**Logotipo

Descripción generada automáticamente**

**Modelo de clasificación de clientes para la unidad de negocio de previsión en una empresa funeraria**

Gloria Patricia Cadavid Zuluaga

Laura Hernandez Rebolledo

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

|  |  |
| --- | --- |
| **Cita** | (Cadavid Zuluaga & Hernandez Rebolledo, 2024) |
| **Referencia**  **Estilo APA (2020)** | Cadavid Zuluaga, G., & Hernández Rebolledo, L. A. (2024). *Modelo de clasificación de clientes para la unidad de negocio de previsión en una empresa funeraria* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  |  |
| --- | --- |
|  | Diagrama  Descripción generada automáticamente con confianza media |

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Tabla de contenido**

[Resumen 7](#_3rdcrjn)

[Abstract 8](#_26in1rg)

[1. Descripción del problema 9](#_35nkun2)

[1.1. Problema de negocio 10](#_1ksv4uv)

[1.2. Aproximación desde la analítica de datos 11](#_44sinio)

[1.3. Origen de los datos 12](#_z337ya)

[1.4. Métricas de desempeño 13](#_3j2qqm3)

[2. Objetivos 14](#_4i7ojhp)

[2.1. Objetivo general 14](#_2xcytpi)

[2.2. Objetivos específicos 14](#_3whwml4)

[3. Datos 15](#_2bn6wsx)

[3.1. Datos originales 15](#_qsh70q)

[3.2. Datasets 17](#_3as4poj)

[3.3. Analítica descriptiva 19](#_1pxezwc)

[4. Referencias](#_49x2ik5) 19

5. Anexo 21

**Lista de tablas**

**Tabla 1**  Agrupación descripción de campos. 19

**Lista de figuras**

[**Figura 1** Distribución de la variable de salida.](#_2r0uhxc)

[**Figura 2** Distribución de las variables categóricas](#_1664s55)

**Figura 3** Mapa de correlación por cluster

**Siglas, acrónimos y abreviaturas**

**IEEE** Instituto de Ingenieros Eléctricos y Electrónicos

**ML** Machine Learning

**UdeA** Universidad de Antioquia

# Resumen

El principal reto de la línea de negocio Territorio de Previsión en Los Olivos Antioquia-Chocó es la renovación de los contratos adquiridos por los clientes, lo que impacta directamente en la sostenibilidad del negocio a largo plazo. La falta de renovaciones representa una amenaza para el crecimiento y la rentabilidad de la empresa. Este proyecto tiene como objetivo implementar métodos de machine learning para desarrollar un modelo de clasificación capaz de predecir si un cliente renovará su contrato o no. Al identificar patrones y características relevantes de los clientes, el modelo permitirá segmentar y priorizar los esfuerzos de retención de manera más eficiente. De esta forma, se optimizarán los recursos en las campañas de fidelización, maximizando el impacto de las acciones comerciales. El modelo también brindará insights valiosos sobre los factores que influyen en la renovación de contratos, facilitando decisiones más informadas y estrategias más efectivas para mejorar la tasa de renovación y garantizar la sostenibilidad del negocio a largo plazo.

*Palabras clave*: Renovación, Clasificación, Asignación de recursos, Machine learning.

# Abstract

The main challenge for the *Territorio de Previsión* business line in Los Olivos Antioquia-Chocó is the renewal of contracts acquired by customers, which directly impacts the long-term sustainability of the business. The lack of renewals poses a threat to the company’s growth and profitability. This project aims to implement machine learning methods to develop a classification model capable of predicting whether a customer will renew their contract. By identifying relevant patterns and characteristics of customers, the model will allow for better segmentation and prioritization of retention efforts. This will optimize the resources allocated to loyalty campaigns, maximizing the impact of commercial actions. The model will also provide valuable insights into the factors influencing contract renewals, enabling more informed decisions and more effective strategies to improve renewal rates and ensure the long-term sustainability of the business.

