DESARROLLO DE UN MODELO PREDICTIVO PARA LA ESTIMACIÓN DEL COSTO DE SEGUROS MÉDICOS MEDIANTE ANÁLISIS DE HISTORIAL DE SINIESTRALIDAD Y SEGMENTACIÓN DE CARTERAS DE USUARIOS

Laura Catalina Guzmán Sandoval - 9026888
Xammy Alexander Victoria González - 9026569

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

Director
Cristian Alejandro Torres Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 2026

FICHA RESUMEN

TITULO: DESARROLLO DE UN MODELO PREDICTIVO PARA LA ESTIMACIÓN DEL COSTO DE SEGUROS MÉDICOS MEDIANTE ANÁLISIS DE HISTORIAL DE SINIESTRALIDAD Y SEGMENTACIÓN DE CARTERAS DE USUARIOS

1. TIPO DE PROYECTO: Aplicado
2. ÁREA DE TRABAJO: Ciencia de datos- sector salud
3. ESTUDIANTE (S): Laura Catalina Guzmán Sandoval, Xammy Alexander Victoria González
4. CORREO ELECTRÓNICO: lcguzmans2@javerianacali.edu.co, alexander3023@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO: Calle 159 No. 7-74 T3 716 Bogotá D.C. y Calle 8c #21b-101 Buga-Valle
6. DIRECTOR: Cristian Alejandro Torres Valencia
7. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
8. CORREO ELECTRÓNICO DEL DIRECTOR: cristian.torres@javerianacali.edu.co
9. CO-DIRECTOR(ES) (Si aplica):
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5):
    - Tarificación
    - Aprendizaje automático
    - Seguros médicos
    - Gestión del riesgo
    - Segmentación de riesgo
13. FECHA DE INICIO: 3/04/2025
14. FECHA DE FINALIZACION: 25/05/2025

RESUMEN

El presente proyecto aplicado aborda la problemática de la estimación del costo de los seguros médicos en el sector asegurador colombiano, considerando las limitaciones actuales en los procesos de tarificación y evaluación del riesgo. El desarrollo del proyecto tuvo como objetivo desarrollar un modelo predictivo basado en técnicas de aprendizaje automático que permitiera apoyar de manera técnica y objetiva la estimación de costos asociados a los asegurados, integrando variables demográficas, clínicas, comportamentales y de uso de servicios médicos. Para ello, se realizó la estructuración y depuración de una base de datos amplia y heterogénea, la cual requirió procesos exhaustivos de limpieza, imputación, estandarización y transformación para garantizar su calidad y consistencia. A partir de este insumo, se realizó un análisis exploratorio detallado que permitió identificar relaciones relevantes, anomalías y variables clave para la construcción del modelo. Posteriormente, se implementaron algoritmos supervisados y no supervisados como Regresión Lineal, Random Forest, XGBoost y técnicas de segmentación de usuarios, con el fin de identificar patrones de siniestralidad y mejorar la precisión en la predicción de costos. Los modelos fueron evaluados mediante métricas cuantitativas como MAE, RMSE y R^2, permitiendo seleccionar la arquitectura con mejor desempeño predictivo. Como resultado, se obtuvo una herramienta analítica orientada a fortalecer los procesos de tarificación y gestión del riesgo, además de un prototipo interactivo de visualización que facilita la interpretación de resultados y la toma de decisiones estratégicas en el sector asegurador. El proyecto aporta una aplicación práctica de la ciencia de datos y el aprendizaje automático en seguros de salud, contribuyendo al fortalecimiento técnico y financiero de las aseguradoras dentro del contexto colombiano.

TABLA DE CONTENIDO

LISTA DE ILUSTRACIONES

LISTA DE TABLAS

1. INTRODUCCIÓN

En el sector asegurador, uno de los principales desafíos operativos es la adecuada tarificación de los productos, especialmente en el ámbito de los seguros médicos. Este proceso se ve afectado por la complejidad de evaluar con precisión el perfil de riesgo de cada asegurado, lo cual puede derivar en primas inadecuadas o en niveles elevados de siniestralidad. En Colombia, esta situación se agrava por las crecientes deficiencias del sistema de salud público, que han impulsado una mayor demanda de seguros privados, haciendo aún más urgente el diseño de mecanismos técnicos que garanticen sostenibilidad y equilibrio financiero para las aseguradoras.

Ante este panorama, el proyecto desarrolló un modelo predictivo basado en técnicas de aprendizaje automático orientado a estimar el costo de los seguros médicos a partir del análisis de perfiles de usuarios y del historial de siniestralidad. Para ello, se estructuraron y depuraron bases de datos de afiliados y siniestros, integrando variables demográficas, clínicas, comportamentales y de uso de servicios médicos. Igualmente, se realizaron procesos de limpieza, transformación, codificación y normalización de la información, junto con un análisis exploratorio de datos que permitió identificar patrones, anomalías y relaciones relevantes para el modelado predictivo.

Durante el desarrollo del proyecto también se implementaron y compararon diferentes algoritmos supervisados y no supervisados, entre ellos Regresión Lineal, Random Forest, XGBoost y técnicas de segmentación de usuarios, con el propósito de mejorar la precisión en la estimación de costos y fortalecer la evaluación del riesgo. Los modelos fueron validados mediante métricas cuantitativas como MAE, RMSE y R^2, permitiendo seleccionar la arquitectura con mejor desempeño predictivo. Finalmente, se desarrolló un prototipo interactivo de visualización orientado a facilitar la interpretación de resultados y apoyar la toma de decisiones estratégicas dentro del sector asegurador colombiano.

2. CONTEXTUALIZACIÓN DEL PROYECTO

2.1 Definición del problema

2.1.1 Planteamiento del problema

En las compañías de seguros, la tarificación de los productos representa uno de los desafíos más significativos. Esto se debe a que el asegurado transfiere su riesgo a la aseguradora, exponiéndola a un posible uso excesivo de servicios o reclamaciones desproporcionadas. Si no se evalúa adecuadamente el perfil de riesgo del asegurado o se desconocen las variables estadísticas asociadas al producto, pueden generarse primas insuficientes o niveles elevados de siniestralidad.

En Colombia, los productos relacionados con la cobertura de salud han ganado mayor acogida, impulsados por las crecientes limitaciones en el acceso al sistema público de salud, caracterizado por deficiencias en la prestación de servicios y una cobertura insuficiente. Esta situación ha llevado a muchos ciudadanos a recurrir al sector privado, asumiendo costos adicionales considerablemente altos. A esto se suman desafíos estructurales en la eficiencia del sistema de salud colombiano, como los cambios demográficos asociados al envejecimiento poblacional, las dificultades de coordinación institucional y las limitaciones en la gestión y modernización del sistema de salud [2]

2.1.2 Formulación del problema

¿Cómo puede la integración de herramientas de análisis de datos mejorar los procesos de tarificación y gestión del riesgo en seguros médicos?

2.1.3 Sistematización

- ¿Qué variables son relevantes para definir el perfil de riesgo de los asegurados?
- ¿Qué técnicas de análisis de datos y aprendizaje automático permiten estimar con mayor precisión el costo de los seguros médicos?
- ¿Qué tan precisos son los modelos desarrollados para predecir el costo del seguro, según las métricas de evaluación?
- ¿Cómo puede visualizarse de forma interactiva la información generada por el modelo predictivo?

3. OBJETIVOS

3.1 Objetivo General

Desarrollar un modelo predictivo para estimar el costo de seguros médicos a partir del análisis de variables demográficas de los asegurados, con el fin de apoyar la toma de decisiones en procesos de tarificación y gestión de riesgo en el sector asegurador.

3.2. Objetivos Específicos

- Estructurar la base de datos de asegurados con variables demográficas, para generar un conjunto de datos apto para el análisis predictivo.
- Implementar algoritmos supervisados y no supervisados para modelar el costo de seguros médicos según los perfiles de los usuarios.
- Validar modelo mediante métricas de desempeño, para evaluar su precisión en la estimación del costo del seguro.
- Proponer un prototipo de visualización básico que permita interactuar con el modelo predictivo, facilitando la exploración de los resultados y la toma de decisiones.

4. MARCO TEÓRICO

4.1. Conceptos clave

Tarificación de seguros

La tarificación es el proceso mediante el cual una aseguradora determina el valor de la prima que un cliente debe pagar, con base en su nivel de riesgo. Este proceso requiere identificar y cuantificar factores de riesgo relevantes, con el fin de establecer precios que garanticen la viabilidad económica del producto y la sostenibilidad del portafolio. La adecuada gestión del riesgo y la suficiencia financiera son elementos fundamentales para la estabilidad del sector asegurador [1].

Siniestralidad

La siniestralidad es la proporción entre el valor de los siniestros pagados por una aseguradora y el valor de las primas recibidas. Altos niveles de siniestralidad pueden indicar problemas en la tarificación, selección adversa o falta de control en el uso del servicio por parte del asegurado. Predecir correctamente el comportamiento de la siniestralidad es crucial para mantener el equilibrio financiero de una aseguradora.

Aprendizaje automático (machine learning)

Es una rama de la inteligencia artificial que permite que los sistemas aprendan de los datos y mejoren sus predicciones con el tiempo sin ser explícitamente programados. Se divide principalmente en dos categorías:

- **Aprendizaje supervisado**, se define como un paradigma de la inteligencia artificial en el cual un algoritmo desarrolla una función de mapeo entre variables de entrada y una salida específica, basándose en un conjunto de datos de entrenamiento previamente etiquetados a continuación se describe algunos algoritmos existentes

    a. **Regresión Lineal Múltiple:** La regresión lineal múltiple es una aproximación paramétrica para predecir una respuesta cuantitativa Y basada en múltiples variables predictoras o características. Este modelo asume que existe una relación lineal aditiva entre las covariables observadas y la variable dependiente. Matemáticamente, la regresión lineal se expresa a través del modelo:

    $$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \epsilon$$

    Donde X_j representa el j-ésimo predictor, β_0 es el intercepto, β_j son los coeficientes que cuantifican la asociación entre cada variable y la respuesta, y ϵ representa el término de error irreductible [12].

b. **Random Forest Regressor:** Los Bosques Aleatorios son un método de ensamble (conjunto) que mejora la precisión predictiva de los árboles de decisión al generar múltiples árboles y promediar sus resultados para reducir la varianza. Para cada árbol k, se utiliza un vector aleatorio Θ_k generado independientemente de los anteriores, y se construye un predictor $h(x, \Theta_k)$. Para evitar la alta correlación entre los árboles, en cada división de los nodos del árbol se selecciona aleatoriamente un subconjunto de $m \approx \sqrt{p}$ predictores. Para la regresión, el predictor del bosque se define tomando el promedio de las predicciones independientes de B árboles:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Donde $\hat{f}_{bag}(x)$ es la predicción del b-ésimo árbol entrenado con una muestra bootstrap de los datos. El error de generalización cuadrático medio converge a medida que el número de árboles tiende a infinito: $E_{X,Y}(Y - E_{\Theta}h(X, \Theta))^2$ [13].

c. **Potenciación del Gradiente (Gradient Boosting):** A diferencia del bosque aleatorio, la potenciación del gradiente (Gradient Boosting) construye árboles de manera secuencial y aditiva, donde cada nuevo árbol intenta corregir el error o señal residual de los árboles anteriores aprendiendo "lentamente". Algoritmos avanzados y escalables como XGBoost optimizan el modelo mediante una función objetivo regularizada:

$$L(\phi) = \sum_{i} l(\hat{y}_i, y_i) + \sum_{i} \Omega(f_k)$$

Donde l es la función de pérdida diferenciable que mide la diferencia entre la predicción $\hat{y}_i$ y el objetivo y_i, y $\Omega(f_k)$ penaliza la complejidad del modelo para evitar el sobreajuste [14].

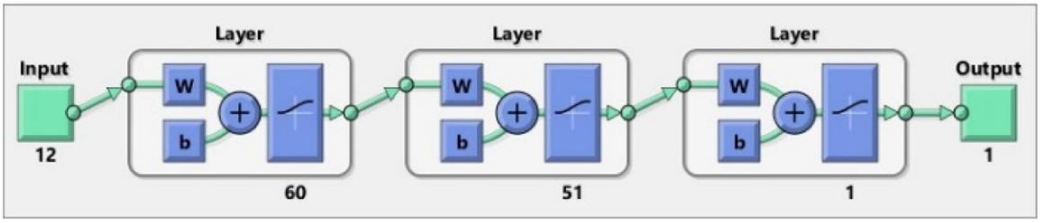
d. **Redes Neuronales Artificiales (Multilayer Perceptron - MLP)** Un Perceptrón Multicapa (MLP) es una clase de red neuronal *feed-forward* que toma un vector de entrada de variables y construye una función no lineal mediante una o varias capas ocultas de nodos (neuronas),. Las activaciones de la capa oculta se calculan como transformaciones no lineales de combinaciones lineales de las entradas. La

arquitectura matemática fundamental para una red neuronal con una única capa oculta compuesta por unidades, se define como:

$$f(x) = \beta_0 + \sum_{k=1}^{K} \beta_{kg} \left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j \right)$$

Donde los parámetros w_{kj} representan los pesos de las conexiones desde la capa de entrada a la capa oculta, β_k son los pesos hacia la capa de salida, y $g(z)$ representa una función de activación no lineal como la función Sigmoide $g(z) = \frac{1}{1+e^{-z}}$ o ReLU $g(z) = \max(0, z)$, esenciales para modelar interacciones y complejidades no lineales [15].

