UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

MAESTRÍA EN ANALÍTICA DE DATOS

FACULTAD DE INGENIERÍA INDUSTRIAL

MODELOS PREDICTIVOS

PROYECTO FINAL

DESARROLLO DE UN MODELO DE ANÁLISIS PREDICTIVO PARA EL COSTO DE SEGUROS MÉDICOS BASADO EN LOS ANCEDENTES MÉDICOS Y DEMOGRÁFICOS DEL ASEGURADO

GRUPO 1AN-216

PROFESOR: Juan Marcos Castillo, PhD

NOMBRE: Juan Torres

CÉDULA: E-8-160393

2025

1. INTRODUCCIÓN

El costo promedio de un día en el hospital en Estados Unidos, sin un seguro médico es de aproximadamente $13,600. Este monto puede variar mucho según el tipo de tratamiento que se necesite, del hospital y el estado del paciente (Calderón, 2025).

Un seguro médico nos garantiza seguridad en nuestro día a día. El saber que contamos con un respaldo en caso de emergencia para nosotros y nuestros seres queridos nos ofrece algo invaluable: tranquilidad. Sin embargo, hoy en día el costo de estos seguros es cada vez más alto, varía de un momento a otro, muchas veces llegando a ser insostenible dentro de nuestra economía personal. La realidad es que hay muchos factores que pueden influir en esto, nuestra edad, el estilo de vida que llevamos, nuestro historial médico, entre otros. Y todo eso, lo tienen que tomar en cuenta las aseguradoras, para asegurarse de que podrán brindarnos la ayuda que necesitamos, pero manteniendo un negocio rentable.

WorldRemit Ltd (2023) nos explica que, en Estados Unidos, el manejo de los costos del sistema de salud depende de programas estatales, federales y de compañías de seguros privadas. Existen planes de salud, amparados por el gobierno, que van dirigidos a poblaciones vulnerables como personas de escasos recursos, con discapacidad o de avanzada edad. Pero, lastimosamente, la realidad es que muchos ciudadanos no tienen acceso a un seguro privado, debido a que no cuentan con los recursos (Bupa Global, 2025).

A continuación, vamos a analizar el impacto de factores clave como el tabaquismo, el género, la cantidad de hijos, género, la región y la edad, en el costo médico que representa una persona para las aseguradoras.

1.1. Tabaquismo

Hay investigaciones que demuestran que el tabaquismo representa entre el 8% y el 11.7% del gasto de salud anual en Estados Unidos. Esto representa, aproximadamente, entre 170 y 226 mil millones al año (McCann, 2020). Estamos hablando de que un fumador puede llegar a representar un gasto de entre seis mil y siete mil dólares en gastos de salud al año, en comparación con un no fumador (Vapor Technology Asociation, 2025).

Está demostrado que los fumadores tienen una esperanza de vida mucho menor, ya que son más susceptibles a enfermedades como bronquitis, cáncer pulmonar y problemas cardíacos (Seguros del Pichincha, s.f.).

1.2. Índice de Masa Corporal (BMI)

De acuerdo con el Instituto Texas Heart (2025), el índice de masa corporal (IMC), que se calcula dividiendo los kilogramos de peso por el cuadrado de la estatura en metros.

Según el Instituto Nacional del Corazón, los Pulmones y la Sangre de los Estados Unidos (NHLBI), el sobrepeso se define como un IMC de más de 25. Se considera que una persona es obesa si su IMC es superior a 30.

Los costos médicos para personas con sobrepeso u obesidad tienden a ser mucho más altos. En Estados Unidos, los costos en atención médica debido a la obesidad alcanzan los 173 mil millones de dólares al año. Una persona con obesidad puede representar el doble que una persona de peso saludable en gastos médicos (Telesford & Schwartz, 2024).

Quantum Pro (2023) nos dice que el índice de masa corporal es utilizado por las aseguradoras como un indicador de salud. Entre más alto es, puede indicar un mayor riesgo de desarrollar enfermedades crónicas como la diabetes, ciertos tipos de cáncer o enfermedades cardíacas.

1.3. Número de Hijos

Las familias que tienen niños se enfrentan a gastos médicos más elevados, lo que las hace un grupo de alto costo para las compañías de seguros, sobre todo por el uso de servicios para niños. De acuerdo con una investigación, estas familias pueden gastar en promedio $1,300 por una hospitalización de un niño. Además, una de cada siete hospitalizaciones puede costar más de $3,000, incluso teniendo un seguro privado (Carlton, Scott, Moniz, & Prescott, 2023).