*Keywords: Renewal, Classification, Resource allocation, Machine learning.*

# Descripción del problema

Un estudio realizado por la Fundación MAPFRE [1] sobre la composición del mercado asegurador en Latinoamérica revela un creciente interés por la adquisición de coberturas de seguros No Vida. Este incremento responde, entre otros factores, a una mayor valoración del bienestar y la protección integral de las familias. Además, se proyecta que la diversificación de productos, como seguros inclusivos y de salud, seguirá siendo una tendencia clave para el sector asegurador en la región durante los próximos años.[2]

Sin embargo en Colombia, el índice de cobertura en el mercado asegurador se mantiene bajo, reflejando los retos estructurales que enfrentan las empresas del sector. [3] Una de las principales causas de este fenómeno es la limitada cultura de previsión financiera entre los ciudadanos, quienes tienden a percibir las pólizas de seguro como un gasto no recuperable. Según un informe de la Fundación MAPFRE, una proporción significativa de las adquisiciones de seguros está motivada por requerimientos legales, en lugar de un interés genuino en la protección financiera, lo que contribuye a un crecimiento lento del sector asegurador en el país. [1]

Al igual que en el sector de los seguros generales, las empresas funerarias han adoptado estrategias de mejora continua para optimizar sus procesos y adaptarse a tendencias emergentes. En ambos mercados, la personalización de los servicios y la satisfacción del cliente son elementos fundamentales para fomentar la fidelización y atraer nuevos consumidores. Por esta razón, muchas compañías concentran sus esfuerzos en estrategias de marketing innovadoras, enfocadas en identificar y satisfacer las necesidades específicas de los clientes. No obstante, la renovación constante de servicios sigue representando un desafío crítico, especialmente en un entorno donde las demandas de los consumidores cambian rápidamente.

En este contexto, el uso de la analítica de datos y las técnicas de machine learning (ML) se ha consolidado como una herramienta esencial para la segmentación y clasificación de clientes. Este enfoque permite identificar patrones en el comportamiento del consumidor y diseñar estrategias más precisas y personalizadas. Los modelos de clasificación, como los árboles de decisión, Random Forest y K-Nearest Neighbors (KNN), han demostrado ser efectivos para procesar datos relevantes, incluyendo atributos demográficos, geográficos y transaccionales, lo que facilita la predicción de comportamientos futuros [5], [6].

Además, estudios recientes subrayan cómo las técnicas avanzadas, como las redes neuronales profundas, mejoran la capacidad de los modelos para estimar el valor de vida del cliente (Customer Lifetime Value), maximizando la rentabilidad y optimizando la asignación de recursos en estrategias de marketing [7], [8]. Este enfoque basado en datos no solo fortalece la capacidad de las empresas para anticipar tendencias, sino que también promueve la toma de decisiones estratégicas en mercados altamente competitivos.

## Problema de negocio

Los Olivos Antioquia-Chocó es una empresa del sector funerario que tiene principalmente dos líneas de negocio, en adelante Territorios, una de ellas Territorio Fúnebre o de homenajes y otra **Territorio Previsión**. Si bien la actividad principal del negocio es el Territorio Fúnebre, ambos Territorios son relevantes para la conservación de la empresa en el tiempo. El Territorio Previsión, es la unidad de negocio que enmarca todo lo relacionado con coberturas de asistencia integral para la familia y las mascotas, así cómo auxilios de diferentes índoles.

Es así como el Territorio Previsión de Los Olivos Antioquia-Chocó tiene características que enmarcan dicha unidad de negocio dentro del sector seguros. De tal forma que hoy el desafío está en encontrar la estabilidad de la empresa en el tiempo, a través de la conservación e incremento de los clientes. En ese orden de ideas, al clasificar los clientes nuevos con la probabilidad de **renovación** de la cobertura, se podrán identificar las personas en las que será más rentable enfocar un programa estructurado de las campañas de mercadeo. Tal como lo resaltan en los trabajos de C. Álvarez Florez y L. T. Molina Ruiz, así como de M. Sánchez Sardaña. Álvarez Florez y Molina Ruiz donde subrayan la importancia de la personalización en la atención al cliente, aún más relevante en el contexto funerario donde las decisiones son altamente influenciadas por factores emocionales.

Inicialmente, este tipo de empresas se enfocan principalmente en ofrecer el servicio de previsión exequial, que abarca todos los actos y procedimientos relacionados con el proceso funerario, incluyendo aspectos legales, familiares, así como acompañamiento y orientación. A pesar de que los servicios funerarios pueden implicar costos elevados en el momento de la necesidad, los precios para adquirir la protección de este servicio suelen ser relativamente bajos. El modelo de negocio se basa en el pago constante a lo largo del tiempo, con el fin de asegurar la permanencia en la empresa hasta que se requiera la cobertura [4].

