forest.

El modelo de Random Forest es un método de aprendizaje que es utilizado principalmente para tareas de clasificación y regresión. Consiste en construir múltiples árboles de decisión y combinarlos para obtener predicciones más precisas y estables.

6.1.6.3. Gradient Boosting.

Es una técnica de aprendizaje automático que construye un modelo predictivo mediante la combinación de múltiples modelos simples, dónde utiliza árboles de decisión débiles, en un conjunto secuencial. Cada árbol se ajusta para corregir los errores del modelo anterior, lo que mejora gradualmente la precisión de las predicciones. El proceso de entrenamiento se realiza de forma iterativa, enfocándose en las instancias mal clasificadas o con residuos más altos en cada iteración, lo que resulta en un modelo final robusto y altamente preciso.

6.1.7. Técnicas para encontrar mejores hiperparámetros.

Cuando se entrena un modelo de machine learning, existen varias características que se pueden ajustar o cambiar en comparación a las definiciones del modelo, y estas características se modifican para poder obtener el mejor rendimiento y datos estadísticos del modelo, pero debido a que existen una gran cantidad de condiciones que se pueden ajustar, se hace complejo poder establecer cuál es la configuración más adecuada para el modelo, para ello existen técnicas de configuración para encontrar los mejores hiperparámetros que se adapten al modelo.

6.1.7.1. Grid Search.

Esta técnica implica definir una cuadrícula de hiperparámetros y probar sistemáticamente todas las combinaciones posibles dentro de esa cuadrícula. Se evalúa el rendimiento del modelo utilizando validación cruzada y se selecciona la combinación de hiperparámetros que maximiza la métrica de evaluación especificada.

6.1.7.2. Random Search.

En lugar de probar todas las combinaciones posibles de hiperparámetros, la búsqueda aleatoria selecciona aleatoriamente un conjunto de combinaciones de hiperparámetros para evaluar. Aunque puede perder algunas combinaciones potencialmente buenas, la búsqueda aleatoria es más eficiente computacionalmente y puede ser más efectiva para espacios de hiperparámetros de alta dimensionalidad.

6.1.7.3. Búsqueda Bayesiana.

Esta técnica utiliza el proceso de optimización bayesiana para encontrar la mejor combinación de hiperparámetros. A través de un proceso iterativo, la búsqueda bayesiana aprende de las iteraciones anteriores para guiar la búsqueda hacia las regiones del espacio de hiperparámetros que tienen un alto potencial de mejorar el rendimiento del modelo.

6.1.7.4. Matriz de Confusión.

La matriz de confusión es una tabla que se utiliza para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba para los cuales se conocen los valores reales. Esta tabla permite comparar los valores predichos por el modelo con los valores reales y consta de cuatro componentes:

- TP (True Positives): Verdaderos positivos, casos correctamente predichos como positivos.
- TN (True Negatives): Verdaderos negativos, casos correctamente predichos como negativos.
- FP (False Positives): Falsos positivos, casos incorrectamente predichos como positivos (error tipo I).
- FN (False Negatives): Falsos negativos, casos incorrectamente predichos como negativos (error tipo II).

6.1.7.5. Métricas Derivadas de la Matriz de Confusión:

Exactitud (Accuracy): Mide la proporción total de predicciones correctas.

$$Exactitud = \frac{TP + TN}{\mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precisión (Precision): Indica cuántos de los casos predichos como positivos son realmente positivos.

$$\text{Precisión} = \frac{TP}{\text{TP} + \text{FP}}$$

Sensibilidad (Recall o Tasa de Verdaderos Positivos): Indica la capacidad del modelo para identificar correctamente los casos positivos.

$$\text{Sensibilidad} = \frac{TP}{\text{TP} + \text{FN}}$$

Especificidad (Specificity o Tasa de Verdaderos Negativos): Indica la capacidad del modelo para identificar correctamente los casos negativos.

$$Especificidad = \frac{TN}{\text{TN} + \text{FP}}$$

F1 Score: Proporciona un balance entre precisión y sensibilidad, siendo útil cuando hay un desbalance entre las clases positivas y negativas.

$$F1\ Score = 2\ \frac{Presición * Exactitud}{\text{Presición} + \text{Exactitud}}$$

6.2. Estado del arte.

Para llevar a cabo la implementación de este proyecto se ha realizado una búsqueda de artículos investigativos, tesis, documentos, etc., que permitan tener mayor claridad en el tema a desarrollar y la mejor formar de abordarlo, dónde se han dividido en 3 aspectos principales:

6.2.1. Análisis de datos.

Nos ayudará a definir y clasificar las variables que utilizaremos en el modelo para su aprendizaje.

6.2.2. Modelos de Datos.

Tipos de modelos que podríamos implementar o técnicas de selección según las clasificaciones que se adaptarían al tema de investigación.

6.2.3. Segmentación de Clientes.

Una vez establecido el modelo se debe establecer clasificaciones para los clientes que presenten diferentes grados de riesgo en segmentación creada por el modelo.

6.2.4. Referencias y artículos de investigación relevantes.

A continuación, se presentan los artículos que se consideran más relevantes dentro de la investigación, sin embargo, no todos los artículos están basados en pólizas de seguros, pero son una guía fundamental para tener un enfoque de la implementación a realizar sin centrarnos en el sector que fueron aplicados.

Tabla 1. Artículos de investigación.

Título del Artículo	Referencia
An introduction to machine learning for classification and prediction. Family Practice	(Black et al., 2023)
Estrategias para el análisis de datos cualitativos. Universidad de Buenos Aires.	(Borda et al., 2020)
Fidelización y rentabilización de usuarios de seguros todo riesgo de vehículos por medio de la venta cruzada y la venta escalonada. Un enfoque promocional para la industria aseguradora.	(Contreras Serrano, 2016)
Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association	(Kruskal & Wallis, 1952)
Deep learning. Nature,	(Lecun et al., 2015)
Optimization of Experimental Teaching System based on ACSI Model	(Liu et al., 2022)

Risk Identification Using Quantum Machine Learning for Fleet Insurance Premium. Communications in Computer and Information Science	(Naik & Bhise, 2022)
Machine Learning Logistic Regression Model for Early Decision Making in Referral of Children with Cervical Lymphadenopathy Suspected of Lymphoma	(Zijtregtop et al., 2023)
A machine learning approach to identifying decision-making styles for managing customer relationships	(Tudoran, 2022)
Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa	(Cam Gensollen, 2022)

7. OBJETIVOS

7.1. Objetivo General.

Entrenar un modelo supervisado de clasificación para establecer el riesgo que presentaría un cliente en la no renovación de las pólizas de seguro de automóvil individuales y mejorar la capacidad de la Gerencia de Cartera en la compañía de seguros para retener a los clientes con riesgo de deserción.

7.2. Objetivos específicos.

- Recopilar y analizar datos históricos y actuales de clientes de las pólizas de seguro de automóvil individuales.

- Aplicar técnicas de análisis exploratorio de datos para identificar patrones y características significativas relacionadas con la pérdida de clientes en las pólizas de seguro de automóvil individuales.

- Entrenar modelos supervisados de aprendizaje automático para clasificar los clientes con riesgo de deserción.

8. METODOLOGÍA

Para cumplir con los objetivos planteados en este proyecto, se utilizó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), que es un modelo estándar utilizado en proyectos de minería de datos y análisis predictivo, dónde se establece la siguiente estructura por etapas para abordar la problemática.

- Recopilar y preparar los datos históricos y actuales de los clientes desde las fuentes de información que la compañía nos suministre.
- Realizar un análisis estadístico y exploratorio de datos para obtener una vista general de sus distribuciones y así poder identificar patrones y clasificar los diferentes tipos de variables.
- Explorar los posibles modelos de aprendizaje automático a implementar con muestras aleatorias para definir cuál es el mejor modelo con datos estadísticos.
- Desarrollar el modelo de aprendizaje automático con los datos que se han utilizado en la etapa exploratoria.
- Validar y evaluar el rendimiento del modelo.
- Proponer un plan de implementación del modelo.