Figura 1. Ejemplo de Red de perceptrón multicapa (MLP)



Fuente: Tomado de Namdari y Durrani [15].

e. **Máquinas de Vectores de Soporte para Regresión (SVR)**

El objetivo de la regresión -SVR es encontrar una función $f(x) = \langle w, x \rangle + b$ que tenga como máximo una desviación $\in$ respecto a los valores reales y_i para todos los datos de entrenamiento, y que al mismo tiempo sea lo más plana posible.

La "planitud" de esta función se asegura matemáticamente buscando un vector de pesos w pequeño, lo cual se logra minimizando su norma $\frac{1}{2}\|w\|^2$. Sin embargo, dado que puede que no exista una función lineal que logre aproximar todos los pares de datos con una precisión exacta de, se introducen las variables de holgura (*slack variables*) ξ_i, ξ_i^* para tolerar errores que superen este límite, manejando así restricciones que de otra manera serían inviables. El problema de optimización convexo (primal) se formula rigurosamente como:

$$minimizar\ \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

Sujeto a las siguientes restricciones:

$$yi - \langle w, xi \rangle - b \leq \epsilon + \xi_i$$

$$\langle w, xi \rangle + b - yi \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

Donde la constante $C > 0$ sirve como parámetro de penalización que determina el *trade-off* (balance) entre la planitud de la función f y la magnitud de las desviaciones mayores a que son toleradas

Mediante el truco del kernel (*Kernel Trick*), SVR mapea los datos a un espacio de mayor dimensionalidad para separar no linealmente la información. La función de decisión del SVR resuelta mediante multiplicadores de Lagrange se expresa como:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

Donde α_i, α_i^* son los multiplicadores duales, b es el sesgo y b es la función del kernel (como un Kernel Lineal, Polinómico o de Base Radial) que cuantifica la similitud matemática entre observaciones [16].

- **Aprendizaje no supervisado**, que analiza datos sin etiquetas para encontrar patrones o agrupar elementos similares (por ejemplo, segmentar asegurados según su perfil) [3]

  a. **Agrupamiento K-Means (K-Means Clustering)** K-Means es un método de aprendizaje no supervisado diseñado para identificar y descubrir subgrupos (clústeres) ocultos dentro de un conjunto de observaciones sin etiquetas. El algoritmo busca particionar la población N-dimensional en K conjuntos distintos de tal forma que la varianza intra-clúster (la dispersión interna) sea lo más pequeña posible. La optimización de K-Means consiste en minimizar la desviación al cuadrado de cada punto respecto al centroide de su clúster, definiéndose el problema matemático como la minimización de:

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2$$

Donde C_k contiene los índices de las observaciones en el k-ésimo clúster y $|C_k|$ representa el número de observaciones en dicho clúster. El algoritmo funciona iterativamente asignando cada observación al centroide más cercano y recalculando las medias hasta alcanzar una convergencia o un óptimo local [17].

4.2. Métricas de error

Son medidas que expresan el error en las mismas unidades de la variable que se está estimando. Dado que condensar un gran conjunto de datos de error en un solo valor numérico elimina información, la selección adecuada de estas métricas es fundamental para obtener una representación precisa del comportamiento del modelo [18].

- **Error Absoluto Medio (MAE)**

  El Error Absoluto Medio (MAE, por sus siglas en inglés) es considerado en la literatura científica como la medida más natural e inequívoca para describir la magnitud del error promedio en las predicciones de un modelo. Esta métrica es en especial útil porque asigna exactamente el mismo peso a todos los errores individuales, siendo ideal para describir conjuntos de errores que presentan una distribución uniforme.

  Matemáticamente, el MAE se calcula sumando las magnitudes (valores absolutos) de los errores individuales (donde $e_i = P_i - O_i$, es decir, la diferencia entre el valor predicho P_i y el observado O_i), para luego dividir este total entre el número de observaciones. Su formulación matemática se expresa como:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

  El MAE presenta una ventaja significativa para evaluar el rendimiento promedio del modelo frente a otras métricas, ya que su valor refleja de manera directa el error y no se ve artificialmente influenciado ni sesgado por la variabilidad dentro de la distribución de las magnitudes de los errores [18]

- **Raíz del Error Cuadrático Medio (RMSE)**

  La Raíz del Error Cuadrático Medio (RMSE) es una métrica de evaluación estándar que, a diferencia del MAE, penaliza la varianza de los errores al otorgar un peso significativamente mayor a aquellos errores con valores absolutos más grandes.

  Para su cálculo, se realiza un proceso matemático de tres pasos: se eleva al cuadrado cada error individual (lo que elimina el signo negativo y magnifica los errores de gran tamaño), se promedian estos valores cuadrados, y finalmente se extrae la raíz cuadrada para devolver la medida a las dimensiones originales del problema. Matemáticamente, se define como [18],[19]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

- **Coeficiente de Determinación (R^2)**

  El Coeficiente de Determinación (R^2) es una métrica estadística que evalúa la proporción de la variabilidad de la variable dependiente que puede ser explicada por el modelo. A diferencia de las métricas de error como MAE o RMSE, el R^2 no mide directamente la magnitud del error, sino la capacidad del modelo para ajustar los datos observados.
  Un valor de R^2 cercano a 1 indica que el modelo explica gran parte de la variabilidad de los datos, mientras que valores cercanos a 0 reflejan un bajo poder explicativo. Sin embargo, es importante señalar que un R^2 elevado no garantiza necesariamente un buen modelo predictivo, ya que puede estar influenciado por problemas como el sobreajuste. [20] [21]
  Matemáticamente, el R^2 se define como:

$$R^2 = 1 - \frac{\sum i = 1n(Oi - Pi)2}{\sum i = 1n(Oi - O^{-})2}$$

  Donde:
  (Oi) son los valores observados,
  (Pi) son los valores predichos por el modelo,
  (O^{-}) es el promedio de los valores observados.

  El R^2 resulta especialmente útil para comparar modelos, ya que permite evaluar cuál de ellos logra explicar mejor la variabilidad de los datos. No obstante, debe emplearse en conjunto con métricas de error como MAE y RMSE para obtener una visión integral del desempeño del modelo

4.3. Analítica predictiva en seguros

La analítica predictiva se refiere al uso de datos históricos, algoritmos estadísticos y técnicas de machine learning para anticipar eventos futuros. En el sector asegurador, su aplicación permite estimar la probabilidad de ocurrencia de eventos como enfermedades, accidentes, o reclamaciones, y ajustar las primas de forma más precisa [2].

4.4. Visualización de datos

La visualización de datos es fundamental para comunicar resultados de modelos complejos de forma clara. Permite a los usuarios no técnicos interactuar con información crítica, facilitando decisiones estratégicas basadas en evidencia. Los dashboards y gráficos interactivos son herramientas clave para monitorear indicadores como siniestralidad, costos proyectados o agrupaciones de clientes [6].

4.5. Transformación digital en el sector asegurador

El uso de tecnologías avanzadas en la industria aseguradora ha generado un cambio significativo en la forma de analizar y gestionar el riesgo. La integración de big data, inteligencia artificial y automatización ha permitido que las aseguradoras pasen de modelos tradicionales, estáticos y basados en promedios, a esquemas dinámicos, individualizados y en tiempo real. [5]

En mercados desarrollados como el de Norteamérica y Europa, las aseguradoras han incorporado modelos predictivos para identificar clientes con alta propensión a usar el servicio, ajustar tarifas de forma individualizada y detectar patrones anómalos que podrían indicar fraude. Estos avances han contribuido a mejorar la rentabilidad sin comprometer la calidad del servicio.

4.6. Contexto y pertinencia del proyecto

El sector de salud enfrenta diversos desafíos estructurales como la fragmentación del sistema, el envejecimiento poblacional y la limitada cobertura del régimen público. La integración de big data e inteligencia artificial ha permitido a las aseguradoras mejorar los procesos de evaluación del riesgo y desarrollar modelos de tarificación más personalizados y precisos [5]

Sin embargo, muchas aseguradoras aún enfrentan dificultades para estimar adecuadamente el costo real asociado a cada cliente, debido a la falta de integración tecnológica y a una gestión limitada de los datos históricos. Este escenario evidencia la necesidad de aplicar modelos

analíticos que permitan optimizar los procesos de tarificación y gestión del riesgo, aprovechando el potencial del aprendizaje automático y de la analítica predictiva.

4.7. Antecedentes

Para fundamentar el desarrollo de este proyecto, se revisaron estudios y experiencias previas que abordan problemáticas relacionadas con la estimación del costo de seguros médicos, la segmentación de usuarios y el uso de técnicas de aprendizaje automático en el sector asegurador. A continuación, se presentan tres antecedentes relevantes, cada uno acompañado de su respectiva referencia, resumen y análisis comparativo.

[4] "Digital insurance in 2018: Driving real impact with digital and analytics.

Este informe expone cómo las aseguradoras han incorporado analítica avanzada y herramientas de inteligencia artificial en procesos como la tarificación, la suscripción de pólizas y la gestión del riesgo. Asimismo, describe cómo el uso de modelos predictivos y análisis de datos contribuye a mejorar la precisión en la evaluación de riesgos y optimizar la toma de decisiones operativas.

Aporte y diferencias con el proyecto:

Este antecedente aporta una visión estratégica del valor del aprendizaje automático en la industria aseguradora y respalda la pertinencia del uso de modelos predictivos para tareas como la estimación de costos. Sin embargo, su enfoque es principalmente global y corporativo, mientras que el presente proyecto se sitúa en un contexto académico y nacional, concentrándose en el análisis de historial médico individual y segmentación de usuarios del sistema de salud colombiano.

[11] "Stochastic Claims Reserving Methods in Insurance. Wiley"

Esta obra desarrolla modelos estadísticos orientados a la estimación de reservas para siniestros ocurridos, pero no reportados (IBNR), utilizando teoría actuarial y métodos estocásticos para el pronóstico de costos futuros en el sector asegurador. El libro destaca la importancia del modelado de datos históricos y de la evaluación cuantitativa del riesgo en procesos actuariales.

Aporte y diferencias con el proyecto:

El Libro Stochastic Claims Reserving Methods in Insurance. Wile de Wüthrich y Merz [11] se distingue por su rigor metodológico y por la incorporación de técnicas de análisis avanzado dentro de un enfoque actuarial. La obra aborda el riesgo agregado asociado a los siniestros no reportados (IBNR) desde una perspectiva institucional, lo que le confiere un enfoque técnico, colectivo y

orientado a la gestión de reservas en compañías aseguradoras. En contraste, el presente proyecto se enfoca en la predicción del costo individual de los seguros de salud, a partir del análisis del historial médico y el uso de servicios por parte de cada usuario.

[8] Salud al alcance de todos: Una década de expansión del seguro médico en Colombia. Banco Interamericano de Desarrollo.

Este artículo analiza los avances y desafíos del sistema de salud colombiano una década después de la Sentencia T-760 de 2008, considerada un hito en la garantía del derecho fundamental a la salud en Colombia. Los autores examinan aspectos relacionados con el acceso a los servicios de salud, la cobertura del aseguramiento, las barreras persistentes para los usuarios, la sostenibilidad financiera del sistema y los retos regulatorios e institucionales. Asimismo, destacan la necesidad de fortalecer los mecanismos de gestión, control y seguimiento para mejorar la eficiencia y calidad de la atención en salud.

Aporte y diferencias con el proyecto:

Este antecedente proporciona una visión integral del contexto del sistema de salud colombiano y de los desafíos asociados al acceso, la sostenibilidad y la gestión de los recursos. Además, resalta la importancia de contar con herramientas que permitan mejorar la toma de decisiones y la eficiencia en la administración del sector salud. Sin embargo, su enfoque se centra en el análisis de políticas públicas y del funcionamiento general del sistema de salud, sin abordar metodologías de analítica de datos o modelos predictivos. En contraste, el presente proyecto se enfoca en el desarrollo de un modelo predictivo para la estimación de costos de seguros médicos, utilizando técnicas de análisis de datos y aprendizaje automático para apoyar los procesos de tarificación y gestión del riesgo.

[9] Forecasting health insurance premium using machine learning approaches.

Esta investigación evaluó variables demográficas y de estilo de vida, como el índice de masa corporal y el tabaquismo, para estimar el costo de primas de seguro de salud mediante algoritmos de machine learning. Los resultados mostraron que variables como el hábito de fumar y el IMC presentaron una alta relevancia en la predicción del costo de las primas.

Aporte y diferencias con el proyecto:

Aunque este antecedente confirma la relevancia del IMC y el tabaquismo en la predicción de primas, el presente proyecto innova al superar su alcance netamente demográfico. A diferencia de Dutta [9], esta investigación incorpora una capa clínica (patologías) y financiera (gasto histórico de siniestros), además de resolver anomalías estadísticas incluyendo el contexto geográfico ("Ciudad") para adaptar los costos a la realidad de las redes médicas en Colombia. Finalmente, el

proyecto trasciende el ámbito teórico al llevar el algoritmo XGBoost a un entorno de producción mediante un prototipo web interactivo orientado al negocio.

[10] "CRISP-DM twenty years later: From data mining processes to data science trajectories"

Este artículo de investigación analiza la vigencia y efectividad de la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) en proyectos contemporáneos de aprendizaje automático. Los autores demuestran que este marco estructurado sigue siendo el estándar industrial más robusto, debido a su capacidad para alinear de forma iterativa el desarrollo técnico de los algoritmos predictivos con las necesidades operativas y estratégicas del negocio.