Este sector fue fuertemente afectado en años recientes por factores imprevistos, como la pandemia, lo que repercutió en los indicadores de rentabilidad, costos y servicios prestados, impactando la sustentabilidad del modelo. En respuesta, las empresas han ampliado su oferta, incluyendo nuevas coberturas orientadas a la protección integral de las familias [14].

El crecimiento del negocio se centra en la expansión de su base de clientes, es decir, en aumentar el número de familias que deseen asegurar a sus seres queridos por un periodo prolongado [4]. Sin embargo, durante 2023 y 2024, se observó una tasa significativa de cancelaciones de contratos de previsión. Estas cancelaciones se dividen en dos causas principales: por políticas de la empresa (80%) y por decisión del titular (20%). Las cancelaciones por políticas de la empresa ocurren cuando el titular no realiza pagos, no renueva su contrato o pierde la relación con las empresas o cooperativas proveedoras del beneficio, tras un periodo de 120 días. Las cancelaciones restantes se deben a diversas razones manifestadas por los titulares, como la percepción de que no es necesario continuar con la protección exequial.

Este fenómeno refleja un comportamiento estático, donde el número de retiros supera al de los nuevos afiliados. Como se mencionó previamente, el crecimiento depende de la capacidad para atraer a nuevas familias que deseen obtener cobertura.

Para reducir las cancelaciones por políticas, se han implementado estrategias de retención. Estas campañas están dirigidas a contactar a los afiliados que no han realizado pagos o que no se han comunicado con la empresa durante más de un mes, con el objetivo de evitar que pasen más de los 120 días. Las estrategias se basan en la asignación óptima de recursos, enfocándose en los registros que deben ser contactados y en la efectividad para garantizar el pago o la renovación del contrato. Es crucial una asignación eficiente para focalizar los esfuerzos en aquellos casos con mayores probabilidades de éxito, y ajustar las estrategias para aquellos titulares que desde el inicio muestran signos de no querer continuar protegidos en su contrato de previsión.

Actualmente, los indicadores como la contactabilidad y la efectividad son factores determinantes para el éxito de las campañas. Sin embargo, la variabilidad del comportamiento mensual y la diversidad de clientes según el canal de venta complican la predicción y asignación de recursos. Este desequilibrio genera problemas de costos, ya que, en lugar de satisfacer las necesidades de retención y permanencia, afectan económicamente a la empresa debido a la falta de estimación adecuada de recursos.

## Aproximación desde la analítica de datos

En el marco de la optimización de recursos para la toma de decisiones, el análisis del comportamiento del consumidor y el desarrollo de sistemas de recomendación destacan como áreas clave de la Ciencia de Datos aplicadas a las problemáticas descritas. Con este enfoque, se propone diseñar un algoritmo basado en técnicas de aprendizaje automático supervisado (ML) para clasificar clientes nuevos en relación con la probabilidad de renovar la cobertura adquirida en Territorio Previsión. Este planteamiento se alinea con metodologías ampliamente utilizadas en diversos sectores para el perfilamiento de clientes y la predicción de comportamientos.

Técnicas como los árboles de decisión, Random Forest, Extreme Gradient Boosting (XGBoost) y K-Nearest Neighbors (KNN) han demostrado ser particularmente eficaces en este tipo de análisis. Estas metodologías no solo ofrecen precisión en los resultados, sino que también permiten manejar adecuadamente problemas de sobreajuste durante el entrenamiento del modelo, según estudios recientes [9], [10]. Álvarez Flórez y Molina Ruiz [11], así como Sánchez Sardaña [12], destacan la importancia de la personalización en la atención al cliente, especialmente en contextos como el sector funerario, donde los factores emocionales tienen un impacto significativo en las decisiones de los consumidores.

Se recomienda la adopción de estas técnicas como un método más efectivo para optimizar la asignación presupuestal en campañas de marketing dirigidas a nuevos clientes. Esto no solo aumentaría la tasa de conversión de dichas campañas, sino que también permitiría focalizar esfuerzos en perfiles de clientes con alta probabilidad de renovar sus coberturas, lo cual contribuye directamente al crecimiento sostenido y la estabilidad del negocio a largo plazo. Además, estas mejoras fortalecerían la relación con los clientes, incrementando el reconocimiento y la competitividad de la empresa en su sector.