La metodología CRISP-DM se utilizó como guía principal para llevar a cabo este proyecto, abordando cada una de sus etapas como se detalla en la Figura 2.

Figura 2. CRISP-DM.

Fuente: The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13—22

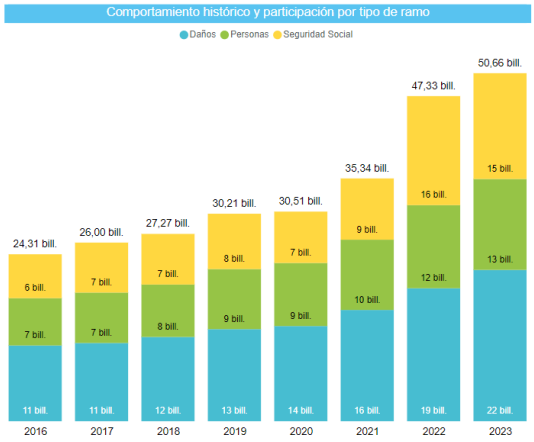
8.1. Comprensión del Negocio.

En Colombia existen varias empresas encargadas de emitir pólizas de seguros y actualmente dichas empresas cuentan con un gran portafolio de productos para las diferentes necesidades de las personas y/o empresas que deseen proteger sus bienes o servicios. Estas empresas están actualmente agremiadas en una entidad sin ánimo de lucro denominada Federación de Aseguradores Colombianos ó FASECOLDA.

En Colombia existe 3 tipos de ramos que agrupan los diferentes sectores o pólizas que se emiten en el país: Los daños, las personas y la seguridad social. Como se observa en la Figura 3, para el año 2023 hubo un incremento en la cantidad de primas emitidas, pasando de $ 47.33 billones de pesos a $ 50.66 billones de pesos, lo que representa un aumento de 7.05% en comparación al año inmediatamente anterior y el ramo que tuvo el mayor incremento fue el de daños que agrupa a las

pólizas que protegen a las personas y empresas contra daños materiales o pérdidas financieras.

Figura 3. Primas emitidas en billones de pesos colombianos

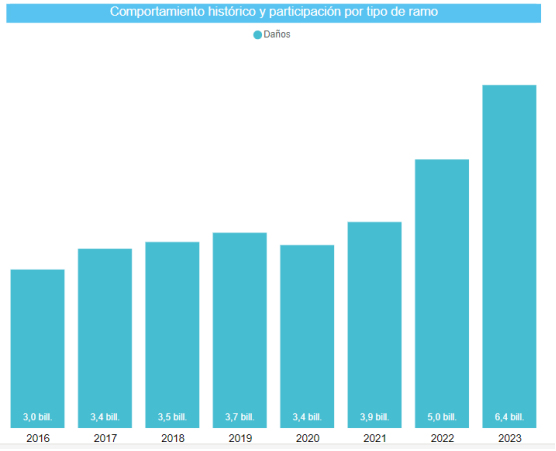


Fuente: Fasecolda, diciembre de 2023

El sector o grupo de pólizas de daños se subdivide en varios sectores, y para efectos de este proyecto nos enfocaremos en el sector de automóviles.

En la figura 4 se observa el comportamiento de los últimos años que ha tenido el sector de las pólizas de automóviles respecto a las primas emitidas, pasando de $5.0 billones de pesos colombianos en 2022, a $6.4 billones de pesos colombianos en el año 2023, lo que representa un incremento del 28% en comparación al año anterior.

Figura 4. Primas emitidas para el sector de Automóviles.



Fuente: Fasecolda, diciembre de 2023

8.2. Comprensión de los datos.

Para la implementación de este proyecto, contamos con una base de datos de los clientes que pertenecen a una compañía aseguradora legalmente constituida en Colombia que por temas de seguridad empresarial no es posible mencionar.

La base de datos corresponde a todos los clientes que han tenido activa una póliza de seguros para el sector de daños sobre el segmento de auto individual que representa a todos los vehículos y camionetas del sector privado. Este tipo de póliza protege a un solo vehículo y a su conductor principal en caso de ocurrencia de algún siniestro.

Para efectos del proyecto, la base de datos de las pólizas que fueron suministradas por la compañía comprende dos periodos anuales con corte mensual.

El data set inicial cuenta con un total de 105 variables que hemos agrupado según sus principales características:

- Información de la póliza activada en la compañía.
- Información de geolocalización de la persona que activó la póliza.
- Información del vehículo asegurado.
- Información de primas y valores asegurado en el momento de la activación de la póliza.
- Información de primas y valores asegurados para la renovación de la póliza.

8.3. Preparación de los datos.

Una vez establecido el data set inicial, se realizaron validaciones y transformaciones sobre los datos. Para este proyecto la preparación fue la siguiente:

- Limpieza de los datos: Se analizaron las variables para detectar valores atípicos y nulos que podrían hacer ruido en el momento de la implementación del modelo.

- Transformación de variables: El data set cuenta con variables numéricas y categóricas, pero se realizaron ajustes sobre algunas variables numéricas que de acuerdo con su contenido podrían ser transformadas a categóricas.

- Creación de nuevas variables: Se analizó la posibilidad de crear nuevas variables a partir de las existentes con el fin de ajustar los datos que va a utilizar el modelo para ser entrenado.

8.4. Selección del Modelo.

Una vez realizado el análisis exploratorio de los datos, observamos que nos enfrentamos a un problema computacional supervisado de clasificación y para establecer la probabilidad de ocurrencia que podrían tener los clientes en el

momento de la no renovación de la póliza de seguros en la compañía, se estableció entrenar los modelos de Regresión Logística, Random Forest, Gradient Boosting para establecer cual tendría el mejor performance estadístico de acuerdo a los datos del data set inicial.

8.5. Evaluación del Modelo.

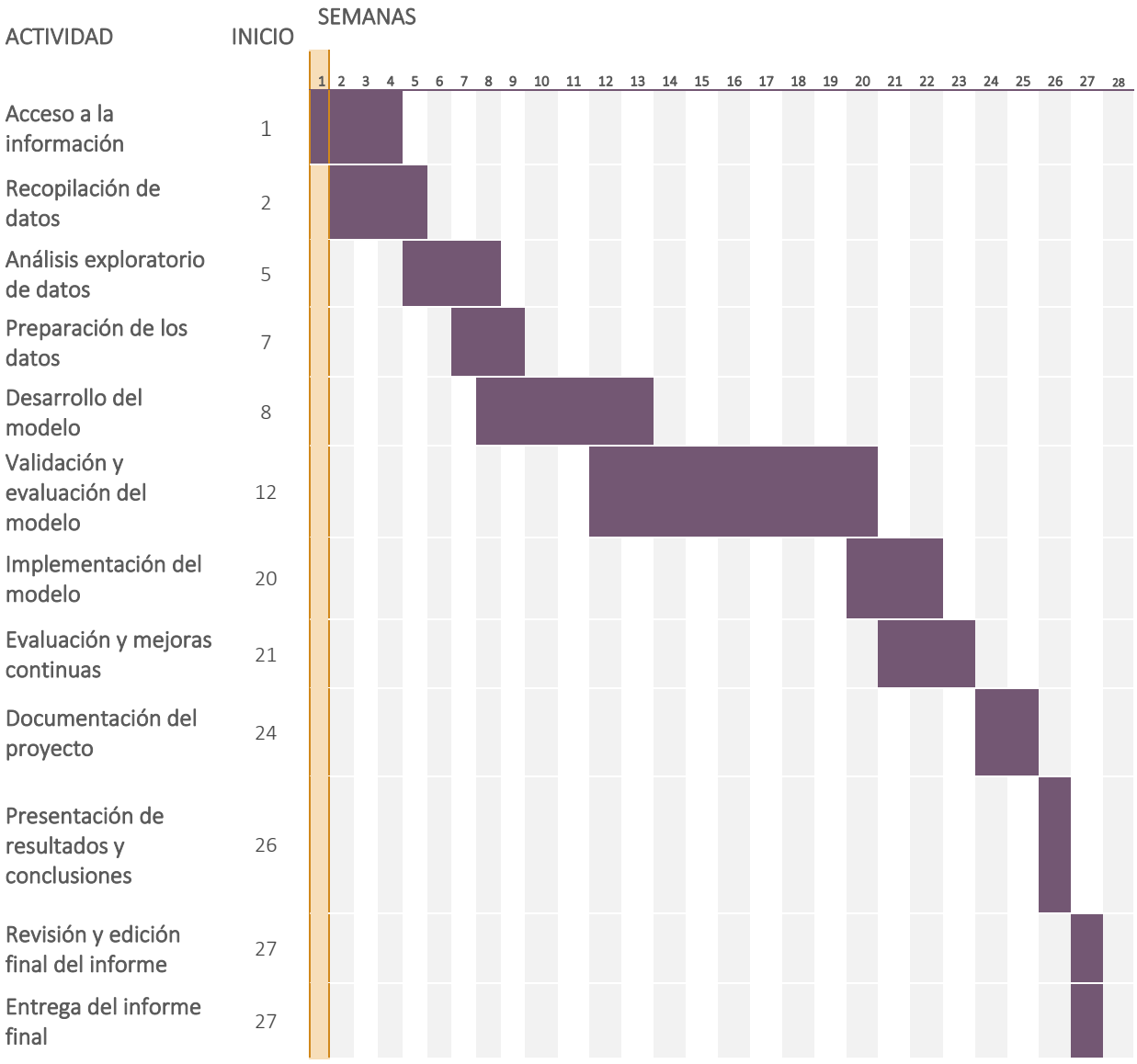
Se validó con la compañía el performance de cada uno de los modelos para establecer cuál podría ser el modelo óptimo de acuerdo con los datos iniciales tomando como referencia los estadísticos obtenidos por cada modelo.