Aporte y diferencias con el proyecto:

Este antecedente aporta un fuerte respaldo académico e internacional, justificando científicamente la elección metodológica de este trabajo. A diferencia de la investigación de Martínez-Plumed et al., que realiza una evaluación teórica y general sobre la aplicabilidad del ciclo de vida de los datos en múltiples industrias, el presente proyecto materializa las seis fases de la metodología CRISP-DM (desde la comprensión del negocio hasta el despliegue) aplicándolas directamente para resolver una necesidad específica del sector salud colombiano: la estructuración de bases de datos médicas y la estimación del costo de seguros mediante algoritmos de *Machine Learning*.

5. MARCO METODOLÓGICO

El presente trabajo de grado se desarrolló bajo un enfoque cuantitativo y de carácter aplicado, orientado a resolver la problemática de la tarificación en el sector asegurador de salud. Para garantizar el rigor técnico y la reproducibilidad del estudio, el ciclo de vida del proyecto se estructuró siguiendo el estándar internacional y las mejores prácticas de la industria en el ámbito de la Inteligencia Artificial.

5.1. Metodología del Proyecto (CRISP-DM)

La investigación adoptó la metodología estructurada CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual proporciona un marco de trabajo robusto y comprobado para el desarrollo de proyectos de Ciencia de Datos y Machine Learning. Esta metodología iterativa aseguró que las decisiones técnicas estuvieran permanentemente alineadas con los objetivos estratégicos de la aseguradora.
El desarrollo del proyecto se ejecutó a través de las seis fases consecutivas que establece este modelo:

Comprensión del Negocio: Se definió el objetivo central del proyecto a partir de la necesidad de la aseguradora de estimar con precisión el costo de las pólizas médicas. Se identificó que la rentabilidad y la mitigación de la siniestralidad dependían de una evaluación objetiva e individualizada del perfil de riesgo de cada usuario.

Comprensión de los Datos: Se realizó un Análisis Exploratorio de Datos (EDA) exhaustivo sobre las bases de afiliados (33.458 registros iniciales) y siniestros (1.796 registros). Mediante matrices de correlación y visualizaciones de distribuciones y atípicos, se identificaron relaciones clave (como el impacto de la edad y el IMC en el costo) y se detectó la necesidad de incorporar variables latentes como la "Ciudad" para explicar discrepancias en las tarifas.

Preparación de los Datos: Se construyó un Pipeline de preprocesamiento matemático para adecuar la información a los requerimientos de los algoritmos. Esta fase incluyó la limpieza de valores nulos, el cruce de bases de datos mediante identificadores únicos (ID y Póliza), la codificación de variables categóricas (One-Hot Encoder), la estandarización de variables numéricas (RobustScaler) y la ingeniería de características (como la transformación logarítmica del siniestro y la interacción de la Edad con el IMC).

Modelado: Se determinó que el problema correspondía a una tarea de regresión continua. Se

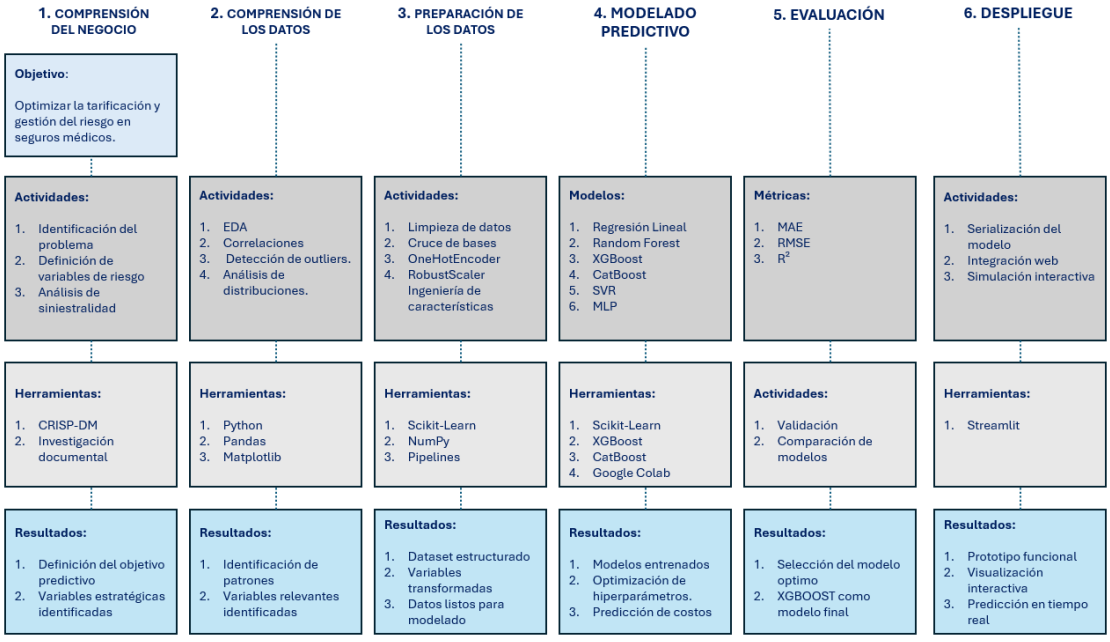
entrenaron y evaluaron paramétricamente diversos algoritmos supervisados con distintos niveles de complejidad, partiendo desde un modelo base de Regresión Lineal Múltiple, escalando hacia Redes Neuronales (MLP) y Máquinas de Vectores de Soporte (SVR), hasta llegar a arquitecturas de Ensamble de última generación (Random Forest y técnicas de Gradient Boosting).

Evaluación: Se validó la capacidad de generalización de los modelos garantizando una partición del conjunto de datos (80% entrenamiento, 20% prueba). El desempeño predictivo fue evaluado bajo un estricto criterio cuantitativo utilizando métricas estándar de error (MAE, RMSE) y ajuste (Coeficiente), lo que permitió seleccionar a la arquitectura XGBoost como el modelo definitivo.
Despliegue: La fase final consistió en trascender el código de experimentación para llevar el modelo a un entorno accesible para el negocio. Se serializó el modelo ganador y se encapsuló todo su flujo de preprocesamiento, integrándolo en una aplicación web interactiva desarrollada para simular y predecir los costos de forma dinámica en tiempo real.

5.2. Flujo metodológico y arquitectura del proyecto basado en CRISP-DM

El siguiente flujo metodológico resume las etapas desarrolladas en el proyecto de investigación, incluyendo las actividades principales, herramientas tecnológicas utilizadas y resultados obtenidos durante la construcción del modelo predictivo para estimación del costo de seguros médicos.

Figura 2. Flujo metodológico y arquitectura del proyecto basado en CRISP-DM



Fuente: Elaboración propia con base en la metodología CRISP-DM y el desarrollo del proyecto.

5.3. Arquitectura Tecnológica

Para el desarrollo computacional de todas las fases descritas, se diseñó e implementó un stack tecnológico basado en lenguajes y librerías de código abierto, garantizando la eficiencia, escalabilidad y reproducibilidad del proyecto. La arquitectura tecnológica se fundamentó en los siguientes componentes:

Lenguaje de Programación y Entorno: Todo el código fuente fue desarrollado en Python, lenguaje estándar en la ciencia de datos por su extenso ecosistema de cálculo científico. La fase de experimentación y entrenamiento fue ejecutada en entornos virtuales utilizando cuadernos de *Google Colab*.

Manipulación y Análisis de Datos: El procesamiento, limpieza y estructuración matricial de la información se realizó utilizando la librería Pandas. Para el manejo eficiente de vectores y cálculos matemáticos avanzados, como las transformaciones logarítmicas, se empleó NumPy.

Visualización de Datos: La exploración de tendencias, matrices de correlación y la explicabilidad del modelo (*Feature Importance*) fueron graficadas combinando herramientas estáticas como Matplotlib y Seaborn, así como librerías dinámicas como Plotly para el análisis temporal de la siniestralidad.

Algoritmos y Preprocesamiento: Se utilizó la librería Scikit-Learn como eje central para estructurar los *Pipelines* de preprocesamiento (ColumnTransformer, RobustScaler, OneHotEncoder). De igual forma, esta librería proveyó los estimadores para los modelos de Regresión Lineal, Redes Neuronales (MLPRegressor), Máquinas de Vectores de Soporte (SVR) y Bosques Aleatorios (RandomForestRegressor).

Modelos de Aprendizaje Avanzado: Para potenciar la predicción de relaciones no lineales, se incorporaron marcos de trabajo independientes y altamente optimizados de potenciación del gradiente (Gradient Boosting), específicamente CatBoost (optimizado para el procesamiento directo de variables categóricas) y XGBoost (algoritmo definitivo que maximizó la precisión predictiva).

Despliegue y Puesta en Producción: Una vez entrenado el modelo óptimo, se utilizó la librería **Joblib** para la persistencia del modelo, exportando los artefactos (.pkl). El control de versiones y alojamiento de los archivos se gestionó mediante un repositorio de **GitHub**. Finalmente, la interfaz visual e interactiva para el usuario final fue desplegada en la nube utilizando el framework Streamlit.

6. ESTRUCTURACIÓN DE LA BASE DE DATOS DE ASEGURADOS

6.1. Descripción general del proceso

Para el desarrollo del modelo predictivo se requirió de una base de datos estructurada que integrara la información demográfica y de uso de servicios de los asegurados. Este proceso incluyó la recopilación, depuración, codificación y normalización de los datos, garantizando la coherencia y calidad de la información.

6.2. Fuentes de información y descripción del conjunto de datos

La información utilizada en este proyecto se obtuvo a partir de registros emitidos por la aseguradora BMI Companies – BMI Compañía de Seguros Colombia S.A. que suministró la base de datos para efectos de prima y las utilizaciones que se han registrado desde el año 2020, esto para representar el perfil de afiliados a seguros médicos en el contexto colombiano. Los datos se consolidaron a partir de registros reales, los cuales fueron anonimizados para proteger la confidencialidad de la información sensible.

Esta base fue desarrollada específicamente con fines académicos y de investigación, buscando reflejar de manera estructurada los factores que inciden en el costo de una prima de seguro médico. Las variables incluidas responden a criterios técnicos y actuariales empleados en la evaluación de riesgo, como aspectos demográficos, antropométricos, hábitos de vida y antecedentes de salud.

La información fue gestionada y procesada en Google Colab, utilizando herramientas de análisis de datos y bibliotecas de Python como pandas, la cual permitió garantizar la trazabilidad, limpieza y reproducibilidad del proceso de preparación de datos.

6.3. Descripción del conjunto de datos

El conjunto de datos está conformado por registros individuales de afiliados, donde cada fila representa un asegurado y cada columna corresponde a una variable relevante para la estimación del costo de su prima médica. A continuación, se describen las principales variables incluidas en la base de afiliados:

Tabla 1. tipificación de base de afiliados

Variable	Descripción	Tipo de dato	Categoría
ID	Identificador único del afiliado, encriptado	Numérico	Identificación
Género	Sexo biológico del afiliado (Masculino / Femenino).	Categórica	Demográfica
Edad	Edad del afiliado expresada en años.	Numérica	Demográfica
Estado Civil	Condición civil del afiliado (Soltero, Casado, Unión libre, etc.).	Categórica	Demográfica
Estatura (m)	Altura del afiliado en metros.	Numérica	Antropométrica
Peso (kg)	Peso corporal en kilogramos.	Numérica	Antropométrica
IMC	Índice de Masa Corporal calculado a partir de peso y estatura.	Numérica	Antropométrica
Fumador	Indica si el afiliado consume tabaco (Sí / No).	Categórica	Estilo de vida
Actividad Física	Nivel de actividad física (Baja, Media, Alta).	Categórica	Estilo de vida
Patología	Presencia de enfermedades diagnosticadas (Ej. diabetes, hipertensión).	Categórica	Salud
Antecedente Familiar	Existencia de enfermedades hereditarias relevantes.	Categórica	Salud
Alergias	Registro de alergias reportadas por el afiliado.	Categórica	Salud
Valor Prima	Costo total de la prima del seguro médico, expresado en moneda local.	Numérica	Variable objetivo

Para el caso de la base de datos de Siniestros está conformada por registros individuales de siniestros médicos o utilizaciones de servicios de salud, en los cuales cada fila representa un evento asociado a un asegurado y su correspondiente gestión dentro de la póliza.

A continuación, se describen las variables principales incluidas en la base:

Tabla 2. tipificación de base de siniestros o utilizaciones

Variable	Descripción	Tipo de dato	Categoría
TIPO CLIENTE	Clasifica al asegurado según su rol dentro de la póliza (titular, beneficiario, dependiente, etc.).	Categórica	Administrativa
PÓLIZA	Código o número único que identifica la póliza contratada por el cliente.	Categórica	Administrativa
BENEFICIO (COBERTURA AFECTADA)	Indica el tipo de cobertura del seguro médico que fue utilizada (hospitalización, cirugía, consulta, medicamentos, etc.).	Categórica	Técnica / Cobertura
INICIO VIGENCIA	Fecha de inicio de la cobertura activa de la póliza.	Fecha	Temporal
FIN VIGENCIA	Fecha de finalización de la cobertura.	Fecha	Temporal
FECHA OCURRENCIA	Fecha en la que ocurrió el siniestro o evento médico.	Fecha	Temporal
FECHA RESERVA REGISTRO	Fecha en la que la aseguradora registró la reserva del siniestro.	Fecha	Temporal
FECHA PAGO	Fecha en la que se efectuó el pago del siniestro.	Fecha	Temporal
FECHA DE NACIMIENTO	Fecha de nacimiento del asegurado, usada para calcular la edad al momento del evento.	Fecha	Demográfica
EDAD	Edad del asegurado al momento del siniestro.	Numérica	Demográfica
SEXO	Género del asegurado (Masculino / Femenino).	Categórica	Demográfica
VALOR SINIESTRO PAGADO (COL)	Valor total pagado por la aseguradora por concepto del siniestro, expresado en pesos colombianos.	Numérica	Financiera
ESTADO SINIESTRO	Estado administrativo del siniestro	Categórica	Administrativa

CÓDIGO DE DIAGNÓSTICO	Código estandarizado que identifica la patología o diagnóstico médico (usualmente según la clasificación CIE-10).	Categórica	Médica
DESCRIPCIÓN DIAGNÓSTICO	Descripción textual del diagnóstico o enfermedad relacionada con el siniestro.	Texto	Médica

La estructura del dataset permite realizar tanto análisis descriptivos como predictivos, al combinar variables de distinta naturaleza (categóricas y numéricas) que reflejan el perfil integral del asegurado. Esta diversidad de atributos facilita la aplicación de algoritmos de aprendizaje supervisado para la predicción del costo de la prima y de aprendizaje no supervisado para la identificación de grupos con características de riesgo similares.