En la implementación de estos modelos predictivos, es crucial considerar el impacto potencial de errores de clasificación, como los falsos negativos y los falsos positivos. Mientras que los falsos negativos pueden limitar la identificación de clientes con alta probabilidad de renovación, los falsos positivos podrían reducir la eficiencia de las campañas, generando un uso inadecuado de los recursos. Por ello, se sugiere establecer cronogramas de seguimiento y control, que incluyan la validación continua de los resultados y el ajuste del modelo según sea necesario para maximizar su efectividad y precisión [13].

Además, la solución que se propone:

* No incluye el diseño y desarrollo de estrategias de mercadeo con los clientes.
* No implica sugerencias de coberturas a priorizar.
* No sugiere la definición de los montos presupuestales a asignar a cada campaña.
* No contempla el despliegue del algoritmo entrenado y ajustado.

Respecto a riesgos para el entrenamiento del modelo, se consideran:

* El conjunto de datos disponible dado el origen transaccional.
* La imputación de valores faltantes, que reducirían la cantidad del dataset.
* El gran volumen de datos que por el tipo de técnicas generaría sobreajuste de modelos.

## Origen de los datos

Los datos utilizados provienen de una empresa del sector fúnebre que ha estado operando en el departamento de Antioquia y Chocó desde el año 1991. Estos datos se componen de registros de los contratos correspondiente a la previsión, así como las coberturas contratadas según el producto (como previsión exequial, auxilio económico, asistencias para personas (hogar, médica, legal y viaje) , previsión exequial para mascotas y asistencias para mascotas (médico veterinario, peluquería, servicio de medicinas ). Los registros para trabajar se recopilaron desde la base de datos de la empresa seleccionando desde enero del 2023 hasta diciembre del 2023.

Los datos seleccionados desde los sistemas internos, incluyen tanto información demográfica de los clientes (como edad, género, y localización geográfica) como detalles de número de veces renovados, valor pagado, otros contratos activos por titular, generación de PQRS o comunicaciones directas en las líneas de atención. Es importante señalar que la información personal de los clientes ha sido anonimizada para proteger la privacidad, y se ha utilizado con fines exclusivamente analíticos, de acuerdo con las políticas de protección de datos.Este conjunto de datos se obtuvo a través de la transformación directa al momento de extraer la información, con el fin de garantizar el anonimato toda la información se privatizó y se uso solo los ID internos, esto refleja una variedad de comportamientos y patrones de los clientes para continuar o no con su contrato activo, lo que permite analizar patrones y tendencias en los clientes.

## Métricas de desempeño

**Del modelo**

* Accuracy (Precisión): Ésta indica cuántas predicciones acertadas se realizan respecto al total de predicciones. En el caso del proyecto permitirá verificar la capacidad del modelo para identificar los clientes que renovarán la cobertura. Generalmente un valor alrededor del 70% indica un buen desempeño del modelo y sería lo mínimo esperado en relación a las técnicas propuestas.
* Recall (Sensibilidad): Indica el número de clientes clasificados correctamente cómo **Renueva** en relación con el total de clientes que efectivamente renovaron. En ese orden de ideas, un recall de un 50% es aceptable para el comportamiento del modelo.
* Matrices de confusión: Permite identificar visualmente la distribución de las predicciones correctas e incorrectas, mejorando la comprensión de la tasa de verdaderos positivos y verdaderos negativos alcanzados.

**Del negocio**

* Tasa de desempeño de conversión: Número de registros propuestos a renovar versus número de renovación. Se espera que con la implementación de los modelos, también supere cómo mínimo un 70% correspondiente con el desempeño del modelo.
* Costo de adquisición: (1- (Valor de la venta/ Valor de campaña)), en la actualidad es de un 9% de recuperación de la inversión realizada, de ahí que con la implementación del modelo y la optimización de los recursos, sería posible aumentar hasta un 3% la recuperación del capital invertido.

# Objetivos

## Objetivo general

Identificar la posibilidad de renovación de los contratos de previsión de clientes de los olivos Antioquia-Chocó a través de modelos de clasificación de machine learning usando registros del 2023 para brindar información que permita la asignación óptima del presupuesto de campañas.