8.6. Despliegue.

Una vez entrenado y seleccionado el modelo se procedió a implementar el modelo en la compañía.

9. CRONOGRAMA DE ACTIVIDADES

En el siguiente diagrama se presentan las diferentes actividades que fueron ejecutadas en el desarrollo del proyecto con una duración total de 27 semanas.



10. RESULTADOS DE LOS EXPERIMENTOS

En la presente sección, abordamos el desarrollo del proyecto mediante la descripción detallada de las actividades realizadas para su implementación. Se llevó a cabo un análisis exhaustivo y exploratorio de los datos recopilados, seguido de un análisis descriptivo para comprender mejor su naturaleza y distribución. Se procedió a la selección cuidadosa de las variables relevantes para la construcción del modelo, considerando su impacto en los resultados finales. Posteriormente, se llevó a cabo el entrenamiento de diversos modelos estadísticos con el fin de obtener métricas y evaluar su rendimiento. Este proceso permitió determinar cuál de los modelos se adapta de manera óptima a los datos recopilados, facilitando así su implementación efectiva.

10.1. Análisis exploratorio de los datos.

El objetivo de este análisis se basó en entender la composición y las características de la base de datos de clientes que tienen activa una póliza de seguro. Se realizaron diversas técnicas exploratorias para obtener información significativa sobre los clientes y las pólizas de seguro asociadas. A continuación, se resumen los hallazgos más relevantes.

En total, la base de datos inicial utilizada en este proyecto contenía 105 variables y 326,644 registros correspondientes a pólizas activas que finalizan su vigencia mes a mes con un total de 23 meses.

Para realizar el análisis exploratorio de los datos, se realizó la validación de cada una de las variables dónde se observó su distribución y comportamiento frente a la renovación o no renovación de la póliza. Una vez realizado este análisis

exploratorio, se ha reducido la cantidad de variables a utilizar dentro del modelo pasando de 105 variables a una base reducida de 48 variables.

El primer data set reducido está compuesto por una (1) variable tipo fecha, doce (12) variables categóricas y treinta y cinco (35) variables numéricas cómo se referencia en la tabla 2.

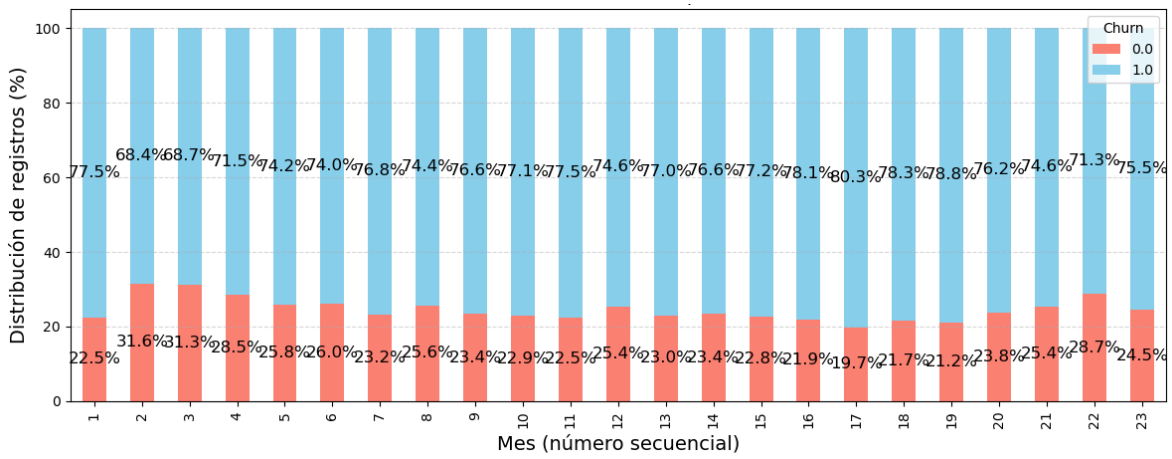
Tabla 2. Dataset reducido.

Nombre Variable	Descripción	Tipo
Cantidad_Renovaciones_Poliza	Cantidad de renovaciones realizadas.	Numérico
Credit_Score_Actual	Score crediticio al activar la póliza	Numérico
Credit_Score_Renovacion	Score crediticio en el momento de la renovación	Numérico
Descuentos_comerciales	Descuentos realizados en el pago de la póliza	Numérico
Edad_Vehiculo	Edad del vehículo	Numérico
Fecha_Corte	Fecha mensual de corte de los datos	Fecha
Id_Ramo	Identificador del tipo de ramo de la póliza	Numérico
Id_Renovacion	Si renovó o no la póliza	Numérico
Puntaje_Actual_Aseguradora	Puntaje actual otorgado por la aseguradora	Numérico
Puntaje_Historico_Aguradora	Puntaje otorgado por la aseguradora de periodos anteriores.	Numérico
Puntaje_Renovación_Aseguradora	Puntaje otorgado por la aseguradora en el momento de la renovación	Numérico
Tipo_Documento	Tipo de documento de la persona que activó la póliza.	Categórico
Tipo_Pago	Periodo acordado para el pago de la póliza.	Numérico
Tipo_Poliza	Descripción del tipo de póliza	Categórico
Tipo_Servicio	Descripción del servicio actual utilizado por el vehículo.	Categórico
Tipo_Vehiculo	Descripción del tipo de vehículo	Categórico
Total_Otras_Polizas_Vigentes	Total de pólizas activas	Numérico
Total_Siniestros_Actual	Total de siniestros registras en la vigencia de la póliza actual	Numérico
Valor_Asegurado_Actual	Valor asegurado al activar la póliza	Numérico
Valor_Asegurado_Renovacion	Valor para asegurar en el momento de la renovación	Numérico
Valor_Prima_Actual	Valor de la prima actual	Numérico
Valor_Prima_Renovación	Valor de la prima en el momento de la renovación	Numérico
Edad_Usuario	Edad del usuario que activó la póliza	Numérico
Sucursal_Venta	Sucursal de venta de la póliza	Numérico
Departamento_Movilidad	Departamento de movilidad que activó la póliza	Categórico
Clase_Vehiculo	Clase de vehículo	Categórico