6.4. Limpieza y transformación de datos

El conjunto de datos original de afiliados estaba conformado por 33.458 registros y 14 atributos de naturaleza demográfica, antropométrica, de estilo de vida y de salud. En la fase de depuración, se realizó una estandarización de las cadenas de texto (homologación a minúsculas y eliminación de espacios) para variables categóricas como "Fumador" y "Antecedente Familiar".

Al evaluar la integridad de los datos, se detectó la presencia de valores nulos. Se tomó la decisión metodológica de prescindir de la imputación estadística en esta fase inicial y, en su lugar, descartar los registros incompletos, dado que la muestra restante seguía siendo altamente significativa. Esto permitió consolidar una base depurada de 6.128 registros válidos para los afiliados. Asimismo, se verificó la ausencia total de registros duplicados redundantes en este conjunto.

En paralelo, se analizó el conjunto de datos de siniestralidad, el cual constaba de 1.796 registros sin presencia de datos faltantes ni duplicados.

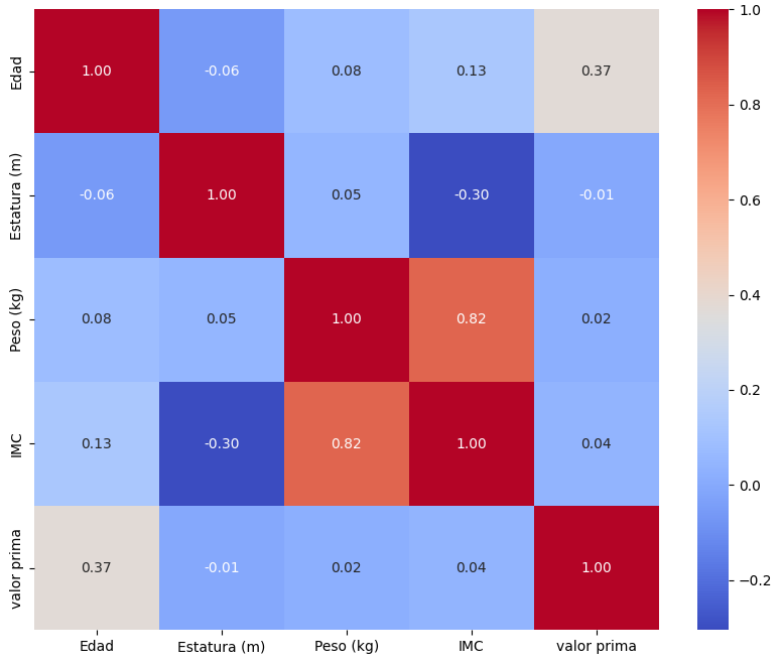
6.5. Matriz de correlación y resultados

Con el objetivo de identificar la magnitud y dirección de las relaciones lineales entre las variables continuas del conjunto de datos, y para prevenir posibles problemas de multicolinealidad en los algoritmos predictivos, se calculó la matriz de correlación de Pearson

El análisis de la **Figura 3**. evidenció hallazgos fundamentales para el modelamiento. En primer lugar, se detectó una correlación positiva de 0.37 entre la edad del asegurado y el valor de la prima. Esto ratifica actuarialmente que el envejecimiento está directamente asociado a un mayor gasto médico recurrente. Por otro lado, se identificó una altísima redundancia estadística (correlación de 0.82) entre las variables "Peso" e "Índice de Masa Corporal (IMC)". Este hallazgo

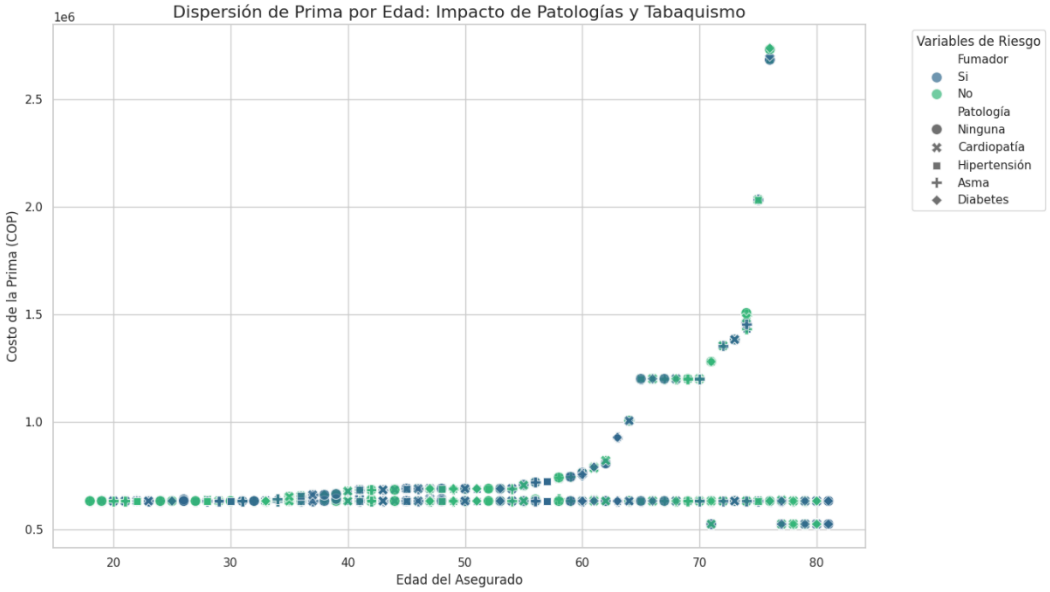
justificó la decisión metodológica de conservar únicamente el IMC como indicador estandarizado de riesgo antropométrico, descartando el peso y la estatura para evitar ruido estadístico en los modelos.

Figura 3. Matriz de correlación de variables numéricas



Fuente: Elaboración propia a partir de resultados en Python

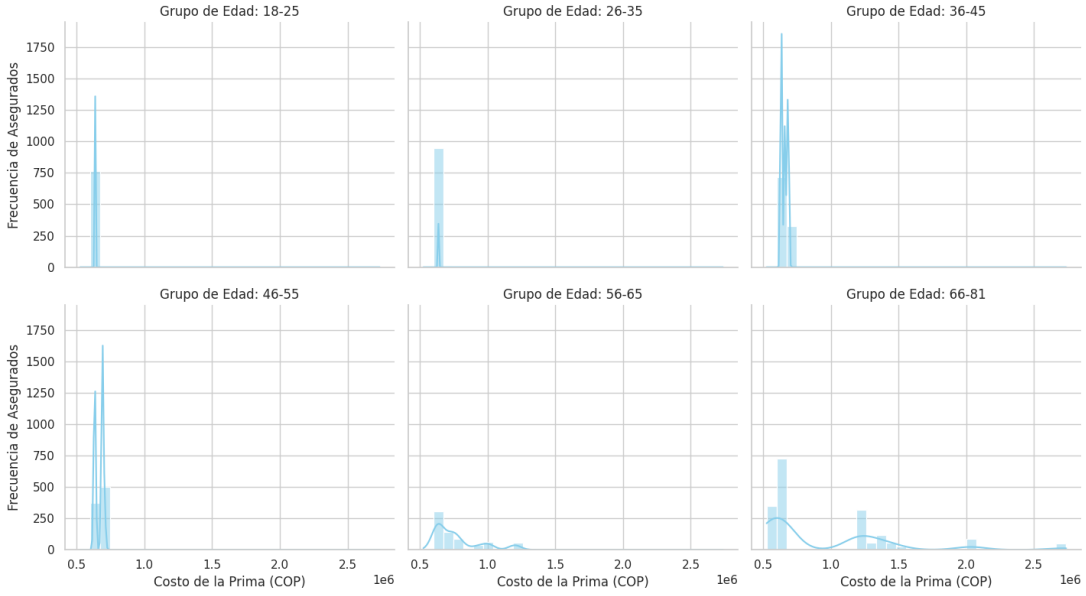
Figura 4. Dispersión de prima por edad



Fuente: Elaboración propia a partir de resultados en Python

Complementando el análisis estadístico, la **Figura 4** permite visualizar la naturaleza de la varianza que el coeficiente de correlación de 0.37 solo resume de forma agregada: una distribución del costo que, si bien asciende con la Edad, presenta una alta dispersión vertical. Se observa una notable concentración de registros en niveles de prima base (aprox. 633,502 COP) que atraviesa casi todo el espectro generacional, lo que sugiere que la edad no actúa como el único determinante del precio. Asimismo, la proyección de variables de riesgo muestra que los costos más elevados (superando los 2.5 millones de COP) se asocian a la convergencia de Patologías crónicas y el hábito de Fumador, evidenciando que el riesgo clínico no sigue una progresión lineal simple, sino que se manifiesta mediante interacciones complejas entre el estado de salud y los hábitos del asegurado.

Figura 5. Distribución de la prima por ciclo de vida del asegurado

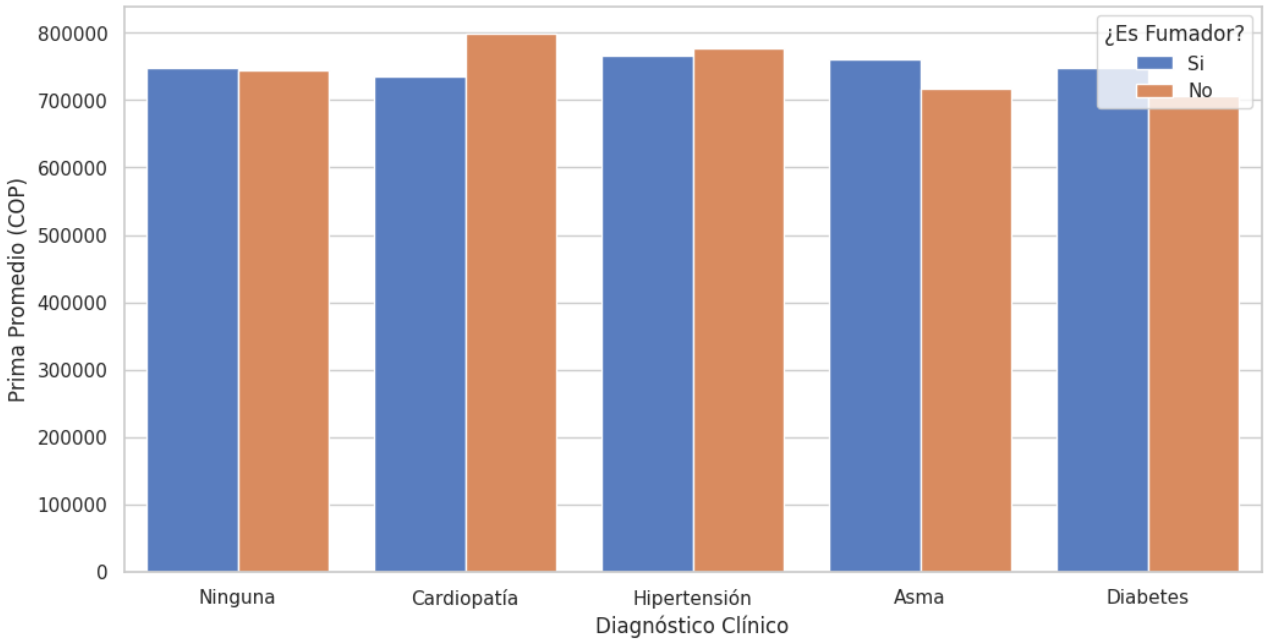


Fuente: Elaboración propia a partir de resultados en Python

La segmentación del costo de la prima por grupos de edad (**Figura 5**) revela una transformación en la morfología de la distribución que expande el análisis de correlación inicial de 0.37, evidenciando que la edad actúa como un multiplicador de la incertidumbre financiera. Mientras que en los grupos de 18-25 y 26-35 años la distribución es altamente leptocúrtica y se concentra en el valor base de 633,502 COP, indicando un riesgo clínico mínimo y uniforme en la juventud, a partir del segmento de 36-45 años aparece una bimodalidad que sugiere la influencia de variables ajenas a la edad cronológica, como hábitos de vida o diagnósticos tempranos. Finalmente, en los grupos de 56-81 años, se vuelve una distribución con curtosis negativa, desplazándose hacia la

derecha con frecuencias superiores a los 2.5 millones de COP, lo que confirma que el riesgo no sigue una progresión lineal, sino que se acelera ante la convergencia de, validando así la necesidad de emplear modelos no lineales para una tarificación técnica y objetiva.