## Objetivos específicos

Los objetivos específicos son las metas a corto plazo que se deben realizar para poder alcanzar el objetivo central o principal, conocido a su vez como objetivo general.

* Consolidar la información de las variables relevantes asociadas a un contrato de previsión, a través de la integración de las tablas de la base de datos, con el fin de generar el dataset.
* Establecer las variables del dataset que tienen mayor relación o incidencia mediante la exploración y preparación del dataset
* Entrenar y evaluar el desempeño de los diferentes modelos implementados para la clasificación de los clientes, utilizando métricas de evaluación estándar para cada modelo.
* Validar el desempeño de los modelos entrenados, comparando sus métricas de evaluación, con el fin de seleccionar el modelo más adecuado para su implementación final.
* Definir el número de personas a abordar basado en las predicciones obtenidas, para optimizar la asignación de recursos en la campaña, maximizando su efectividad y resultados.

# Datos

Este apartado, presenta la estructura de datos y la correspondiente adecuación para llevar a cabo el proyecto propuesto, con el análisis de las variables relevantes. Además, se realiza un análisis descriptivo de sus características para identificar las relaciones entre variables y seleccionar las más adecuadas para aplicar modelos de machine learning.

## Datos originales

Los datos utilizados corresponden a los clientes actuales del territorio de previsión, extraídos desde la base de datos de la compañía a través de SQL Server, con un periodo comprendido entre el año 2023. El dataset fue construido mediante la unión de diversas tablas, considerando variables como la edad, número de pagos, el mecanismo de pago, la composición del producto actual y la zona geográfica. Cada una de las variables seleccionadas no está relacionada con información sensible, lo que contribuye a mitigar el riesgo asociado al tratamiento de datos, dado que se considera la naturaleza del negocio. Además, se trabajará desde un archivo en formato CSV de un tamaño de 8.02MB, ya que la base de datos es transaccional, lo que asegura que se utilizarán datos reales del periodo sin ser alterados por posibles cambios posteriores.

El dataset contiene un total de 57,462 registros y 28 columnas, abarcando tanto datos cuantitativos como cualitativos. Cada registro representa a un titular de contrato creado en el periodo indicado, y las columnas corresponden a características del comportamiento del cliente y del producto. En términos de seguridad, la base de datos contiene únicamente dos datos sensibles según la normativa vigente: la edad y el género. Para garantizar la protección de la información, se asigna un Id\_Titular generado aleatoriamente desde la base de datos y se utiliza el primer nombre de manera genérica. Los datos se montan en un repositorio privado, y el acceso a estos se realiza mediante token en el modelo, asegurando su integridad y confidencialidad.

Tabla 1

|  |  |  |  |
| --- | --- | --- | --- |
| Grupo Variables | Descripción | Variables | Tipo |
| Sociodemográficas | Consolida todas las variables descriptivas del usuario titular | Edad  Género  Estado civil  Tipo municipio | Int  Object  Int  Object |
| Características del contrato | Unifica las características que identifican el tipo de contrato como son: la vigencia, la forma de pago, el canal de venta, póliza grupal, que indica si está afiliado bajo un contrato que vincula varios usuarios y la fecha de renovación | Id\_Titular  Ingreso\_titular  Contrato  Forma Pago  Vigencia contrato  Método de pago  Canal  Póliza grupal  Fecha renovación | Int  Date  Int  Object  Int  Object  Object  Int  Date |
| Características del productos y protegidos | Agrupa la cantidad de productos de la UEN Previsión, los cuáles pueden estar vinculados hasta 3 entre ellos | Beneficiarios  Mascotas  Auxilio  Asistencia  Previsión exequial  Prima mes | Int |
| Relación cliente empresa | Son las diferentes interacciones que tiene el cliente titular con la empresa durante la vigencia del contrato | Cambios de periodicidad  # de pagos  Valor pagado ct vigente  # de renovaciones  #de contratos activos  Valor pagado\_otros ct  PQRS  Comunicación | Int |

Tabla 1 Agrupación descripción de campos. Dada la cantidad de variables se describe el agrupamiento de los datos, considerando 4 aspectos importantes y sus correspondientes tipos.