Cilindraje_Vehiculo	Cilindraje de vehículo	Numérico
Caballos_Fuerza_Vehiculo	Caballos de Fuerza de vehículo	Numérico
Tipo_Combustible_Vehiculo	Tipo de Combustible de vehículo	Categórico
Peso_Vehiculo	Peso de vehículo	Numérico
Pasajeros_Vehiculo	Pasajeros de vehículo	Numérico
Puertas_Vehiculo	Puertas de vehículo	Numérico
Capacidad_Carga_Vehiculo	Capacidad de Carga de vehículo	Numérico
Tipo_Caja_Vehiculo	Tipo de Caja de vehículo	Categórico
Valor_por_Partes_Vehiculo	Valor por Partes del vehículo	Numérico
Aireacondicionado_Vehiculo	Aire acondicionado de vehículo	Numérico
Tipo_Transmision_Vehiculo	Tipo de transmisión de vehículo	Categórico
Pesocategoria_Vehiculo	Peso de la categoría de vehículo	Categórico
Ultimomodelo_Vehiculo	clasificación de Ultimo modelo de vehículo	Numérico
Departamento_Usuario_Central_Riesgo	Departamento del usuario que activó la póliza en centrales del riesgo	Categórico
Genero_Usuario	Genero de usuario	Categórico
Ultimomodelo_Vehiculo_Renovacion	Clasificación de último modelo de vehículo en el momento de la renovación	Numérico
Valor_Prima_Pura_Riesgo	Valor de la prima de riesgo	Numérico
Valor_Prima_Tecnica	Valor de la prima de técnica	Numérico
Valor_Prima_Cobrada	Valor de la prima de cobrada	Numérico
Valor_Prima_Mercado	Valor de la prima de mercado	Numérico
Valor_Prima_Mercado_Prediccion	Valor de la prima de predicción	Numérico
Valor_Prima_Mercado_Prediccion_DESV	Valor de la prima en mercado con desviación estándar	Numérico

Luego, de haber definido el conjunto de datos, se ha determinado que la variable objetivo a predecir es 'Id_Renovacion', la cual tiene dos valores posibles: si el cliente renueva la póliza para la próxima vigencia o no. En la figura 5 podemos observar que la distribución del churn o pólizas de vehículo que no son renovadas en promedio representan el 24.5% de los datos.

Figura 5. Distribución mensual de churn por fecha de corte.



10.2. Data Cleaning.

En esta sección, nos enfocaremos en analizar algunas de las variables del data set. Examinaremos las cantidades de registros por categoría y evaluaremos si se requiere filtrar o eliminar ciertos valores.

Comenzando con el "Tipo de Vehículo", como se observa en la tabla 3, notamos que el 99% de los datos corresponden a la categoría 2, mientras que hay dos valores adicionales que no son relevantes para el análisis presente. Por lo tanto, procederemos a filtrar estos valores para quedarnos únicamente con el más representativo. Posteriormente, dado que esta variable tendrá un único valor después del filtrado, no será necesario incluirla en el modelo.

Tabla 3. Variable: Tipo de Vehículo

Descripción	Conteo	% Participación
Categoría 1	14	0.004
Categoría 2	326,628	99.995
Categoría 3	2	0.001

Nota: Las categorías de la variable han sido anonimizadas por solicitud de la empresa.

La variable "Tipo de Servicio" indica el tipo de uso que tiene el automóvil y sobre esta variable como podemos observar en la tabla 4, notamos que existe un pequeño porcentaje de datos asociados a la categoría 2, representando sólo el 0.23% del total de los datos. Por lo tanto, procederemos a filtrar este valor para mantener únicamente el más representativo. Tras aplicar este filtro, la variable ya no será necesaria incluirla en el modelo, dado que tendrá solo un valor.

Tabla 4. Variable: Tipo de Servicio

Descripción	Conteo	% Participación
Categoría 1	325,887	99.77
Categoría 2	753	0.23

Nota: Las categorías de la variable han sido anonimizadas por solicitud de la empresa.

La variable "Tipo de Identificación" indica los tipos de documento de las personas que activaron la póliza. En la tabla 5 observamos que hay 11 registros marcados en la categoría 3, lo que representa menos del 0.01% de los datos. Por ende, procederemos a excluir estos registros de los datos.

Tabla 5. Variable: Tipo de Identificación

Descripción	Conteo	% Participación
Categoría 1	317,289	97.14
Categoría 2	8,459	2.59
Categoría 3	11	0.00
Categoría 4	885	0.27

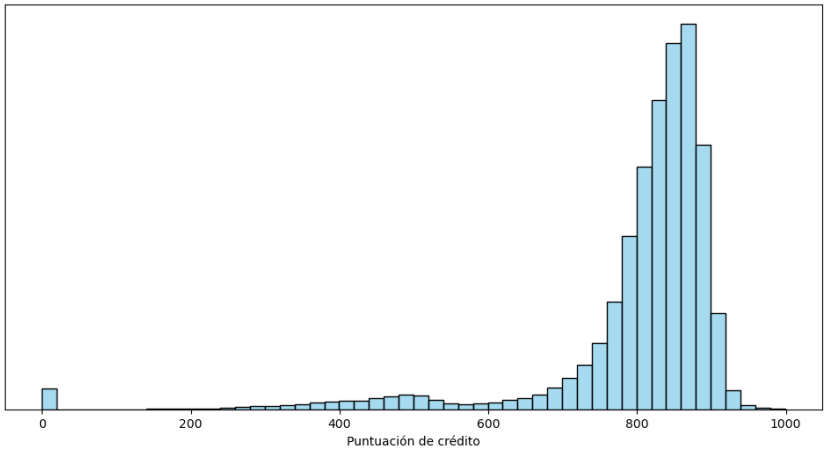
Nota: Las categorías de la variable han sido anonimizadas por solicitud de la empresa.

10.3. Data Transformation.

En esta sección, analizaremos la necesidad de realizar modificaciones o ajustes en el tipo de variables para el entrenamiento del modelo. Esto contribuirá a mejorar la precisión y eficiencia durante el proceso de entrenamiento, así como facilitará la búsqueda eficiente de hiperparámetros.

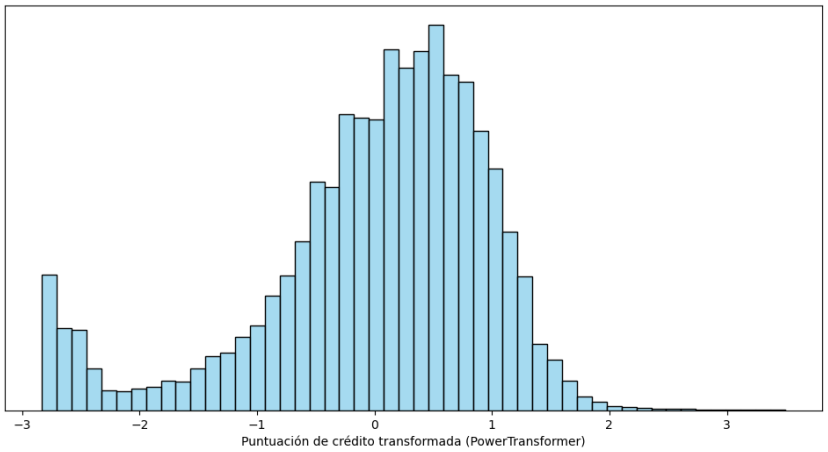
Cómo se observa en la Figura 6, la variable "Score de Crédito" muestra un sesgo significativo hacia la izquierda. Dado que presenta una amplia gama de valores únicos en la base de datos, el modelo podría enfrentar desafíos al realizar predicciones. Por esta razón, se ha decidido aplicar la transformación PowerTransformer. Esta técnica de preprocesamiento de datos se utiliza para normalizar variables que no siguen una distribución normal, lo que puede mejorar la capacidad del modelo para realizar predicciones precisas y el resultado de la transformación de esta variable se puede observar en la figura 7.

Figura 6. Distribución de registros por puntuación de Credit Score



Nota: La escala vertical ha sido removida por solicitud de la empresa.