Figura 6. Comparativa de prima promedio por patología y hábito de fumar



Fuente: Elaboración propia a partir de resultados en Python

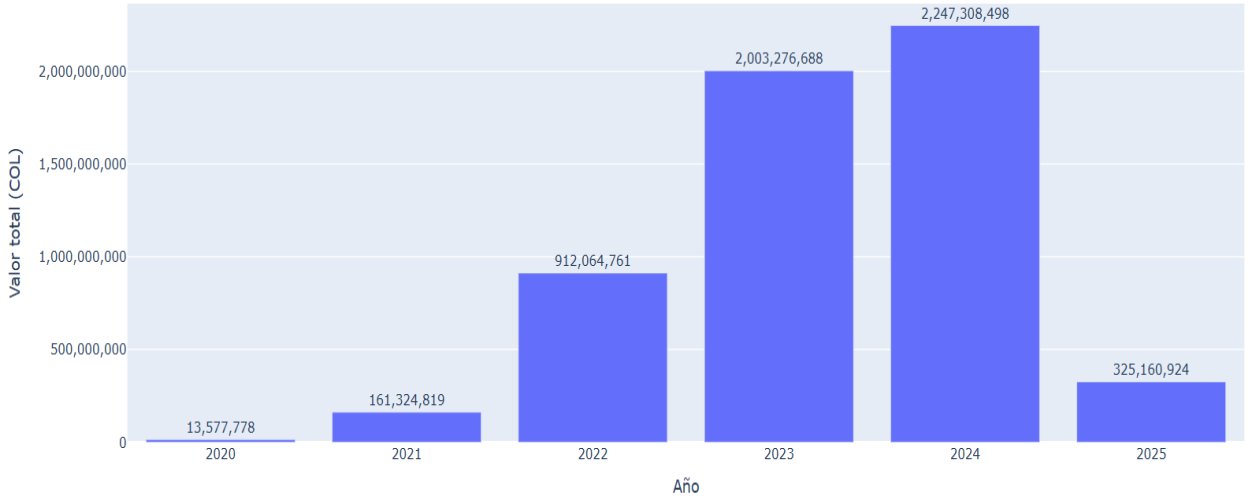
El análisis del costo promedio de la prima según el Diagnóstico Clínico y el hábito de Fumador (**Figura 6**) muestra un comportamiento relativamente uniforme, con valores que oscilan entre los 700,000 y 800,000 COP. Se observa que no existe una diferencia drástica en el promedio de la mayoría de las categorías, lo que indica que la aseguradora mantiene una base de cobro estandarizada para gran parte de su población. Sin embargo, destaca que los pacientes con Cardiopatía presentan el promedio más alto del grupo, incluso por encima de quienes fuman, lo que posiciona a esta patología como el factor de mayor peso individual en el valor de la póliza. En términos generales, la gráfica confirma que, aunque el hábito de fumar influye, es el estado de salud previo el que define los niveles superiores del costo promedio de aseguramiento.

6.6. Análisis de Siniestralidad y Diagnósticos

Asimismo, se evidenció un aumento significativo en el número de siniestros durante los años 2022 y 2023, lo que sugiere la necesidad de analizar factores contextuales propios de la compañía y del entorno nacional. Es posible que este comportamiento esté asociado a los efectos del periodo

pospandemia de COVID-19, ya sea por un incremento en la demanda de servicios médicos preventivos o por la atención de secuelas derivadas de la enfermedad. Esta tendencia resalta la importancia de incorporar variables temporales y contextuales en el análisis predictivo, con el fin de comprender mejor los cambios en los patrones de utilización y su impacto en los costos del seguro.

Figura 7. Siniestros pagados por año



Fuente: Elaboración propia a partir de resultados en Python

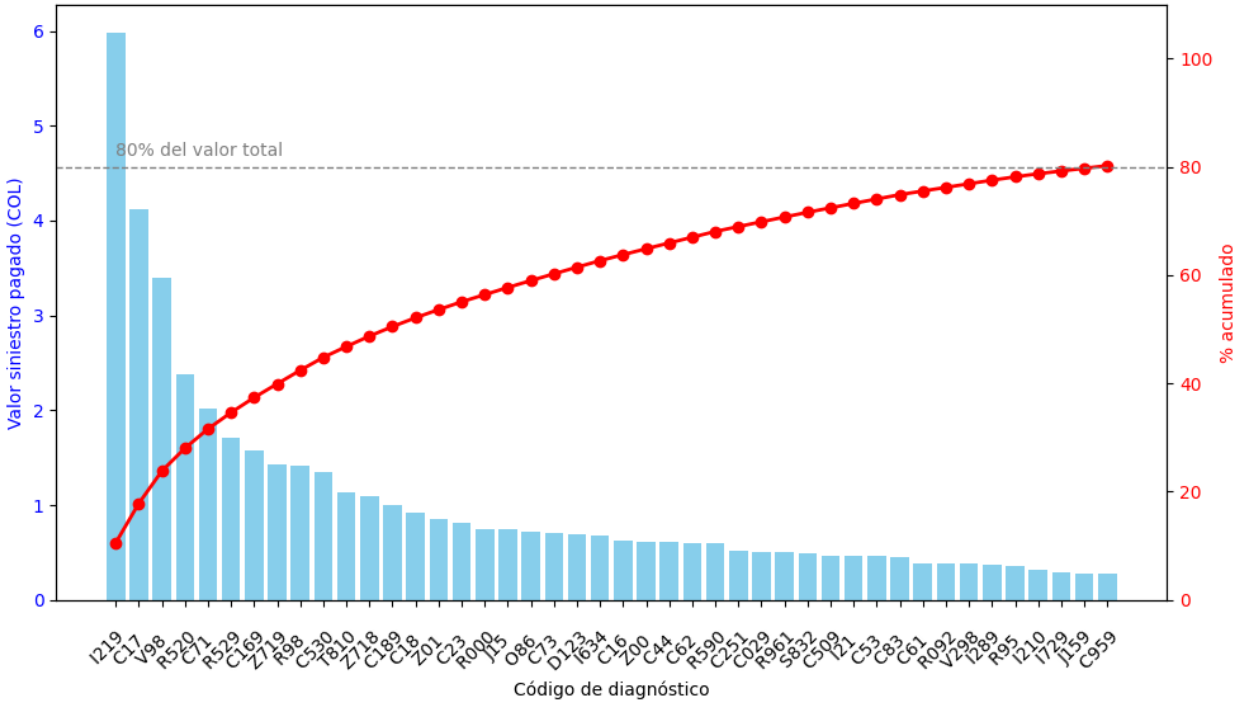
Para comprender la distribución del riesgo financiero dentro de la cartera de afiliados, se ejecutó un análisis de concentración de costos sobre la base de datos de siniestros. Se evaluó el impacto económico de cada patología agrupando los eventos médicos por su Código de Diagnóstico Internacional (CIE-10) y calculando la suma total del "Valor Siniestro Pagado COL".

Dado que el portafolio de reclamaciones médicas suele presentar una alta dispersión, se aplicó el principio de Pareto para identificar las patologías críticas que concentran la mayor carga financiera para la compañía aseguradora. El análisis reveló un patrón de alta severidad: del total de diagnósticos registrados, el modelo identificó que el 80% del valor total de los siniestros pagados se concentra en un subgrupo muy reducido de apenas 44 diagnósticos (Top 44).

Al desglosar este grupo crítico, se observó que las enfermedades crónicas, cardiovasculares y oncológicas dominan abrumadoramente el gasto. El diagnóstico de mayor impacto financiero fue el Infarto Agudo de Miocardio (Código I219), el cual representó un desembolso superior a los 597 millones de pesos colombianos para la aseguradora, seguido por patologías oncológicas severas, como el Tumor Maligno del Intestino Delgado (Código C17), con un costo acumulado superior a

los 412 millones de pesos. Otros diagnósticos que encabezan la lista incluyen el manejo del dolor agudo (R520) y tumores malignos del encéfalo (C71).

Figura 8. Diagrama de Pareto de concentración de costos por código de diagnóstico médico



Fuente: Elaboración propia a partir de resultados en Python.

La identificación de este comportamiento de Pareto es un hito analítico para la investigación, ya que justifica matemáticamente la inclusión de la variable categórica "Patología" y "Antecedente Familiar" dentro del Pipeline de Machine Learning. Queda demostrado que la presencia de morbilidades preexistentes no afecta el costo de manera lineal o equitativa, sino que expone a la compañía a siniestros atípicos de altísimo costo, los cuales deben ser capturados y penalizados algorítmicamente por modelos avanzados para garantizar la suficiencia técnica de la prima.

6.7. Identificación de anomalías estructurales

Durante la investigación se identificó una anomalía estructural mediante un proceso de validación iterativa. Pues al intentar establecer una línea base de modelado, se observó que los algoritmos no superaban un R^2 de 0.43, presentando incluso valores negativos en modelos lineales, lo que evidenció una inconsistencia profunda en los datos. Para identificar la causa, se realizó un análisis de colisiones en el espacio de características, agrupando los registros por la totalidad de sus covariables demográficas y de salud, y se detectaron múltiples casos en los que dos o más asegurados presentaban características exactamente idénticas en todas las variables observadas

(edad, género, IMC, patologías, entre otras), pero registraban valores objetivo (Valor Prima) significativamente dispares, llegando en algunos casos a representar casi el doble del costo para el mismo perfil.

Desde la perspectiva del aprendizaje automático, la presencia de entradas idénticas con salidas divergentes introduce ruido estadístico y limita la capacidad del algoritmo para converger de manera óptima. Para resolver esta discrepancia, el equipo investigador tomó dos decisiones fundamentales, alineadas tanto con la viabilidad técnica del modelo como con la estrategia del negocio asegurador:

6.7.1. Criterio de protección del negocio:

Ante la incertidumbre de tener perfiles idénticos con primas distintas, se adoptó una decisión metodológica conservadora centrada en la gestión del riesgo. Se procedió a unificar estos registros conservando únicamente el valor máximo de la prima. Esta medida se justifica desde la perspectiva de la sostenibilidad financiera de la aseguradora: subestimar el riesgo y cobrar una prima insuficiente genera pérdidas directas, mientras que estimar la cota superior protege el patrimonio de la compañía frente a posibles volatilidades en la siniestralidad.

6.7.2. Identificación e inclusión de variable latente (Ciudad):

Simultáneamente, el análisis analítico dedujo que esta severa variación en el costo de la prima no era un error de digitación, sino la consecuencia de una variable latente omitida en la matriz de datos original. Se concluyó que la estructura de costos médicos estaba fuertemente influenciada por la ubicación geográfica del asegurado, ya que el valor de los honorarios médicos, insumos e infraestructura hospitalaria varía significativamente dependiendo de la región y la red de prestadores locales.

Por consiguiente, se solicitó a la compañía aseguradora la extracción y suministro de la información correspondiente a la ubicación de los afiliados. La exitosa incorporación de la variable "Ciudad" (que categoriza a los usuarios en regiones como Bogotá, Cali, Medellín, Barranquilla, entre otras) permitió desagregar los perfiles previamente colisionados. Esta inclusión resolvió la varianza no explicada, permitiendo que los algoritmos predictivos ahora sean capaces de ajustar la tarifa de la póliza no solo según el riesgo clínico del usuario, sino también según el contexto de los costos médicos de su ciudad de residencia.

6.8. Variables Seleccionadas para el Modelamiento

Una vez concluida la fase de depuración y el Análisis Exploratorio de Datos (EDA), se llevó a cabo un proceso de selección de características (Feature Selection) orientado a construir modelos predictivos robustos y libres de multicolinealidad. El equipo investigador determinó modelar el

perfil de riesgo y costo de los asegurados utilizando un subconjunto de 10 variables críticas, clasificadas estrictamente según su naturaleza estadística de la siguiente manera:

6.8.1. Variables Numéricas

Edad: Es la variable predictora cuantitativa más determinante en la tarificación médica. El análisis de la matriz de correlación evidenció que la edad presenta la mayor correlación positiva con el valor de la prima (0.37) y una correlación moderada con el valor del siniestro pagado (0.16). Esto confirma actuarialmente que el incremento en la edad está directamente asociado a una mayor frecuencia de enfermedades crónicas y a la utilización recurrente de servicios médicos especializados, elevando el costo para la aseguradora.

Índice de Masa Corporal (IMC): Se integró como variable numérica continua para medir el riesgo antropométrico. Desde la perspectiva técnica del modelado, se decidió conservar únicamente el IMC y descartar sus variables base ("Peso" y "Estatura"). La matriz de correlación demostró una altísima redundancia estadística (0.82) entre el peso y el IMC. Suprimir el peso previene problemas de multicolinealidad en algoritmos como la Regresión Lineal, manteniendo al IMC como un indicador estandarizado suficiente para evaluar riesgos de salud.

Valor Siniestro Pagado COL: Es la variable numérica integradora fundamental, ya que cuantifica el desembolso económico real (en pesos colombianos) que la aseguradora ha efectuado por los eventos médicos del afiliado. Su inclusión mediante el cruce de las bases de datos de afiliados y siniestros proporciona a los algoritmos el valor histórico necesario para estimar y proyectar la prima esperada de forma matemática y basada en el gasto.

6.8.2. Variables Categóricas

Género: Aunque el análisis exploratorio determinó que el sexo biológico no presenta una correlación estadística sustancial con el gasto médico, su inclusión como variable nominal es un estándar técnico indispensable en la industria aseguradora para la segmentación poblacional y la identificación de riesgos específicos asociados a la siniestralidad.