1.4. Género

Las investigaciones sugieren que las mujeres suelen tener un costo más elevado para las empresas de seguros médicos, hasta un 42% más que los hombres, debido factores biológicos, sociales y demográficos En Estados Unidos, el gasto médico de las mujeres es más alto que el de los hombres, incluso si se excluyen los gastos relacionados con la maternidad (Palacio, 2025).. Además, tienen una mayor esperanza de vida mayor, lo que significa que necesitan atención médica durante más años, y son más propensas a sufrir enfermedades con tratamientos costosos como el cáncer de mama, la osteoporosis, la artritis y diversas condiciones ginecológicas (Plenilunia Salud Mujer, 2025).

Asimismo, requieren chequeos médicos más frecuentes, como mamografías, exámenes de Papanicolaou y controles hormonales, lo que aumenta la utilización de servicios de salud y, por lo tanto, el costo del seguro (Aflac, 2025).

1.5. Región

La cantidad de dinero que invierten las compañías de seguros de salud en Estados Unidos para cada individuo varía dependiendo de la región en la que se encuentren. Existen varias explicaciones para esto: las tarifas de los hospitales locales, la disponibilidad de médicos, entre otros factores (NIHCM Foundation, 2025). Así lo comprueba la información publicada por Centers for Medicare & Medicaid Services (2025). Para este estudio nos enfocamos en las regiones: **noreste**, **noroeste**, **sureste** y **suroeste**.

a. Noreste

En la región noreste del país, que incluye estados como Nueva York o Massachusetts, los gastos tienden a ser altos debido a la presencia de grandes hospitales con tecnología avanzada, salarios más altos para el personal médico y una mayor cantidad de personas mayores que suelen requerir más atención médica.

b. Noroeste

Por otro lado, en el noroeste, en lugares como Utah o Idaho, los costos tienden a ser más bajos porque hay menos hospitales grandes, un menor uso de servicios y precios de atención más accesibles.

c. Sureste

En el sureste, los gastos también pueden ser bajos, ya que los costos de atención son más reducidos y hay menor acceso a programas públicos, lo que disminuye el gasto directo del seguro privado. Ahora, cabe resaltar que en esta región predominan problemas de salud como la obesidad o la diabetes.

d. Suroeste

La región suroeste es muy variada, ya que incluye estados como Texas y Arizona, donde también prevalecen costos más bajos, debido a una población más joven, y se les da un menor uso a los servicios más especializados. Sin embargo, aquí también podemos encontrar grandes ciudades como Houston o Dallas, donde los costos pueden ser más altos.

1.6. Edad

La Fundación NIHCM (2025) nos explica cómo y por qué la edad juega un papel importante en el costo del seguro que representa una persona. Los jóvenes (menores de 35 años) normalmente son personas sanas y requieren pocos servicios médicos, por lo que su costo de seguro suele ser menor. Conforman una gran parte de la población, pero solo un pequeño porcentaje del gasto en salud.

Sin embargo, a medida que las personas envejecen, tienden a desarrollar más problemas de salud crónicos como la hipertensión, el colesterol elevado o la diabetes, y la necesidad de servicios médicos aumenta, especialmente entre los adultos de 50 a 64 años.

Los mayores de 65 años, aunque representan solo alrededor del 17% de la población, acaparan casi el 40% del gasto total de la salud. Esto se debe a que utilizan con mayor frecuencia hospitales, medicamentos y tratamientos complejos.

Por eso, para el seguro médico, una persona mayor representa un costo mucho mayor que una persona joven, aun cuando ambos paguen primas similares.

2. OBJETIVO

El objetivo principal del proyecto será desarrollar un modelo predictivo que permita calcular de forma precisa el costo médico que representa un individuo para una aseguradora, conociendo sus antecedentes y características.

3. METODOLOGÍA

El modelo predictivo se desarrolló aplicando un enfoque cuantitativo basado en aprendizaje automático. El proceso incluyó una exploración y limpieza inicial del *dataset*, un análisis estadístico descriptivo de los datos y las variables y el entrenamiento, prueba y comparación de diferentes modelos.