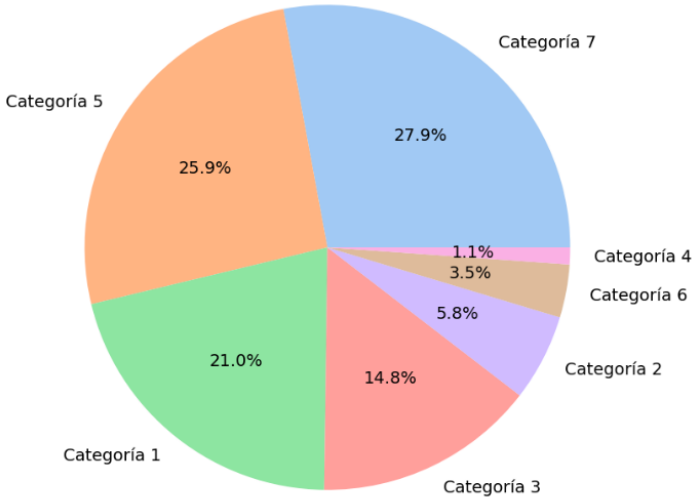
Figura 7. Distribución PowerTrasnformer sobre la variable Credit Score



Nota: La escala vertical ha sido removida por solicitud de la empresa.

La variable "Edad del Vehículo" representa la antigüedad del vehículo asegurado. Sin embargo, en la exploración inicial de la variable, observamos que existen vehículos asegurados que representan menos del 1.1% de los datos. Por esta razón, hemos generado grupos por rangos de edad los cuales se pueden visualizar en la figura 8.

Figura 8. Porcentaje de participación por Edad del Vehículo



Nota: Las categorías de la variable han sido anonimizadas por solicitud de la empresa.

Durante el análisis de los datos, en otras variables se identificaron valores nulos en el conjunto de datos. Se evaluó si estos valores deberían ser o no eliminados en relación con la cantidad nulos en comparación con el tamaño total del conjunto de datos, dónde encontramos que representan un pequeño porcentaje de los datos cómo se observa en la tabla 7, por lo que su eliminación no alteraría significativamente la integridad del conjunto de datos en el momento del entrenamiento del modelo.

Tabla 6. Conteo valores nulos

Variable	Valores Nulos
Credit_Score_Actual	1,471
Credit_Score_Renovacion	864
Puntaje_Renovación_Aseguradora	8
Puntaje_Actual_Aseguradora	5
Valor_Asegurado_Actual	5
Descuentos_comerciales	4
Total_Siniestros_Actual	3
Valor_Prima_Actual	3

Para las variables numéricas que contienen información con datos de moneda o presentan valores muy dispares en su contenido, se aplicó la siguiente transformación:

Variables: Valor_Asegurado_Actual y Valor_Asegurado_Renovación.

- Creación de una nueva variable del porcentaje de cambio relativo.
- Se excluye del modelo la variable Valor_Asegurado_Actual.

Variables: Valor_Prima_Actual y Valor_Prima_Renovacion.

- Creación de una nueva variable del porcentaje de cambio relativo.

- Se excluye del modelo la variable Valor_Prima_Actual.

Variables: Credit_Score_Actual y Credit_Score_Renovacion.

- Creación de una nueva variable del porcentaje de cambio relativo.
- Se excluye del modelo la variable Credit_Score_Actual.

Luego de haber realizado estas trasformaciones en los datos, las variables finales a utilizar en el modelo son 43:

Tabla 7. Data set final para entrenar el modelo

	Nombre Variable	Tipo
1	Cantidad_Renovaciones_Poliza	Categórica
2	Credit_Score_Renovacion	Categórica
3	Descuento_puntaje_actual	Categórica
4	Descuento_puntaje_renovacion	Categórica
5	Descuentos_comerciales	Categórica
6	Edad_Vehiculo_GRUPO	Categórica
7	Id_Renovacion	Numérica
8	Porcentaje_Diferencia_Credit_Score	Numérica
9	Porcentaje_Diferencia_Valor_Asegurado	Numérica
10	Porcentaje_Diferencia_Valor_Prima_actual	Numérica
11	Puntaje_renovacion	Categórica
12	Tipo_Documento	Categórica
13	Tipo_Pago	Categórica
14	Total_Otras_Polizas_Vigentes_GRUPO	Categórica
15	Total_Siniestros_Actual	Categórica
16	Valor_Asegurado_Renovacion	Numérica
17	Valor_Prima_Renovacion	Numérica
18	Cilindraje_Vehiculo	Numérica
19	Caballos_Fuerza_Vehiculo	Numérica
20	Peso_Vehiculo	Numérica
21	Capacidad_Carga_Vehiculo	Numérica
22	Valor_por_Partes_Vehiculo	Numérica
23	Pesocategoria_Vehiculo	Numérica
24	Ultimomodelo_Vehiculo	Numérica
25	Edad_Usuario	Numérica
26	Ultimomodelo_Vehiculo_Renovacion	Numérica

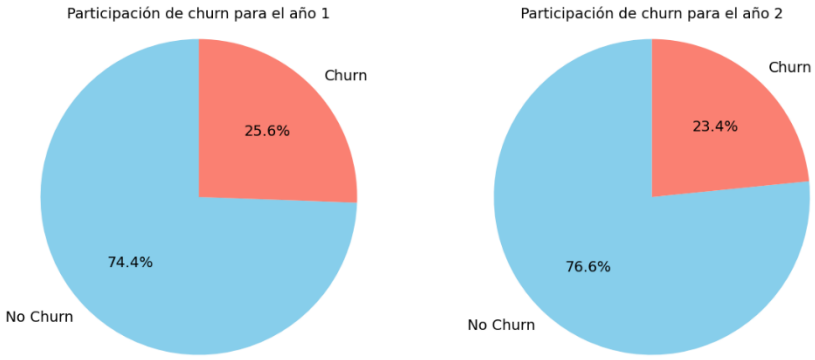
27	Valor_Prima_Pura_Riesgo	Numérica
28	Valor_Prima_Tecnica	Numérica
29	Valor_Prima_Cobrada	Numérica
30	Valor_Prima_Mercado	Numérica
31	Valor_Prima_Mercado_Prediccion	Numérica
32	Valor_Prima_Mercado_Prediccion_DESV	Numérica
33	Tipo_Poliza	Numérica
34	Sucursal_Venta	Categórica
35	Tipo_Combustible_Vehiculo	Categórica
36	Pasajeros_Vehiculo	Categórica
37	Puertas_Vehiculo	Categórica
38	Tipo_Caja_Vehiculo	Categórica
39	Aireacondicionado_Vehiculo	Categórica
40	Tipo_Transmision_Vehiculo	Categórica
41	Genero_Usuario	Categórica
42	Departamento_Grupo	Categórica
43	Clase_Vehiculo_Grupo	Categórica

10.4. Variable Objetivo.

La variable objetivo o variable dependiente denominada 'Id_Renovacion', representa los datos que se van a predecir dentro del modelo. En nuestro caso, esta variable está representada de forma binaria, dónde “0” corresponde a las pólizas que no fueron renovadas y “1” corresponde a las pólizas que sí fueron renovadas.