Fumador: Variable binaria que captura el consumo de tabaco. Es un predictor exógeno altamente incidente en el desarrollo a largo plazo de patologías crónicas (respiratorias y cardiovasculares), permitiendo al modelo penalizar el riesgo del asegurado objetivamente.

Actividad Física: Variable categórica de tipo ordinal que evalúa el nivel de sedentarismo del usuario (ej. sedentario, moderado, activo). Actúa como un factor comportamental que mitiga o incrementa el nivel de riesgo de salud general del afiliado.

Patología: Variable nominal que señala la presencia de enfermedades diagnosticadas preexistentes (como diabetes, hipertensión o cardiopatías). Refleja la morbilidad base del usuario y tiene un impacto directo sobre la probabilidad de reclamaciones y utilizaciones médicas futuras.

Antecedente Familiar: Variable nominal que modela la existencia de enfermedades hereditarias (cáncer, diabetes, etc.). Aporta al modelo la capacidad de prever el riesgo genético latente que puede desencadenar altos costos a largo plazo.

Alergias: Variable nominal que registra condiciones alérgicas del asegurado (medicamentos, polen, lácteos, etc.), contribuyendo a perfilar la complejidad clínica que el afiliado pueda presentar al momento de utilizar los servicios médicos.

Ciudad: Identificación e inclusión de variable latente (Ciudad): Se clasifica como categórica porque agrupa a los afiliados en zonas geográficas discretas (como Bogotá, Cali, Medellín, Barranquilla, etc.). Su función dentro del modelo es actuar como un factor de ajuste geográfico de costos. Aunque dos pacientes tengan exactamente el mismo perfil clínico y demográfico, la categoría "Ciudad" le permite al algoritmo comprender y ponderar las diferencias económicas regionales, ya que los honorarios de la red de especialistas, los costos hospitalarios y las tarifas de insumos médicos varían significativamente según la ubicación geográfica del asegurado.

7. MODELADO

7.1. Regresión Lineal Múltiple (Modelo Base)

Como punto de partida, se implementó un modelo paramétrico de regresión lineal. Matemáticamente, el modelo asume una relación lineal aditiva entre las covariables del asegurado (X) y la prima (Y), representada por la ecuación:

Resultados: Para garantizar la correcta evaluación y reproducibilidad del algoritmo, el conjunto de datos se dividió destinando el 80% de los registros para la fase de entrenamiento y el 20% restante como conjunto de prueba (test), fijando una semilla de aleatoriedad (random state) de 42. Se utilizó la implementación estándar del algoritmo (LinearRegression).

El modelo arrojó un ajuste deficiente (R^2 de 0.1670), lo que justificó la transición hacia algoritmos de mayor complejidad.

7.2. Random Forest Regressor

Para superar las limitaciones lineales, se implementó un ensamble de árboles de decisión basado en la técnica de Bagging. La predicción final del modelo es el promedio de las predicciones independientes de B árboles de regresión.

Al igual que en el modelo base, la validación se estructuró mediante una partición del 80% de los datos para entrenamiento y el 20% para pruebas, utilizando el random state de 42. La arquitectura del bosque se configuró con 400 árboles independientes y, para controlar la varianza, se exigió un mínimo de 50 muestras para dividir un nodo intermedio y al menos 3 muestras en cada nodo hoja final. Además, en cada división solo se evaluó un subconjunto de variables correspondiente a la raíz cuadrada del total de características. Este modelo alcanzó un R^2 de 0.6966.

7.3. Algoritmos de Potenciación del Gradiente (Gradient Boosting)

A diferencia de los bosques aleatorios, la potenciación del gradiente construye árboles de manera secuencial, donde cada nuevo árbol intenta corregir el error residual del ensamble anterior.

Para evaluar esta arquitectura, se implementaron dos algoritmos del estado del arte, ambos entrenados bajo el mismo esquema de validación cruzada simple: 80% de los datos para

entrenamiento y 20% para la prueba de generalización, manteniendo la semilla de aleatoriedad en 42:

7.4. CatBoost Regressor

Diseñado para manejar variables categóricas de forma eficiente. Su entrenamiento inicial se programó para 2000 iteraciones con una tasa de aprendizaje de 0.03 y una profundidad máxima de 6 niveles por árbol. Para evitar el sobreajuste, el entrenamiento se detuvo prematuramente conservando los primeros 432 árboles iterados. Logró un R^2 de 0.7457.

7.5. XGBoost Regressor

Fue el modelo de mayor rendimiento. Su estructura ensambló 500 árboles de decisión con una profundidad de 6 niveles. Para estabilizar las hojas, se exigió un peso mínimo en los nodos hijos de 3. El modelo integró un submuestreo estocástico, utilizando el 80% de los datos y el 80% de las columnas por cada árbol, apoyado por términos de regularización L1 ($\alpha = 0.1$) y L2 ($\lambda = 1.0$). Esta configuración maximizó el R^2 a 0.7827, convirtiéndose en el modelo definitivo.

7.6. Redes Neuronales Artificiales (Multilayer Perceptron - MLP)

Siguiendo la rigurosidad metodológica del estudio, se segmentó la matriz de diseño reservando un 80% para la fase de aprendizaje de la red neuronal y un 20% para validación de predicciones, con un random state estandarizado en 42. Mediante una búsqueda aleatoria de hiperparámetros (RandomizedSearchCV), la topología óptima seleccionada constó de 3 capas ocultas (128, 128 y 64 neuronas). La propagación de la señal utilizó una activación de tangente hiperbólica y los pesos se actualizaron con el optimizador Adam. Se aplicó una regularización L2 con un de 0.001199, logrando un R^2 de 0.6379.

7.7. Máquinas de Vectores de Soporte para Regresión (SVR)

Finalmente, se exploró el mapeo del conjunto de datos a un espacio de características de mayor dimensionalidad para encontrar el hiperplano óptimo de regresión. Matemáticamente, SVR busca una función $f(x)$ que tenga como máximo una desviación ϵ de los valores reales de la prima (y_i), minimizando a su vez la norma de los coeficientes de los pesos para evitar la complejidad excesiva del modelo.

Este algoritmo también fue sometido a la partición de datos estándar de la investigación, utilizando el 80% de los registros para el ajuste del modelo y el 20% para evaluar el error de predicción, con la semilla random state de 42. Se configuró el algoritmo SVR utilizando un kernel radial de base gaussiana (kernel="rbf"), aplicando un coeficiente de penalización C igual a 50 y definiendo un margen de tolerancia hiper-tubular ($\in$) de 0.05, garantizando que los errores dentro de ese margen no fuesen penalizados. En este algoritmo se logró un R^2 de 0.5083.

7.8. Modelos no Supervisado K-means

Para abordar el objetivo de segmentar a los usuarios en perfiles de riesgo, se implementó un enfoque de aprendizaje no supervisado, el cual permite analizar los datos sin etiquetas predefinidas para encontrar patrones matemáticos ocultos y agrupar elementos con características similares. En el contexto de este proyecto, esta técnica resulta fundamental para identificar nichos poblacionales (clústeres) que comparten un nivel de riesgo homogéneo, fortaleciendo la capacidad analítica del sistema asegurador.

Selección de Variables y Preprocesamiento: Para ejecutar el algoritmo de agrupamiento (Clustering), se seleccionó un subconjunto integral de características demográficas, clínicas y de estilo de vida, compuesto por las variables: Edad, IMC, Género, Fumador, Actividad Física, Patología, Antecedente Familiar y Ciudad.

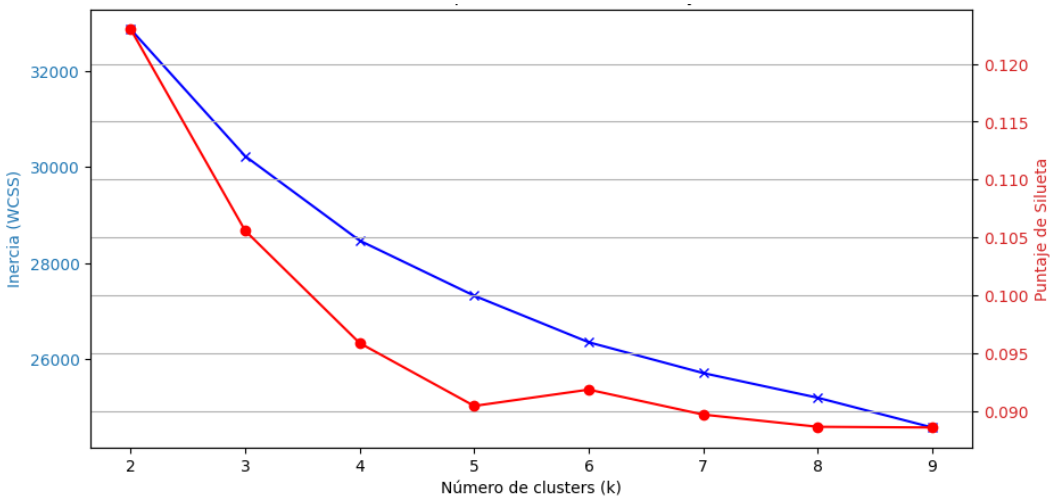
Dado que el algoritmo K-Means se basa en el cálculo de distancias y es altamente sensible a la escala en la que se encuentran los datos, se hizo imperativa la construcción de un Pipeline de preprocesamiento. Este consistió en aplicar una estandarización de varianza (StandardScaler) a las variables numéricas continuas (Edad e IMC) para llevarlas a una misma escala matemática, y una codificación binaria (One-Hot Encoder) para vectorizar los atributos categóricos de los pacientes.

Optimización y Selección del Número de Clústeres (K): Se implementó el algoritmo K-Means evaluando iterativamente un rango desde 2 hasta 9 posibles agrupaciones (K). Para evitar asignaciones empíricas o arbitrarias, el número óptimo de clústeres se definió midiendo el desempeño del algoritmo a través de dos métricas de validación simultáneas:

Inercia (WCSS - Within-Cluster Sum of Squares): Se calculó para medir la suma de las distancias al cuadrado de cada punto respecto al centroide de su clúster, evaluando así la compacidad interna de los grupos creados.

Puntaje de Silueta (Silhouette Score): Se utilizó para evaluar de forma simultánea la cohesión interna frente a la separación externa, determinando qué tan bien asignado está cada perfil de riesgo a su propio clúster.

Figura 9. Selección del número de clústeres mediante Inercia y Puntaje de Silueta



Fuente: Elaboración propia a partir de resultados en Python.

En la Figura 9. Se demostró gráficamente que la curva de inercia (WCSS) experimenta un punto de inflexión o "codo" de estabilización, el cual coincide con el momento en que el puntaje de silueta alcanza una estabilización óptima. A partir de este comportamiento matemático, se determinó algorítmicamente que la partición ideal de la base de datos consolida 4 grupos de riesgo distintos.

La identificación de estos 4 clústeres confirma que la cartera asegurada presenta nichos claramente diferenciados por sus hábitos y comorbilidades. La implementación de este modelo no supervisado dota a la compañía aseguradora de una herramienta capaz de clasificar automáticamente a los usuarios en segmentos estructurados, lo cual resulta indispensable para apoyar el diseño de estrategias comerciales, planes preventivos de salud a la medida y una gestión del riesgo mucho más precisa.

Una vez determinada la partición óptima de la cartera en cuatro segmentos (k=4), se procedió a extraer y analizar los centroides de cada clúster para interpretar su composición demográfica y financiera. La Tabla 3 consolida los valores promedio de la Edad, el Índice de Masa Corporal (IMC) y el costo de la prima para cada agrupación, junto con la distribución del número de afiliados (n). A partir de estos promedios matemáticos, fue posible perfilar el nivel de riesgo de cada segmento

de la siguiente manera:

Tabla 3. Caracterización de los perfiles de riesgo obtenidos mediante K-Means (k=4).

Clúster / Perfil de Riesgo	Cantidad de Afiliados (n)	Edad Promedio (Años)	IMC Promedio	Valor Prima Promedio (COP)
0: Riesgo Latente (Jóvenes con Obesidad)	1453	35,9	36,4	$ 654.753,22
1: Riesgo Crítico (Adultos Mayores con Obesidad)	1137	69,3	38,7	$ 877.846,95
2: Bajo Riesgo (Jóvenes Saludables)	1846	32,6	22,7	$ 646.651,25
3: Riesgo Moderado-Alto (Adultos Mayores Peso Normal)	1782	65,9	24,6	$ 856.421,81

Fuente: Elaboración propia a partir de los resultados experimentales en Python.

Clúster 2: Perfil de Bajo Riesgo (Jóvenes Saludables)

- **Características:** Edad promedio de 32.6 años e IMC de 22.6 (peso normal).
- **Impacto financiero:** Presentan el costo promedio de prima más bajo de la cartera ($646.651 COP) y representan el segmento demográfico más grande, con 1.846 afiliados.
- **Interpretación:** Es el segmento más rentable y estable para la aseguradora. Su juventud y ausencia de factores de riesgo antropométricos mantienen la siniestralidad al mínimo. La estrategia comercial para este grupo debe enfocarse en la retención y la fidelización temprana.

Clúster 0: Perfil de Riesgo Latente (Jóvenes con Obesidad)

- **Características:** Edad promedio de 35.9 años e IMC de 36.4 (obesidad tipo II/III).
- **Impacto financiero:** Prima promedio de **$**654.753 COP, agrupando a 1.453 afiliados.
- **Interpretación:** Aunque su prima actual es relativamente baja debido a su juventud biológica, su alto índice de masa corporal los convierte en una "bomba de tiempo" actuarial. Este clúster tiene una alta probabilidad de desarrollar comorbilidades crónicas

a mediano plazo (como diabetes o hipertensión). Representan el nicho ideal para que la aseguradora implemente programas proactivos de prevención nutricional y acondicionamiento físico.