La mayor parte del proyecto, se llevó a cabo por medio de la herramienta de Jupyter Notebook, utilizando el lenguaje de programación Python, apoyándonos en programas como Microsoft Excel.

4. JUSTIFICACIÓN

El desarrollo de modelos predictivos para costos de seguros médicos ha adquirido importancia en la industria médica, especialmente en naciones que cuentan con sistemas de seguros privados como es el caso de Estados Unidos. La habilidad de anticipar estos gastos con exactitud ayuda a las compañías de seguros a ajustar sus tarifas y a los clientes a comprender qué aspectos tienen mayor influencia en sus costos de salud (Fei, 2023).

El sistema de salud Estados Unidos es uno los más costosos del mundo (MSH International, 2022). El aumento en los gastos de la atención médica supone un reto para las compañías de seguros y para los servicios de salud, tanto públicos como privados.

Estados Unidos gasta cerca del doble en atención a la salud que otros países ricos porque todo, desde los medicamentos, los equipos y los salarios del personal médico, es más caro (SWI, 2018). Por eso, es fundamental que las compañías de seguros puedan prever y estimar cuánto podrían necesitar gastar en la atención médica de cada cliente, no solo para establecer tarifas más justas y viables, sino también para crear planes de prevención, gestión de riesgos y personalización de seguros.

En el caso de los pacientes, también es importante que logren llevar un control de sus gastos de salud, las aseguradoras no siempre cubrirán todo, habrá casos donde se tendrán que asumir ciertos costos. Conocer de antemano qué elementos incrementan esos costos les ayudaría a tomar decisiones más conscientes sobre su modo de vida y su planificación financiera.

En este sentido, un modelo de predicción basado en las características y antecedentes del individuo, puede ser una herramienta muy útil para que ambas partes puedan llevar un control de sus gastos médicos.

Además, conocer los factores que inciden directamente en los costos médicos de los seguros en Estados Unidos puede brindarnos información útil que podemos aplicar y utilizar como guía para un análisis similar en Panamá.

5. ANTECEDENTES

De acuerdo con Fei (2023), el conjunto de datos **Medical Cost Personal Datasets de Kaggle** se ha establecido como una fuente valiosa, ya que incluye variables importantes tanto demográficas como médicas de 1,338 personas. Estudios previos (Djebali, 2025) con este dataset han demostrado que:

- Los fumadores tienen costos médicos mucho mayores que los no fumadores en todas las regiones.
- El IMC contribuye significativamente al aumento de costos
- Las variables como región y sexo muestran un impacto mínimo en comparación.

Se han utilizado modelos de regresión líneal, transformaciones logarítmicas e ingeniería de características para el estudio del dataset, identificando que existe una menor precisión para casos de costos muy elevados, mayores a $50,000, y no siempre capturan relaciones no lineales complejas.

Sin embargo, existen oportunidades que no se han desarrollado a profundidad anteriormente. Por ejemplo, la mayoría de los modelos no han explorado exhaustivamente algoritmos avanzados como Random Forest o Gradient Boosting.

6. DEFINICIÓN DEL PROBLEMA

En la actualidad, las compañías de seguros enfrentan el reto de estimar de manera precisa los costos asociados a la salud de cada persona, tomando en cuenta sus antecedentes personales y de salud. Esta evaluación es crucial tanto para la gestión financiera de las empresas de seguros como para distribuir las primas de manera justa entre los asegurados.

El problema de investigación que se plantea consiste en desarrollar un modelo predictivo capaz de estimar el costo del seguro médico (*charges*) a partir de datos demográficos y de salud, como: la edad, el índice de masa corporal, la cantidad de hijos, el género, el hábito de fumar y la ubicación geográfica. El objetivo es evaluar el desempeño de distintos modelos para identificar cuál ofrece la mejor precisión en la predicción.

Este desafío se aborda utilizando métodos de análisis estadístico, visualización de datos y machine learning, considerando distintos enfoques de modelado, desde regresión lineal hasta modelos más avanzados como XGBoost y Gradient Boosting.

Como ya se presentó, existen muchos aspectos que pueden influir en el costo que representa un individuo para las aseguradoras, en términos de salud. Sin embargo, muchos de ellos tienen su causa raíz en factores demográficos y de estilo de vida. Es por eso que para abordar el problema, se utilizará la información del *dataset*: **Medical Cost Personal *Dataset***, ya que cuenta con información muy útil

en ese sentido. El *dataset* se descargó de la plataforma Kaggle. con el nombre *insurance.csv*.