Mediante una investigación de información de la DIAN, se genera un archivo con el listado de municipios de Colombia y como estos se consideran en base a su estructura. Si puede ser considerado rural o urbano. Este archivo CSV con 757 registros de 12,7 KB se usa como apoyo para reducir la variable sociodemográfica de la base.

Se identifican 3 variables con datos nulos, el estado con 83, la edad del titular con 436 y el género con 34.442 de estas 3 variables la edad y el estado se eliminan directamente. Mientras que con el género se decide continuar para que mediante tratamiento de datos se pueda disminuir el número de datos nulos

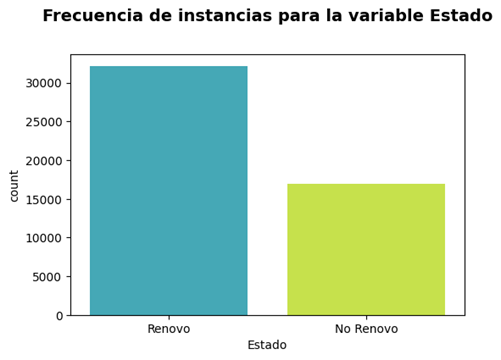


Figura 1. Distribución de variable de salida

Nuestra variable de salida es de tipo categórica donde podemos identificar quienes continúan o no con su contrato de previsión exequial. Es importante destacar que la variable que mayor información tenemos es para los contratos que renuevan, esto nos permitirá tener una mayor diversidad en la probabilidad de renovación.

## Datasets

La información se extrajo de diversas tablas de la base de datos principal de la empresa, las cuales se agrupan en cuatro tipos que corresponden a los diferentes grupos de variables mencionadas. La tabla principal contenía las características de los contratos, mientras que otras tablas proporcionaban información sobre los productos y sus protegidos, incluyendo las tarifas asociadas. Además, se incluyó una tabla con datos sociodemográficos que completaba la información a nivel de cliente. Adicionalmente, se utilizaron tres tablas auxiliares para traducir información codificada, como el segmento de venta, el mecanismo de venta y el municipio.

En la consulta, estas tablas se utilizaron para caracterizar los productos y sus protegidos a través de conteos, con el fin de unificar la información en un solo registro por fila. La información relacionada con el contrato y los datos sociodemográficos se integraron directamente, generando una columna para cada variable. Además, el comportamiento del cliente con la empresa se modeló en función del número de interacciones, donde un valor de 0 indica la ausencia de interacciones, y cualquier valor superior a 0 refleja la cantidad de interacciones registradas.

La selección de los datos para el proceso de entrenamiento y prueba del modelo, se divide en dos líneas de tiempo. Para el entrenamiento, se utilizó el conjunto de datos recopilado a lo largo del año 2023, el cual incluye información histórica de campañas previas, comportamiento de clientes, y otros factores relevantes para dimensionar futuras campañas. Este conjunto permite que el modelo aprenda patrones y relaciones clave entre las características de las campañas y los registros obtenidos. En cuanto a la prueba, se selecciona un conjunto de datos específico de personas que deben renovar sus registros entre enero y febrero de 2024. Este grupo se usa para evaluar la capacidad del modelo para predecir el número ajustado de registros para ese mes, comparando las predicciones con los resultados reales obtenidos tras la ejecución de la campaña. Este enfoque asegura una evaluación de la capacidad de generalización del modelo en un contexto temporal futuro.

*Imputación de datos para la variable género faltantes*

En el conjunto de datos se observan valores nulos en las variables como edad, estado y primer nombre, los cuales no pueden ser imputados y serán descartados directamente. Por lo tanto, se eliminarán los valores nulos en estas variables.

En cuanto al género, se puede imputar utilizando el “primer\_nombre”, a través de la librería gender\_guesser.detector tal y como se observa en **el anexo 3.**

Es importante destacar que esta librería sólo puede identificar el género si la entrada se encuentra en un formato específico: la primera letra debe estar en mayúsculas y el resto en minúsculas.

Para garantizar la correcta imputación, se realiza una limpieza previa de los datos: se eliminan los caracteres especiales (**ver anexo 1**), se convierte todo a minúsculas y, mediante un ciclo, se ajusta cada nombre para que su primera letra sea mayúscula como se hizo (**ver anexo 2**).