Clúster 3: Perfil de Riesgo Moderado-Alto (Adultos Mayores con Peso Normal)

- **Características:** Edad promedio de 65.8 años e IMC de 24.5 (peso normal).
- **Impacto financiero:** Prima promedio alta de $856.421 COP, abarcando a 1.782 afiliados.
- **Interpretación:** En este segmento, el encarecimiento de la póliza obedece estrictamente al desgaste biológico natural (edad) y no a malos hábitos de vida, ya que mantienen un IMC óptimo. Este hallazgo valida el análisis de variables del modelo supervisado (*XGBoost*), el cual demostró que el envejecimiento por sí solo dispara los costos de atención médica.

Clúster 1: Perfil de Riesgo Crítico (Adultos Mayores con Obesidad)

- **Características:** Edad promedio de 69.3 años e IMC de 38.7 (obesidad severa).
- **Impacto financiero:** Generan la prima promedio más alta del portafolio ($877.846 COP), conformando el grupo más reducido con 1.137 afiliados.
- **Interpretación:** Es el segmento que concentra la mayor severidad y carga financiera para la compañía. La convergencia simultánea de una edad avanzada y una obesidad mórbida maximiza la probabilidad de siniestros catastróficos u hospitalizaciones recurrentes. Para este clúster, la aseguradora debe aplicar políticas de gestión de riesgo estrictas, provisionar mayores reservas técnicas y asignar planes de atención médica domiciliaria y control crónico riguroso.

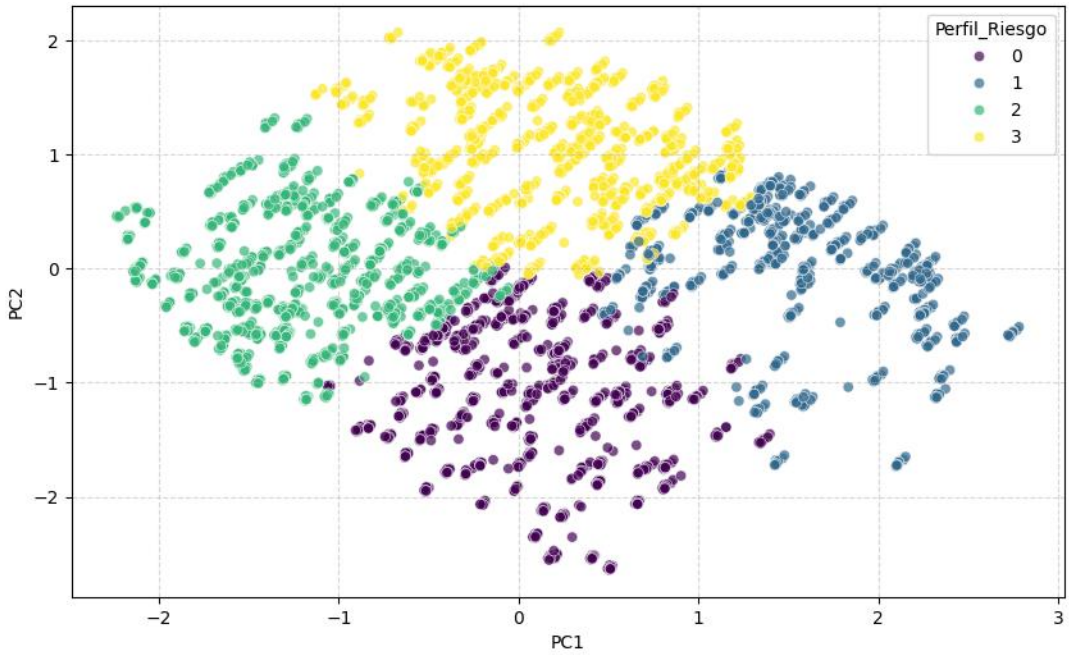
La caracterización de estos cuatro perfiles comprueba que el aprendizaje no supervisado es una herramienta estratégica de alto valor. Al cruzar los clústeres resultantes, se observa que una persona joven con obesidad (Clúster 0) paga casi lo mismo que un joven sano (Clúster 2), pero al llegar a la tercera edad, la diferencia entre mantener un peso normal (Clúster 3) o padecer obesidad (Clúster 1) encarece significativamente el riesgo. Esta segmentación le permite a la compañía aseguradora transitar de una tarificación masiva a una gestión de carteras hiper-personalizada.

Para complementar la caracterización cuantitativa de los perfiles de riesgo y ofrecer una representación visual interpretable de la segmentación, se implementó una técnica de reducción de dimensionalidad conocida como **Análisis de Componentes Principales (PCA - *Principal***

Component Analysis).

Dado que el algoritmo K-Means agrupó a los asegurados evaluando simultáneamente múltiples dimensiones (edad, IMC, variables demográficas y clínicas), resulta matemáticamente imposible proyectar los clústeres resultantes en su espacio multidimensional original. Para solucionar este desafío, el algoritmo PCA transformó y comprimió las variables originales en dos nuevos ejes ortogonales, denominados Componente Principal 1 (PC1) y Componente Principal 2 (PC2), los cuales retienen la mayor cantidad de varianza y estructura de los datos originales.

Figura 10. Agrupación perfiles de riesgo obtenida mediante clustering y proyectada con PCA



Fuente: Elaboración propia a partir de los resultados experimentales en Python.

Como se evidencia en la **Figura 10**, la proyección bidimensional corrobora gráficamente el éxito del modelo de aprendizaje no supervisado. Cada punto en el plano representa a un asegurado, y los colores indican el perfil de riesgo (Clúster 0, 1, 2 y 3) asignado por K-Means. La gráfica muestra una agrupación espacial altamente coherente y contigua, lo que demuestra que el algoritmo logró establecer fronteras de decisión claras entre los distintos fenotipos poblacionales.

Esta visualización confirma que los perfiles no se traslapan de forma caótica, sino que obedecen a patrones latentes fuertemente estructurados en la base de datos, justificando la transición de la aseguradora hacia un modelo de gestión y tarificación hiper-segmentado.

8. EVALUACIÓN DEL MODELO

Para dar cumplimiento al objetivo de validar los modelos desarrollados y evaluar su precisión en la estimación del costo del seguro médico, se estableció un marco de evaluación cuantitativo riguroso. Siguiendo las buenas prácticas en la validación de algoritmos de Machine Learning, el desempeño de todos los modelos se calculó exclusivamente sobre el conjunto de prueba aislado (correspondiente al 20% de los datos), garantizando la capacidad de generalización sobre información no vista durante la fase de entrenamiento.

La evaluación cuantitativa se fundamentó en tres métricas estadísticas estándar para problemas de regresión continua:

- Error Absoluto Medio (MAE)
- Raíz del Error Cuadrático Medio (RMSE)
- Coeficiente de Determinación (R^2)

8.1. Análisis Comparativo de Resultados

La evaluación de los cinco enfoques algorítmicos evidenció una evolución progresiva en la precisión predictiva a medida que se implementaron arquitecturas con mayor capacidad para procesar relaciones complejas y no lineales. En la **Tabla 4** se consolidan los resultados de desempeño obtenidos sobre el conjunto de validación:

Tabla 4. Comparativa de métricas de desempeño de los algoritmos evaluados.

Modelo Predictivo	MAE (COP)	RMSE (COP)	Coeficiente R2
Regresión Lineal Múltiple	$ 168.924,10	$ 301.773,19	0.1670
Redes Neuronales (MLP Regressor)	$ 99.079,49	$ 222.468,02	0.6379
Random Forest	$ 91.887,61	$ 203.583,75	0.6966
CatBoost Regressor	$ 82.446,96	$ 186.367,44	0.7457
XGBoost Regressor (Seleccionado)	**$ 70.312,81**	**$ 172.262,72**	**0.7827**

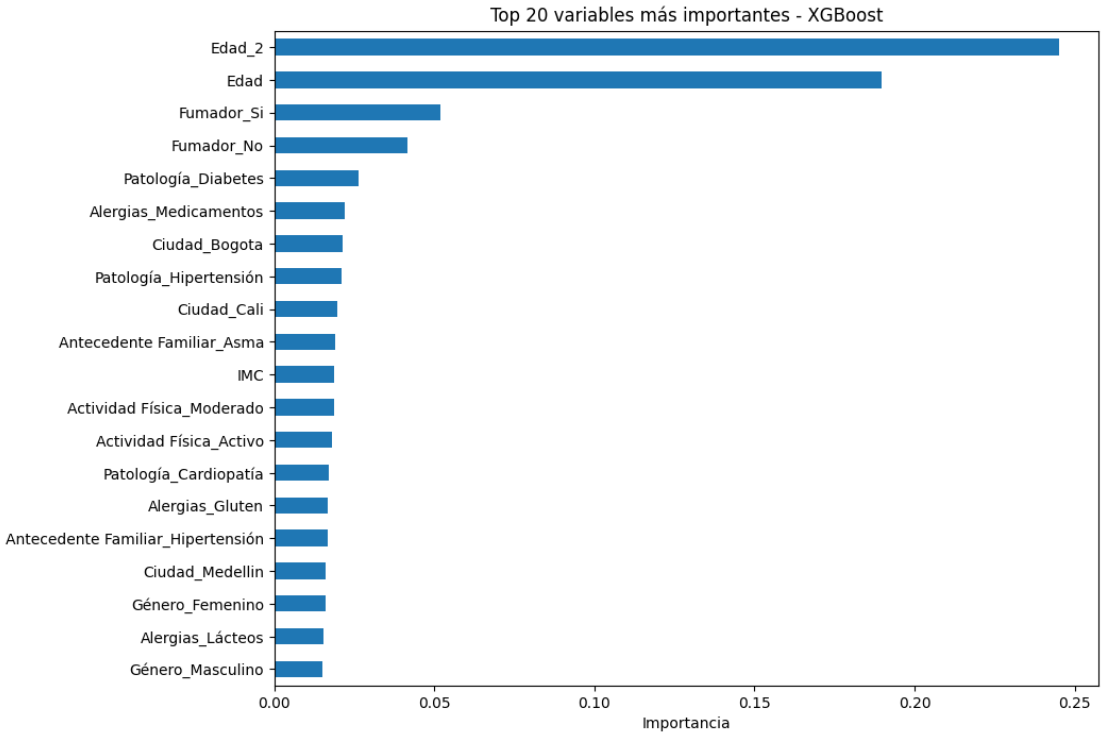
Fuente: Elaboración propia a partir de los resultados experimentales en Python.

Como se observa en la Tabla 4, el modelo paramétrico de Regresión Lineal exhibió el desempeño predictivo más bajo, ratificando que el problema de tarificación médica requiere enfoques analíticos más sofisticados. La implementación de Redes Neuronales (MLP) y Bosques Aleatorios (Random Forest) redujo significativamente los márgenes de error, superando la barrera del 60% de varianza explicada.

Sin embargo, fueron las técnicas de potenciación del gradiente (Gradient Boosting) las que demostraron el mejor ajuste. El algoritmo XGBoost se consolidó como la arquitectura predictiva superior de la investigación, presentando la mayor capacidad de generalización al explicar el 78,27% de la varianza del costo del seguro ($R^2 = 0.7827$).

Uno de los principales desafíos en la implementación de modelos avanzados de Machine Learning es la percepción de que actúan como una "caja negra", lo cual dificulta su adopción en sectores altamente regulados como el asegurador. Para garantizar la transparencia técnica y justificar los criterios de tarificación, se extrajo la importancia de las características (Feature Importance) del algoritmo XGBoost. Esta métrica cuantifica el peso matemático y la contribución relativa de cada variable de entrada en la estimación final de la prima.

Figura 11. Top 20 variables más importantes del modelo XGBoost



Fuente: Elaboración propia a partir de resultados en Python.

Como se evidencia en la **Figura 11**, las decisiones del algoritmo son altamente interpretables y guardan una profunda coherencia con los principios actuariales y clínicos de la industria de la salud. El análisis del modelo revela tres grandes hallazgos en torno a sus predictores más fuertes:

La no linealidad del envejecimiento (Edad_2 y Edad): Las dos variables con mayor poder predictivo absoluto son la Edad al cuadrado (Edad_2) y la Edad cronológica estándar. El hecho de

que la transformación polinómica (Edad_2) encabece la lista, superando a la variable original, demuestra que el incremento de los costos médicos no es proporcional ni lineal a medida que el afiliado envejece. Por el contrario, el riesgo de presentar siniestros de alto costo crece de manera acelerada en las etapas adultas avanzadas, lo cual justifica y valida metodológicamente la fase de ingeniería de características (Feature Engineering) realizada previo al modelado.

Penalización de hábitos de riesgo (Fumador_Si): Excluyendo el factor demográfico de la edad, la condición de ser fumador (Fumador_Si) se posiciona como el tercer predictor más importante del modelo, seguido de su contraparte (Fumador_No). Esto indica que el algoritmo aprendió a identificar de forma autónoma que el tabaquismo es el factor de estilo de vida que más encarece el costo proyectado de una póliza, debido a su estrecha relación con el desarrollo futuro de patologías crónicas respiratorias y cardiovasculares severas.

El impacto de las morbilidades preexistentes: En posiciones subsecuentes, el modelo asigna gran relevancia a condiciones clínicas específicas como la diabetes y la hipertensión, así como a las variaciones geográficas capturadas por la ciudad de residencia.