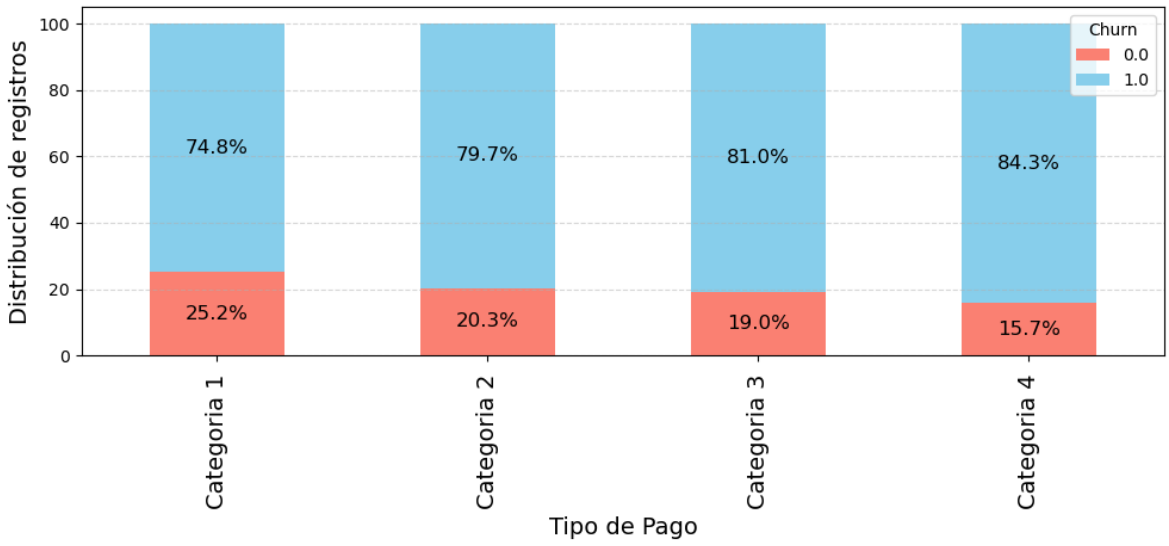
En los primeros 12 meses, la cantidad de pólizas que no fueron renovadas representa el 25.6% de los datos. Los siguientes 11 meses, este valor había disminuido a un 23.5% de los datos.

Figura 9. Porcentaje de participación por churn en los años 1 y 2



Los clientes pueden realizar el pago de la póliza en cuatro modalidades diferentes, en la figura 10, se observa que las pólizas con una mayor tasa de no renovación corresponden al tipo de pago anual, con un 45.8% de participación. Sin embargo, es importante destacar que este tipo de pago representa el porcentaje más significativo de participación en el total de pólizas según la modalidad de pago.

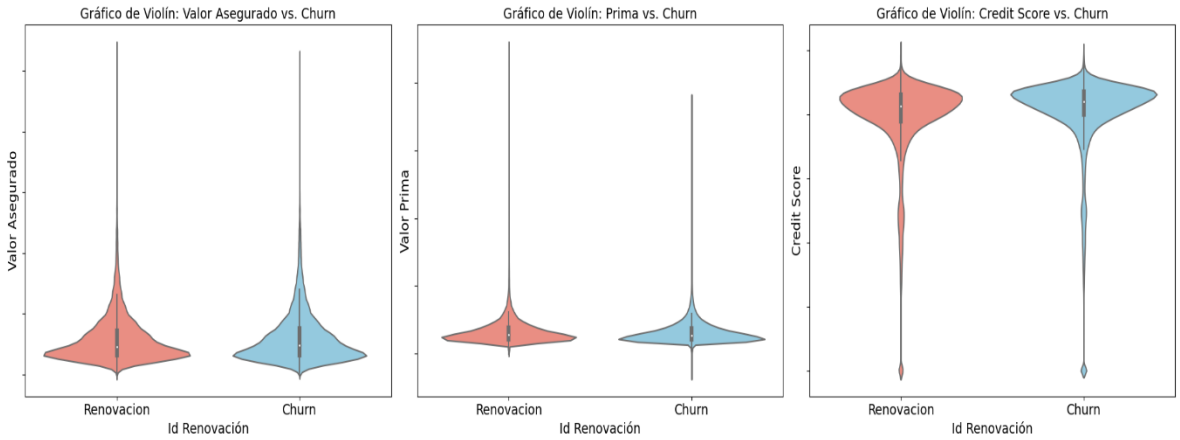
Figura 10. Distribución de churn por tipo de pago de la póliza



Nota: La escala vertical ha sido removida por solicitud de la empresa.

Cuando queremos entender cómo se comportan las variables numéricas en relación con la variable objetivo, el diagrama de tipo Violín nos muestra una distribución simétrica entre las pólizas renovadas y no renovadas. No obstante, es importante tener en cuenta que existen varios datos atípicos fuera de esta distribución. En la figura 11 podemos observar la distribución de algunas variables numéricas y su comportamiento respecto a la variable objetivo para comparar sobre la renovación o cancelación de la póliza.

Figura 11. Variable de churn en comparación a las variables numéricas.



Nota: La escala vertical ha sido removida por solicitud de la empresa.

10.5. Modelo Supervisado de Clasificación.

Una vez definido el conjunto de datos final, el siguiente paso fue buscar y entrenar el modelo que mejor se adaptara a los datos. En este caso, se identificó que se trataba de un problema de clasificación, por lo que se exploraron modelos de aprendizaje supervisado de clasificación.

Las variables independientes (X) representan los atributos o características utilizadas para realizar las predicciones. Por otro lado, la variable dependiente (y) corresponde a la clase o categoría que el modelo intentará predecir utilizando las características proporcionadas.

Se exploraron y entrenaron los modelos de Regresión Logística, Gradient Boosting y Random Forest, con el objetivo de determinar cuál de ellos ofrecía el mejor desempeño en términos de precisión y capacidad predictiva para el conjunto de datos analizado.

10.5.1. Preparación y Preprocesamiento de Datos.

El preprocesamiento de los datos es crucial para asegurar que estén en el formato adecuado y libres de anomalías que pudieran afectar el rendimiento de los modelos. Se definió un pipeline de preprocesamiento que incluyó:

- Power Transformation para la variable 'Credit_Score_Renovacion' para estabilizar la varianza y normalizar la distribución.

- Estandarización de las variables numéricas usando StandardScaler para garantizar que todas las características numéricas tuvieran una media de 0 y una desviación estándar de 1.

- Codificación One-Hot en las variables categóricas mediante OneHotEncoder para convertir las categorías en un formato numérico binario.

10.5.2. División de Datos.

Los datos se dividieron en conjuntos de entrenamiento y prueba con una proporción de 80% para los datos de Training y 20% para los datos de Test, asegurando la estratificación por la variable “y” para mantener la misma distribución de la variable objetivo en ambos conjuntos.

10.5.3. Búsqueda del Mejor Modelo.

Luego de definir las variables a utilizar en el modelo, se procedió a optimizar sus hiperparámetros mediante la técnica GridSearchCV de la biblioteca scikit-learn. Esta técnica realiza una búsqueda exhaustiva dentro de una grilla predefinida de valores posibles a utilizar para cada hiperparámetro del modelo. Para evaluar el desempeño de cada configuración de hiperparámetros, se utilizó la técnica de validación cruzada k-folds (cv=7) junto con el scoring AUC-ROC (Área Bajo la Curva de la Característica Operativa del Receptor), que permite encontrar el rendimiento del modelo para este tipo de clasificación binaria basado en la probabilidad de que una instancia pertenezca a la clase positiva o negativa. La búsqueda se llevó a cabo durante 1000 iteraciones para cada modelo, con los siguientes resultados en el grupo test:

Tabla 8. AUC-ROC en la búsqueda del mejor modelo con GridSearchCV.

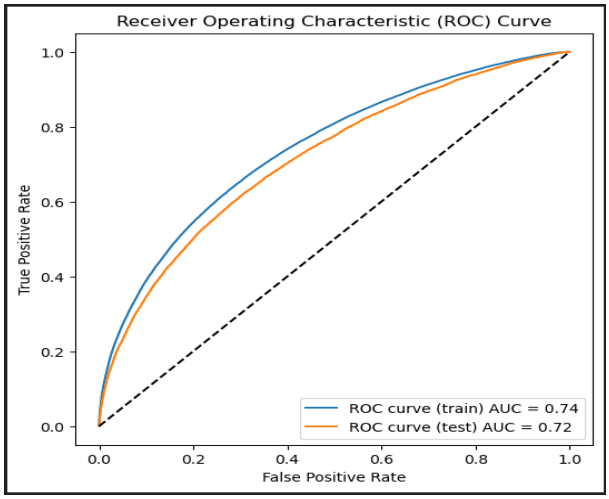
Modelo	AUC -ROC
Regresión Logística	0.6990
Random Forest	0.6912
Gradient Boosting Classifier	0.7155

Como se observa en la tabla 8, el modelo que obtuvo mejor rendimiento utilizando la puntuación de AUC-ROC fue Gradient Boosting, seguido de Random Forest. Esto nos permitió establecer que el modelo que mejor se adapta a los datos es el Gradient Boosting Classifier. Además, como se muestra en la figura 12, donde se presenta la curva ROC para los conjuntos de entrenamiento y prueba, se puede descartar la posibilidad de overfitting debido a la similitud o cercanía entre ambas curvas.