Tras esta limpieza, se aplica la librería mencionada. Inicialmente, teníamos aproximadamente 34,000 valores nulos, lo que equivalía al 60% del conjunto de datos. Después de imputar el género, solo quedaban 8,000 valores no identificados por la librería, lo que representa el 14% del dataset. Para unificar las categorías del género según el resultado de la librería, se utilizó un condicional que agrupa las variables como "M" (masculino) y "F" (femenino).

Finalmente, después de limpiar e imputar los datos, se obtuvo un conjunto de datos con 48,977 registros.

*Escalado y normalización de los datos*

Para el escalado de las variables, se utiliza la librería StandardScaler(). Esta librería se emplea para estandarizar las variables numéricas, unificando sus valores en una escala estándar. Algunas variables presentan valores muy altos, mientras que otras tienen valores muy bajos debido a la naturaleza de las mismas. StandardScaler ayuda a estandarizar estas variables, transformándolas en una medida común.

Se adjunta una base de datos adicional de la DIAN en formato CSV, que contiene un total de 727 registros, clasificados por tipo de zona (rural o urbana). Esto tiene como objetivo reducir el volumen de las variables a través de la técnica de dummyficación.

Para la dummyficación, se emplea One-Hot Encoding para las variables tipo de geografía y género. Para la variable mecanismo de pago, se utiliza Label Encoding, mientras que para el segmento de venta, se aplica una jerarquización de la empresa, agrupando los mecanismos de pago según su importancia o preferencia, al igual que el segmento de venta.

*Cálculo y eliminación de outiler*

Para identificar los outliers, se utiliza el modelo de K vecinos más cercanos (KNN). El valor de K se determina mediante una búsqueda de hiperparámetros (Grid Search), lo que permite encontrar el número óptimo de vecinos para el modelo. En este caso, se obtuvo 5 como el mejor valor de K. Este proceso se aplica únicamente a las variables numéricas. Una vez determinado el valor óptimo de K, se identifican los outliers, los cuales son eliminados del conjunto de datos de acuerdo con los resultados obtenidos a partir del mejor hiperparámetro.

## Analítica descriptiva

# En el análisis exploratorio se detalla en el notebook realizado. Donde primero se evalúa la distribución de nuestras variables categóricas tal como segmento de venta, género y tipo de municipio.

# 

# Figura 2. Distribución de las variable categóricas

# En la distribución de segmento de venta tiene mayor información los relacionados al segmento de venta de asociadas por lo que se hace necesario que al momento de evaluar el modelo se deberá conocer el ajuste por los segmentos de venta

# Luego se evalúan las variables numéricas Forma de pago, vigencia del contrato, número de pagos, valor pagado, cambio de periodicidad, número de renovación, contratos activos adicionales, valor pagado en otros contratos, número de beneficiarios, mascotas, exequial, auxilio, asistencia, valor de la prima mensual, número de pqrs y el numero de comunicación realizadas por el cliente. Estas variables se evalúan mediante un gráfico de mapa de relación por cluster.

# 

# Se observa que la variable tener mascota tiene una incidencia directa en el valor de la prima. Además, se identifica que el número de renovaciones es inversamente proporcional al modo de pago, considerando este factor como una característica del contrato. Solo el producto de previsión y la prima tienen una relación directa, mientras que las coberturas de auxilio y asistencia se toman principalmente como adicionales. También se nota que el número de pagos es inversamente proporcional a la presencia de asistencia. Este hallazgo sugiere un aspecto negativo, ya que la inclusión de asistencia puede influir en la decisión de si se realiza o no el pago. Finalmente, este análisis permite descartar variables adicionales

# 

# Figura 3. Mapa de correlación por cluster

# Se observa que la variable tener mascota tiene una incidencia directa en el valor de la prima. Además, se identifica que el número de renovaciones es inversamente proporcional al modo de pago, considerando este factor como una característica del contrato. Solo el producto de previsión y la prima tienen una relación directa, mientras que las coberturas de auxilio y asistencia se toman principalmente como adicionales. También se nota que el número de pagos es inversamente proporcional a la presencia de asistencia. Este hallazgo sugiere un aspecto negativo, ya que la inclusión de asistencia puede influir en la decisión de si se realiza o no el pago. Finalmente, este análisis permite descartar variables adicionales que no aportan información relevante, como el número de comunicaciones durante el periodo.