En conclusión, este análisis de importancia de características demuestra que la arquitectura XGBoost implementada no es una "caja negra". Por el contrario, el modelo evalúa de manera lógica, transparente y objetiva el riesgo de cada usuario, penalizando matemáticamente los factores controlables como el tabaquismo y ajustando la tarifa según el ciclo de vida del asegurado. Esta explicación dota a la compañía aseguradora de una herramienta confiable para fundamentar sus políticas de precios.

9. DESARROLLO DE PROTOTIPO DE VISUALIZACIÓN

Se diseñó e implementó una aplicación web orientada a facilitar la exploración de resultados y la toma de decisiones. El desarrollo de esta herramienta buscó cerrar la brecha entre el modelo matemático complejo y los usuarios finales, permitiendo a perfiles no técnicos (como analistas de riesgo o equipos de tarificación) interactuar con la información crítica de manera intuitiva.

El proceso de construcción y despliegue del prototipo se dividió en las siguientes fases metodológicas:

9.1. Serialización y encapsulamiento del modelo

Tras validar que la arquitectura XGBoost presentaba el mejor rendimiento predictivo, la primera fase consistió en la serialización computacional del modelo para su puesta en producción. Utilizando la librería joblib de Python, se exportó el Pipeline completo bajo el archivo pipeline_modelo_primas.pkl. Esta práctica metodológica es fundamental, ya que el archivo encapsula no solo el estimador final, sino todos los transformadores de preprocesamiento (imputación, estandarización RobustScaler y codificación One-Hot Encoder) asegurando que los nuevos datos ingresados en el prototipo reciban exactamente el mismo tratamiento matemático que los datos de entrenamiento.

De manera complementaria, se exportó la lista estructurada de las características (features_list.pkl) para garantizar matemáticamente que el vector de entrada conserve la dimensionalidad y el orden exacto de las variables requeridas por el algoritmo al momento de predecir.

9.2. Integración y despliegue del entorno web

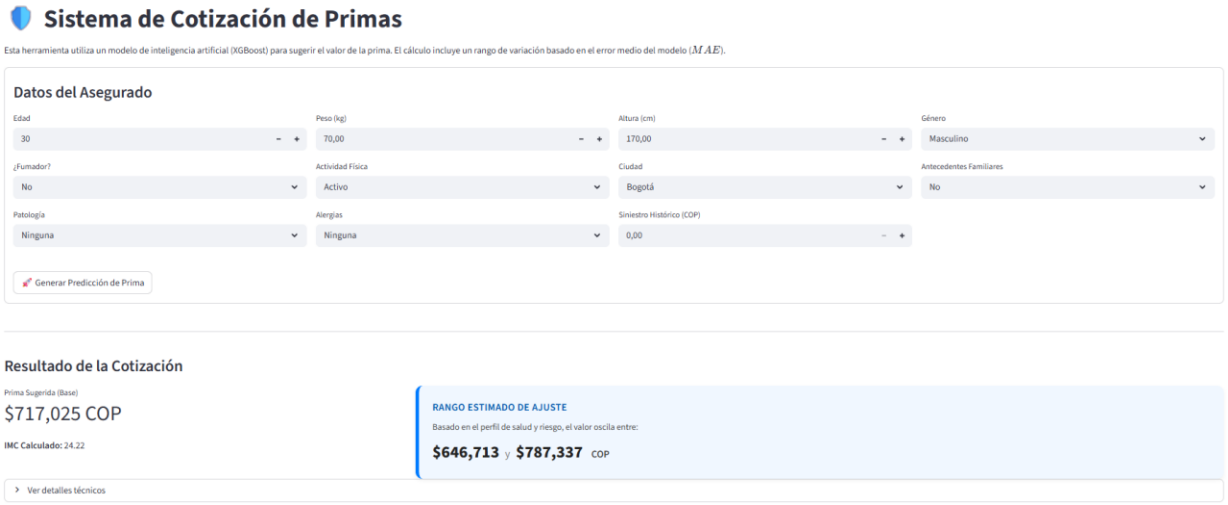
Para la construcción de la interfaz gráfica, los modelos serializados y los documentos operativos fueron versionados y alojados en un repositorio en GitHub. A partir de este repositorio, se desarrolló el frontend interactivo utilizando Streamlit, un framework de código abierto en Python diseñado específicamente para desplegar aplicaciones de Machine Learning y Ciencia de Datos.

El prototipo funcional se encuentra desplegado y accesible en la nube (a través del dominio modeloreseguros-z6tshrttjmj5m9fa5agglw.streamlit.app). La interfaz permite al usuario ingresar de forma ágil y estructurada los parámetros del perfil de un asegurado, tales como su edad, antropometría (peso y altura para calcular dinámicamente el IMC), ubicación geográfica (Ciudad), hábitos de vida y antecedentes clínicos (Patologías y Alergias).

Una vez el usuario ingresa las variables, la plataforma ejecuta el Pipeline en el backend, procesa la información y retorna en tiempo real el valor predicho de la prima comercial. En la **Figura 12**.

Se puede evidenciar la interfaz del prototipo.

Figura 12. Interfaz del Prototipo de visualización



Fuente: Elaboración Propia

Al permitir la alteración interactiva de las variables de entrada, la herramienta funciona como un simulador de escenarios, mediante el cual la aseguradora puede evaluar de forma visual cómo los diferentes factores de riesgo (por ejemplo, el consumo de tabaco o la presencia de una comorbilidad) impactan directamente en el costo estimado del seguro. Esto materializa el valor estratégico del Machine Learning, transformando los datos en decisiones informadas, equitativas y basadas en evidencia para la gestión del riesgo de la compañía.

10. CONCLUSIONES Y TRABAJOS FUTUROS

10.1. Conclusiones

Se concluye que la calidad predictiva en el sector asegurador depende directamente de una estructuración rigurosa de los datos y del descubrimiento de variables latentes. Por un lado, el exhaustivo proceso de depuración (que redujo 33.458 registros iniciales a una base consolidada de 6.128 registros válidos) y la ingeniería de características (como el cálculo del IMC) demostraron que los perfiles médicos básicos son insuficientes por sí solos para estimar riesgos. Por otro lado, durante el modelado se realizó un hallazgo metodológico clave: la existencia de perfiles demográficos idénticos con primas significativamente distintas. Esto permitió identificar que los costos médicos están fuertemente condicionados por la ubicación geográfica del asegurado. La incorporación de la variable 'Ciudad' fue decisiva para eliminar el ruido estadístico, comprobando que el contexto regional de la red de salud es un predictor fundamental para garantizar una tarificación justa y precisa.

La implementación comparativa de múltiples algoritmos comprobó que la relación entre el perfil de salud de un afiliado y el costo de su póliza médica posee una naturaleza altamente compleja y no lineal. El deficiente desempeño del modelo paramétrico base (Regresión Lineal Múltiple), el cual solo alcanzó un ajuste del 16,70%, evidenció que los métodos tradicionales aditivos son insuficientes para modelar este fenómeno. Por el contrario, se concluye que los algoritmos de Ensamble basados en árboles de decisión, particularmente aquellos construidos bajo la arquitectura de Potenciación del Gradiente (Gradient Boosting como CatBoost y XGBoost), poseen la robustez matemática necesaria para capturar las interacciones jerárquicas y no lineales entre variables demográficas, clínicas y de estilo de vida de los asegurados.

La evaluación cuantitativa del desempeño predictivo permitió validar que el algoritmo XGBoost Regressor es la solución técnica óptima para la estimación de primas, superando significativamente a las Redes Neuronales (MLP) y Máquinas de Vectores de Soporte (SVR). Se concluye que su aplicación mitiga drásticamente el riesgo financiero de la aseguradora al explicar el 78,27% de la varianza poblacional () y lograr una reducción del Error Absoluto Medio a $70.312,81 COP. La validación técnica sobre datos aislados demuestra que es factible trasladar el riesgo médico a un entorno algorítmico confiable, logrando que la estimación del costo del seguro sea una tarea precisa y basada en el gasto real histórico de siniestros, protegiendo así el patrimonio de la compañía.

Finalmente, se concluye que el valor de los modelos matemáticos complejos se maximiza únicamente cuando son accesibles para la toma de decisiones empresariales. El desarrollo y despliegue en la nube (Streamlit) de un prototipo interactivo comprobó ser una herramienta estratégica que cierra la brecha entre la ciencia de datos y los usuarios de negocio (como analistas de riesgo). Al encapsular el Pipeline completo (preprocesamiento y predicción), el prototipo funciona exitosamente como un simulador de escenarios ("What-If"), permitiendo evaluar en tiempo real y de forma visual cómo las variaciones en el estilo de vida o morbilidad de un afiliado impactan el costo proyectado, promoviendo una tarificación personalizada, ágil y transparente.

10.2. Trabajos Futuros

Como extensión natural de los hallazgos y desarrollos obtenidos en la presente investigación, se proponen las siguientes líneas de trabajo futuro orientadas a escalar la capacidad analítica, tecnológica y de negocio de la compañía aseguradora:

Enriquecimiento del Conjunto de Datos y Variables Macroeconómicas: Aunque el modelo actual logró un alto nivel de precisión () utilizando variables demográficas, clínicas y geográficas, se recomienda enriquecer la matriz de datos con variables externas. A nivel socioeconómico, la inclusión del estrato o nivel de ingresos del afiliado podría afinar la predicción del tipo de red hospitalaria que este preferirá utilizar. Asimismo, dado que los costos de salud fluctúan con el tiempo (como se evidenció en el análisis temporal de siniestros de 2020 a 2025), se sugiere integrar un índice de inflación médica anual para que el modelo ajuste námicamente sus predicciones a la realidad macroeconómica del país.

Transición hacia una Arquitectura MLOps (Machine Learning Operations): El prototipo funcional desplegado en Streamlit cumple a cabalidad el objetivo de validación visual e interactiva. No obstante, para un entorno de producción masiva, se recomienda la transición del sistema hacia una arquitectura MLOps en la nube (ej. AWS, Google Cloud o Azure). Esto implica automatizar los flujos de ingesta de datos de nuevos afiliados y establecer un sistema de monitoreo de Data Drift (deriva de datos). Dado que los costos médicos y los perfiles epidemiológicos cambian, este monitoreo alertará automáticamente a la aseguradora cuando el modelo XGBoost comience a perder precisión, disparando un proceso de reentrenamiento autónomo.

Integración de Técnicas de Explicabilidad de Modelos (XAI - Explainable AI): Aunque el algoritmo XGBoost Regressor demostró ser altamente preciso, los modelos de ensamble de árboles suelen ser considerados "cajas negras" debido a su complejidad interna. En el sector asegurador, la transparencia es un requisito fundamental tanto por ética profesional como por cumplimiento regulatorio. Por lo tanto, se sugiere integrar en el prototipo interactivo librerías de explicabilidad de Inteligencia Artificial, como SHAP (SHapley Additive exPlanations) o LIME. Esto permitiría que

el dashboard no solo entregue el valor numérico de la prima predicha, sino que desglose visualmente y explique al analista (o al cliente) exactamente qué variables aumentaron o redujeron la tarifa en cada caso particular. Esta mejora incrementaría significativamente la confianza, la auditabilidad y la adopción de la herramienta por parte de los tomadores de decisiones.

11. REFERENCIAS BIBLIOGRÁFICAS

[1] A. L. Glassman, M. L. Escobar, A. Giuffrida, and U. Giedion, Eds., *Salud al alcance de todos: Una década de expansión del seguro médico en Colombia*. Washington, DC, USA: Banco Interamericano de Desarrollo, 2010.

[2] J. D. Cummins and M. A. Weiss, "Systemic Risk and the U.S. Insurance Sector," *Journal of Risk and Insurance*, vol. 81, no. 3, pp. 489–528, 2014.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[4] McKinsey & Company, *Digital Insurance in 2018: Driving Real Impact with Digital and Analytics*. McKinsey & Company, 2018.

[5] OECD, *The Impact of Big Data and AI on the Insurance Sector*. Paris, France: OECD Publishing, 2020.

[6] S. Few, *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*, 2nd ed. Burlingame, CA, USA: Analytics Press, 2013.

[7] McKinsey & Company, *Digital Insurance in 2018: Driving Real Impact with Digital and Analytics*. McKinsey & Company, 2018.

[8] Restrepo-Zea, J. H., & Silva-Maya, C. (2018). ¿Qué ha pasado en Colombia después de diez años de la Sentencia T-760 de 2008 Revista de Salud Pública, 20(6), 670-676.

[9] S. Dutta et al., "Forecasting health insurance premium using machine learning approaches," *Applied Science and Technology Program*, 2023.

[10] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach, "CRISP-DM twenty years later: From data mining processes to data science trajectories," in *Knowledge Discovery and Data Mining*, M. Ceci et al., Eds. Cham, Switzerland: Springer, 2019, pp. 273–292.

[11] M. V. Wüthrich and M. Merz, *Stochastic Claims Reserving Methods in Insurance*. Chichester, UK: John Wiley & Sons, 2008. ISBN: 978-0-470-72346-3.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2nd ed. New York, NY, USA: Springer, 2021.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16)*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.

[15] A. Namdari and T. S. Durrani, "A multilayer feedforward perceptron model in neural networks for predicting stock market short-term trends," *Operations Research Forum*, vol. 2, no. 38, 2021.

[16] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[17] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.
[18] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Climate Research*, vol. 30, pp. 79–82, 2005.
[19] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” *Geoscientific Model Development*, vol. 7, pp. 1247–1250, 2014.
[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken, NJ, USA: Wiley, 2012.
[21] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, 4th ed. Chicago, IL, USA: Irwin, 1996.