7. ANÁLISIS PREDICTIVO

7.1. Determinación de la Base de Datos

El proyecto estará basado en el conjunto de datos: Medical Cost Personal Dataset, que se obtuvo de la plataforma Kaggle con el nombre insurance.csv. Este conjunto de datos incluye información sobre 1,338 individuos, que viven en Estados Unidos y cuentan con un seguro médico privado. Se tiene registro de su información, incluyendo características personales y de su estilo de vida, así como el monto en dólares que gastaron en su seguro de salud (variable objetivo).

Este conjunto de datos fue elegido por su estructura clara y ordenada, que incluye tanto variables numéricas como categóricas, perfectas para llevar a cabo un análisis estadístico y predictivo completo. Además, su tamaño moderado permite explorarlo de manera eficiente utilizando Python como lenguaje de programación, especialmente a través de Jupyter Notebook.

7.1.1. Descripción del *Dataset*

El *dataset* contiene información demográfica, médica y económica para 1,338 beneficiarios de seguro médico residentes en los Estados Unidos. Las columnas que contiene son las siguientes:

- **age:** Edad del beneficiario principal.
- **sex:** Género del asegurado (masculino o femenino)
- **bmi:** Índice de Masa Corporal (Body Mass Index)
- **children:** Número de hijos o dependientes cubiertos por el seguro médico.
- **smoker:** Indicador de si la persona es fumadora (Sí/No)
- **region:** Región geográfica de residencia dentro de Estados Unidos.
- **charges:** Costo estimado que representa el individuo para el seguro.

A partir de esta información, se espera poder desarrollar un modelo predictivo que nos permita calcular el costo que representa una persona para una compañía de seguros médicos. Pero antes, se debe evaluar la calidad y utilidad del *dataset*.

7.1.2. Exploración Inicial del *Dataset*

Comenzamos verificando el tamaño del *dataset*, confirmando que estaremos trabajando con 7 columnas: age, sex, bmi, children, smoker, region y charges, cada una con 1,338 registros. Ninguna de ellas tenía valores faltantes. Trabajamos con dos tipos de variables:

- **Variables Numéricas**: *age, children, bmi* y *charges*. Es importante recordar que charges es nuestra variable objetivo.
- **Variables Categóricas**: *smoker, región, sex*. Estas tendrán que ser codificadas para poder ser analizadas e incluidas dentro del modelo.

7.2. Análisis Descriptivo

7.2.1. Variables Numéricas

a. Estadísticas Descriptivas

Se calcularon estadísticas descriptivas para las variables numéricas, obteniendo específicamente valores como la media, la mediana, la desviación estándar y rango intercuartílico. El resumen de estas medidas se encuentra en la Tabla 1.

	Media	Desviación Estándar	min	25%	50%	75%	max
age	39.21	14.05	18.00	27.00	39.00	51.00	64.00
bmi	30.66	6.10	15.96	26.30	30.40	34.69	53.13
children	1.09	1.21	0	0	1	2	5.00
charges	13,270.42	12,110.01	1,121.87	4,740.29	9,382.03	16,639.91	63,770.43

Tabla 1: Estadísticas Descriptivas para las variables numéricas

A partir de estos datos, podemos sacar varias conclusiones iniciales. Por ejemplo:

- **Edad**: Estamos trabajando únicamente con adultos. Se obtuvieron los datos de personas entre 18 y 64 años, con una edad promedio de 39 años.

- **Hijos**: Tenemos un máximo de 5 hijos por familia. Podemos concluir que hay muchos individuos sin hijos.
- **BMI**: En general, estamos trabajando con personas que sufren de obesidad, ya que tenemos un índice promedio de 30.66. Incluso tenemos casos severos, con un valor máximo de 53.13.