# Referencias

[1] Servicio de Estudios de Fundación MAPFRE, Elementos para la expansión del seguro en América Latina, 2017.

[2] "Tendencias del mercado de seguros en América Latina para 2024", Fundación MAPFRE (2024).

[3] "Seguros de vida ahorro," MAPFRE Colombia, discute los retos en educación financiera y la baja tasa de penetración de seguros en el mercado local.

[4] J. Casado Herrero, "Más allá de la venta: El consumidor y el sector funerario en previsión", *Trabajo de Fin de Grado*, tutelado por J. Gutiérrez Cillán, Universidad de Valladolid, 8 de julio de 2019.

[5] M. Kaushik y R. Kaur, "Customer Segmentation for Targeted Marketing Using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 13, no. 5, pp. 112-119, 2022.

[6] X. Zhao et al., "Application of Data Mining in Customer Behavior Analysis," Journal of Big Data Research, vol. 5, no. 2, pp. 89-95, 2023.

[7] S. Gupta and L. R. Lehmann, "Modeling Customer Lifetime Value: Concepts and Applications," Marketing Science, vol. 42, no. 3, pp. 1-10, 2023.

[8] K. Deepa et al., "Predictive Models for Customer Retention Using Machine Learning," IEEE Access, vol. 10, pp. 56789-56802, 2024.

[9] T. Hastie, R. Tibshirani, y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.

[10] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[11] C. Álvarez Florez y L. T. Molina Ruiz, "Modelos de aprendizaje automático para la predicción de la preferencia en el uso de canales de atención para un fondo de pensiones y cesantías," Tesis, Universidad de Antioquia, Medellín, Colombia, 2023

[12] M. Sánchez Sardaña, "Modelo predictivo de venta cruzada en productos de Vida y Salud: Random Forest vs XGBoost," Trabajo Fin de Máster, Máster en Ciencias Actuariales y Financieras, Universidad Carlos III de Madrid, Madrid, España, junio 2022.

[13] Han, M. Kamber, y J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Elsevier, 2011.

[14] Funerales La Aurora, "Efectos de la pandemia: La mirada desde una funeraria", *Funerales La Aurora*, 4 de enero de 2022. [Enlace] Disponible en:<https://funeraleslaaurora.com/2022/01/04/efectos-de-la-pandemia-la-mirada-desde-una-funeraria/>

**Anexos**

**Anexo 1. Algoritmo usado para colocar nombre en minúscula y eliminar acentuación**

|  |
| --- |
| **# Función para limpiar el nombre def limpiar nombre(nombre):  if isinstance(nombre, str):   nombre = nombre.lower()  nombre = re.sub(r'[^a-záéíóúüñ\s]', '', nombre)   return nombre  else:  return nombre df['Primer\_nombre'] = df['Primer\_nombre'].apply(limpiar\_nombre)** |

**Anexo 2. Algoritmo para colocar la primera letra en mayúscula y el resto en minúscula**

|  |
| --- |
| **def formatear\_nombre(nombre):  if isinstance(nombre, str):  Letras = nombre.split()  Letras\_formateadas = [Letras.capitalize() for Letras in Letras]  return " ".join(Letras\_formateadas)  else:  return nombre df['Primer\_nombre'] = df['Primer\_nombre'].apply(formatear\_nombre)** |

**Anexo 3. Algoritmo para lograr la imputación de género por primer nombre**

|  |
| --- |
| **# inicializarlo d = gender.Detector() def obtener\_genero(nombre):  gender\_result = d.get\_gender(nombre)  return gender\_result  # Como el género viene en cuatro posibles respuestas se ajusta a lo esperado**  **def convertir\_genero(genero):  if genero == 'male':  return 'M'  elif genero == 'female':  return 'F'  elif genero == 'mostly\_male':  return 'M'  elif genero == 'mostly\_female':  return 'F'  elif genero == 'andy':**  **return 'NULL'  elif genero == 'unknown':  return 'NULL'  else:  return genero  df.loc[df['Genero'].isnull(), 'Genero'] = df.loc[df['Genero'].isnull(), 'Primer\_nombre'].apply(obtener\_genero)  df['Genero'] = df['Genero'].apply(convertir\_genero)** |