La elección de AUC-ROC para medir la capacidad del modelo se fundamenta en su solidez para proporcionar una medida robusta del performance sobre el modelo,

especialmente en contextos de clasificación binaria. Esta métrica evalúa la capacidad del modelo para distinguir entre las clases positivas y negativas, integrando tanto la tasa de verdaderos positivos (TP) como la tasa de falsos positivos (FP) a través de diferentes umbrales de decisión.

Figura 12. Curva AUC-ROC para Gradient Boosting Classifier



10.5.4. Hiperparámetros.

A continuación, se describen los hiperparámetros configurados para entrenar el modelo Gradient Boosting Classifier:

GradientBoostingClassifier(
n_estimators=200,
loss= "log_loss",
learning_rate=0.1,
max_depth=10,
min_samples_split=10,
min_samples_leaf=2,
random_state=42)

Dónde:

- *n_estimators*: El número de etapas de potenciación a realizar. En este caso, n_estimators=200 indica que el modelo construirá hasta 200 árboles de decisión secuenciales.

- *learning_rate*: Este parámetro reduce la contribución de cada árbol mediante un valor que ayuda a controlar el sobreajuste; un valor más bajo requiere más árboles para modelar todas las relaciones complejas en los datos. Aquí se ha establecido en 0.1.

- *max_depth*: La profundidad máxima de cada estimador individual. max_depth=5 limita la profundidad de cada árbol a 5 nodos, lo cual controla el sobreajuste al limitar la complejidad de los árboles.

- *min_samples_split*: El número mínimo de muestras necesarias para dividir un nodo interno. min_samples_split=5 significa que cada nodo debe tener al menos 5 muestras antes de considerar una nueva división en ese nodo.

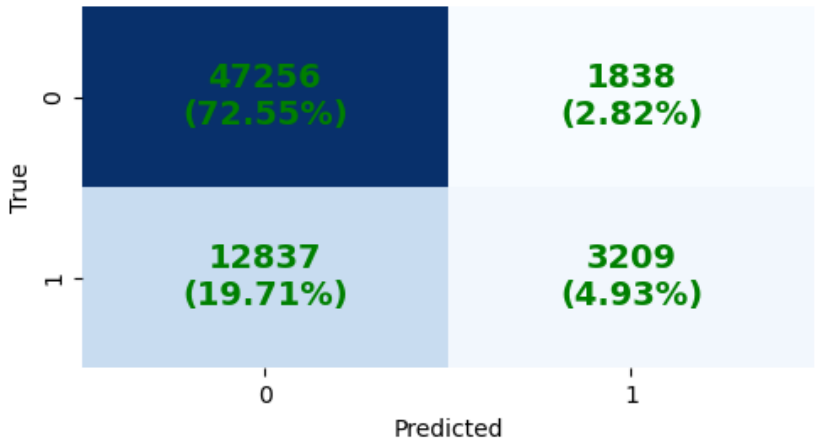
- *min_samples_leaf*: El número mínimo de muestras que un nodo hoja debe tener. min_samples_leaf=2 asegura que cada hoja final de los árboles debe contener al menos 2 muestras, suavizando el modelo y reduciendo el riesgo de sobreajuste.

- *random_state*: Un entero utilizado como semilla para el generador de números aleatorios, asegurando que los resultados sean reproducibles. Aquí se usa random_state=42.

10.6. Optimización del umbral de decisión.

Modelos de clasificación como el Gradient Boosting Classifier no predicen directamente la categoría a la que pertenece una muestra, sino que generan un score entre 0 y 1, y dependiendo si el score es mayor a un umbral de decisión se decide la categoría. Por defecto, este umbral de decisión se fija en 0.5. Como se observa en la figura 13, utilizando el umbral decisión por defecto, se identificó correctamente el 4.93% de los datos sobre el grupo test, los falsos positivos fueron el 2.82% y los falsos negativos tuvieron una tasa del 19.71%. Con respecto a las métricas más relevantes a la hora de evaluar modelos de clasificación, se obtuvieron los siguientes valores: exactitud (accuracy) = 0.775, precisión (precision) = 0.636, sensibilidad (recall) = 0.200 y un f1-score = 0.304.

Figura 13. Matriz de confusión con umbral de decisión de 0.5 en Test



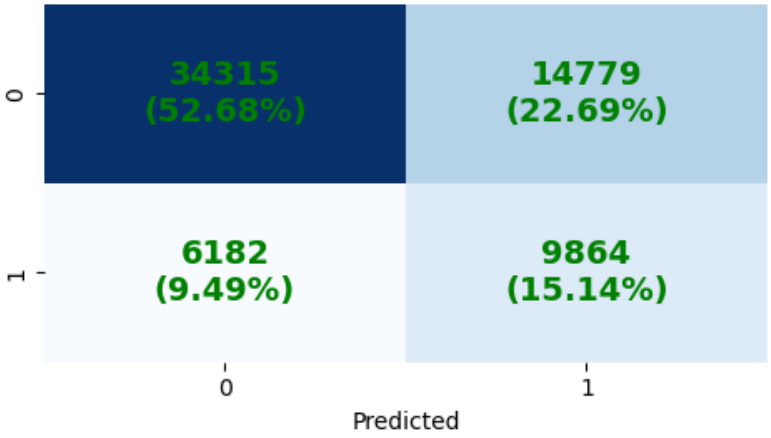
El F1 Score es especialmente útil en contextos de clasificación binaria donde hay una distribución desbalanceada de clases, ya que proporciona una medida balanceada de precisión y sensibilidad, recordemos que la proporción de clientes es de 75.5% para quienes renuevan la póliza y de 24.5% para quienes no realizan la renovación.

Para la predicción de la clase minoritaria de quienes no realizan la renovación, con el modelo Gradient Boosting Classifier inicialmente se obtuvo un F1 score de 0.304. Este valor sugiere que, aunque el modelo es capaz de distinguir entre las clases positivas y negativas, la precisión y sensibilidad para la clase minoritaria eran inicialmente limitadas. Dado que la identificación de registros que no renuevan la póliza ó churn es crítica para nuestro análisis, se decidió ajustar el umbral de decisión del modelo para mejorar el F1 score en esta clase.

Para atrapar más registros en la clase de churn, se realizó un ajuste en el umbral de decisión del modelo. En este caso, se disminuyó el umbral para aumentar la sensibilidad del modelo hacia la clase minoritaria, lo que permitió mejorar la identificación de casos de churn pasando de un umbral de decisión de 0.5 a un umbral de decisión óptimo de 0.24.

El ajuste del umbral tuvo un impacto positivo en la capacidad del modelo para predecir la clase minoritaria mejorando significativamente la cantidad de registros en la clase positiva sobre los datos de prueba, se ha pasado de obtener 4.93% a 15.14% de los registros clasificados como se observa en la matriz de confusión con el umbral optimizado en la figura 14.

Figura 14. Matriz de confusión con umbral optimizado de 0.24 en Test



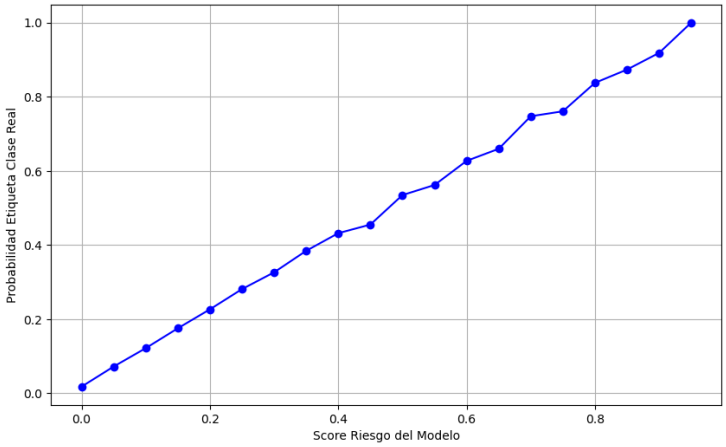
Con respecto a las métricas más relevantes a la hora de evaluar modelos de clasificación, luego de obtener el umbral óptimo para generar la matriz de confusión, se obtuvieron los siguientes valores: exactitud (accuracy) = 0.678, precisión (precision) = 0.400, sensibilidad (recall) = 0.615 y un f1-score = 0.485.