- **Costo:** El costo que representa una persona para las compañías de seguro varía entre $1,121.87 y el máximo es $63,770.43. Considerando que el promedio es de 13,270.42, podríamos decir que el *dataset* incluye varios valores atípicos (outliers) para el costo. Algunos valores extremadamente altos en comparación a la media.

b. Moda:

Con respecto al **género**, la mayoría de los registros son de **hombres**. Si analizamos la **región**, vemos que predomina el **sureste**. Los datos del **bmi** nos dicen muchos sufren de **obesidad (bmi > 30)**. Y en cuanto a la **edad,** estamos trabajando en su mayoría con jóvenes de **18 años**.

c. IQR (Rango Intercuartílico)

Podemos ver que la variable edad tiene un IQR de 24.0, lo que indica que el 50% de las edades registradas rondan en un rango de 24 años, ya que el IQR evalúa la variabilidad de los valores entre el percentil 25% y el percentil 75%. Por otra parte, podemos afirmar que el bmi presenta una dispersión moderada, con IQR de 8.397.

El IQR de la variable children es 2, con lo cual podemos concluir que la mitad de las personas registradas tienen entre 1 y 3 hijos. Una distribución bastante concentrada.

Por último, la variable costos muestra un IQR de 11,899.62. Esto quiere decir que algunos representan costos muy bajos para las aseguradoras y otros, un costo muy alto, y la diferencia gira alrededor de los 11,900.

Esto comprueba que existe una gran variabilidad en los costos de salud y puede indicar una **distribución asimétrica**. Esto se ve más claro en la Figura 1, donde nos lo confirma su histograma de frecuencia.

Ahora, también se determinó que el coeficiente de variación de charges es 0.913, lo que indica una alta dispersión relativa respecto a su media. Esto indica que los valores de gastos médicos varían considerablemente entre individuos, lo cual puede tener implicaciones importantes para los modelos predictivos. Es

probable que las predicciones no sean tan precisas. Hay que tomar esto en cuenta, a la hora de desarrollar los modelos.

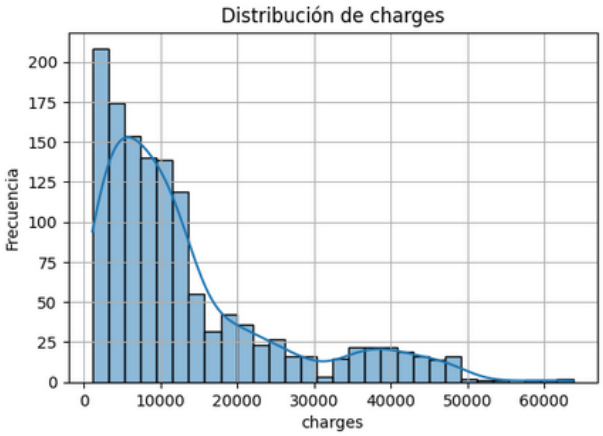

Figura 1: Histograma de la variable objetivo: charges

Este histograma, nos permite ver también que existe cierto sesgo hacia la derecha. Esto significa que la mayoría de las personas no representa un costo tan alto para las aseguradoras, ya que sus gastos se mantienen entre los 1000 y 15,000 dólares más o menos. Mientras que solo una pequeña parte de la población concentra costos más altos, entre los cuarenta y los sesenta mil dólares.

Una posible explicación podría ser que la mayoría de los registros correspondían a jóvenes de 18 años, como lo vemos en el histograma de la variable *age* en la Figura 2, quienes, como se ha demostrado, tienden a generar menos gastos médicos para las aseguradoras.

Los histogramas para las otras variables nos confirman lo que nos decía la moda, que en el estudio predominan las familias sin hijos. Y llama la atención que la variable bmi, a simple vista, presentó una distribución aparentemente normal, a pesar de la presencia de valores atípicos. Esto abre la oportunidad a futuros estudios, que, con más información de los usuarios, podrían ayudar a determinar los factores que influyan en su índice de masa corporal.