Como se observa el efecto de la optimización del umbral es lograr un mejor balance entre la sensibilidad y la precisión del modelo para reducir el número de Falsos Negativos. Desde el punto de vista del negocio esta reducción es muy importante porque disminuye la probabilidad de no detectar potenciales casos de churn.

10.7. Puntaje de clasificación de los clientes.

Otra posible forma de utilización del modelo no se centra en la predicción positiva o negativa de churn, sino en establecer la probabilidad de hacer churn de acuerdo con el score recibido por el modelo. A partir de los datos obtenidos en el modelo procedemos a analizar la relación del puntaje o score y la respectiva probabilidad de hacer churn de acuerdo con la etiqueta real de los datos. Esta probabilidad se obtiene al analizar el porcentaje de las muestras con el valor asignado del score y comparándolos con los registros que finalmente terminan haciendo churn con la etiqueta real de los datos. El resultado de este análisis se muestra en la figura 15, dónde se observa una relación lineal y prácticamente diagonal entre las dos cantidades, lo cual nos permite inferir que el modelo está bien calibrado y que, en la práctica, podemos relacionar directamente el scoring obtenido con la probabilidad de churn. Para cada registro en la base, se propone generar una etiqueta con el fin de clasificarlos y, de esta forma, agrupar a los clientes según el riesgo que tengan de no renovar.

Figura 15. Probabilidad Etiqueta Clase Real vs Score del Modelo.



Una vez el modelo fue entrenado, se calculan las probabilidades de que cada cliente no renueve su póliza y basado en esas probabilidades se asignan etiquetas de riesgo a cada cliente tomando como referencia el umbral optimo. Estas etiquetas permiten clasificar a los clientes en diferentes grupos de riesgo lo cual facilitará la identificación de segmentos de clientes con mayor probabilidad de no renovar sus pólizas, permitiendo una gestión proactiva sobre la base.

Tabla 9. Etiquetas para las probabilidades de riesgo

Rango Riesgo	Scoring	% Participación
Bajo	Entre 0,00 y 0,14	29,3%
Medio	Entre 0,15 y 0,23	35,2%
Alto	Entre 0,24 y 0,69	33,5%
Muy Alto	Entre 0,70 y 1.00	2,0%

11. IMPACTO ESPERADO

La generación de etiquetas de riesgo y la agrupación de clientes permiten una segmentación precisa, enfocándose en los clientes que más probablemente no renovarán sus pólizas. Con esta clasificación, las estrategias de retención pueden ser más efectivas, dirigiendo esfuerzos y recursos a los clientes que más lo necesitan. Los puntajes de riesgo proporcionan información valiosa para la toma de decisiones, ayudando a priorizar acciones y mejorar la retención de clientes.

Con la implementación de este modelo esperamos reducir la tasa de abandono de los clientes que tienen una póliza de seguro en la línea de auto individual en la compañía de seguros. Con el modelo de clasificación de riesgo, se podrá identificar de manera temprana aquellos clientes que presenten patrones los cuales puedan llevar un mayor riesgo de deserción, lo que permitirá a la Gerencia de Cartera implementar estrategias proactivas de retención, brindando ofertas personalizadas, servicios mejorados y/o atención especializada para satisfacer a sus clientes.

Con este modelo se espera mejorar la rentabilidad de la cartera de pólizas para que de esta forma se logren asignar de manera más eficiente los recursos financieros y operativos de la compañía.

12. CONCLUSIONES

El desarrollo de un modelo supervisado de clasificación para predecir la renovación de pólizas de seguro de autos individuales, guiado por la metodología CRISP-DM, ha permitido obtener resultados significativos que pueden beneficiar a la compañía de seguros para la gestión de los clientes en sus diferentes áreas.

La investigación y aplicación de diversas técnicas de análisis de datos y modelos de aprendizaje automático, como el Gradient Boosting, permitieron identificar patrones de deserción de clientes con una precisión notable. El modelo final, con un AUC de 0.71, demostró ser una herramienta eficaz para predecir la probabilidad de abandono de los clientes.

La metodología CRISP-DM proporcionó un marco estructurado que facilitó la comprensión del negocio, la comprensión y preparación de los datos, el modelado, la evaluación y el despliegue. Esta metodología aseguró que todas las fases del proyecto se realizaran de forma exitosa.

Con base en la clasificación de los clientes en cuatro niveles de riesgo (muy alto, alto, medio y bajo), la compañía puede ahora implementar campañas de retención personalizadas. Estas campañas deben centrarse en los clientes clasificados como de muy alto y alto riesgo, mientras que se mejoran los incentivos y la comunicación con los clientes de riesgo medio y bajo.

13. REFERENCIAS BIBLIOGRÁFICAS

Black, J. E., Kueper, J. K., & Williamson, T. S. (2023). An introduction to machine learning for classification and prediction. *Family Practice*, *40*(1), 200–204. https://doi.org/10.1093/FAMPRA/CMAC104

Borda, P., Dabenigno, V., Freidin, B., & Güelman, M. (2020). Estrategias para el análisis de datos cualitativos. *Universidad de Buenos Aires*.

Cam Gensollen, C. R. (2022). Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa. *Ingeniería Industrial*. https://doi.org/10.26439/ing.ind2022.n.5808

Contreras Serrano, C. G. (2016). Fidelización y rentabilización de usuarios de seguros todo riesgo de vehículos por medio de la venta cruzada y la venta escalonada. Un enfoque promocional para la industria aseguradora. *Universidad & Empresa, ISSN 0124-4639, ISSN-e 0124-4639, Vol. 18, Nº. 30, 2016, Págs. 143-157*, *18*(30), 143–157. https://doi.org/10.12804/rev.univ.empresa.30.2016.07

Hush, J. (2020). *Python Para el Análisis de Datos* (ndependently Published, Ed.; Vol. 1).

IBM. (n.d.). *¿Qué es machine learning?* Https://Www.Ibm.Com/Mx-Es/Analytics/Machine-Learning.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/NATURE14539

Liu, M., Chen, W., & Yang, L. (2022). Optimization of Experimental Teaching System based on ACSI Model. *Proceedings - 2022 International Conference on Information System, Computing and Educational Technology, ICISCET 2022*, 106–109. https://doi.org/10.1109/ICISCET56785.2022.00035

Mathworks. (n.d.). *Análisis predictivo Tres cosas que es necesario saber*. Https://La.Mathworks.Com/Discovery/Predictive-Analytics.Html.

Naik, K. S., & Bhise, A. (2022). Risk Identification Using Quantum Machine Learning for Fleet Insurance Premium. *Communications in Computer and Information Science*, *1729 CCIS*, 277–288. https://doi.org/10.1007/978-3-031-21750-0_24

Tudoran, A. A. (2022). A machine learning approach to identifying decision-making styles for managing customer relationships. *Electronic Markets*, *32*(1), 351–374. https://doi.org/10.1007/S12525-021-00515-X

Zijtregtop, E. A. M., Winterswijk, L. A., Beishuizen, T. P. A., Zwaan, C. M., Nievelstein, R. A. J., Meyer-Wentrup, F. A. G., & Beishuizen, A. (2023). Machine Learning Logistic Regression Model for Early Decision Making in Referral of Children with Cervical Lymphadenopathy Suspected of Lymphoma. *Cancers*, *15*(4). https://doi.org/10.3390/CANCERS15041178