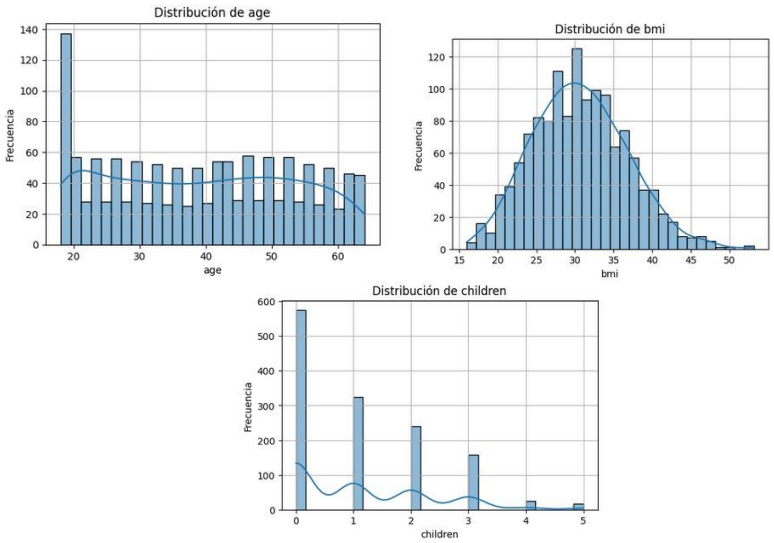

Figura 2: Histogramas de las variables age, bmi y children

d. Relación con la Variable Objetivo

A pesar de que estudios pasados lograron estimar ciertas relaciones entre las variables y la variable objetivo, se decidió explorarlas desde cero en este estudio. En la Figura 3, podemos ver que las gráficas de relación entre charges y las diferentes variables numéricas nos dan una idea de cómo influyen en los gastos médicos.

Además, contamos con un mapa de calor en la Figura 4, que muestra las correlaciones y los coeficientes de variación, que facilitan la comprensión de la fuerza de las relaciones y la dispersión de cada variable.

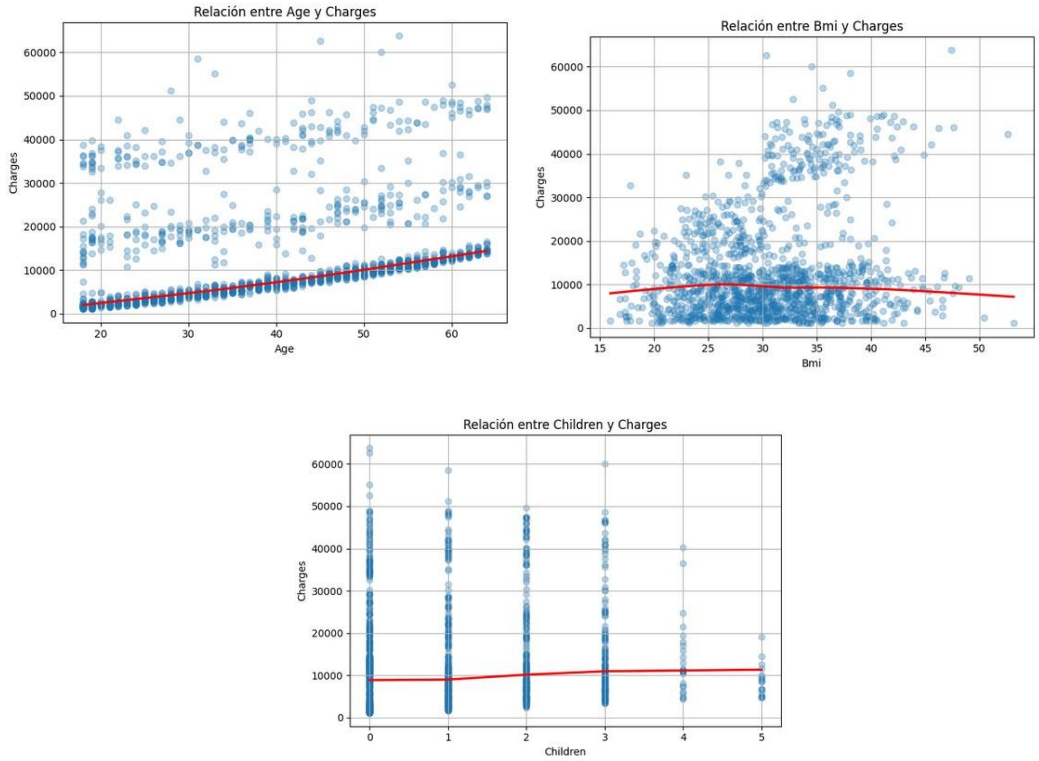

Figura 3: Relación entre la Variable Objetivo y las Variables Numéricas

La relación entre la edad y los gastos médicos muestra una tendencia creciente. Con el paso del tiempo, los gastos parecen aumentar rápidamente. Esto implica que la edad es un elemento clave para anticipar los costos de salud. Esta conclusión es reforzada por el mapa de calor, donde la variable “edad” muestra una correlación positiva con los “costos”, y su coeficiente de variación (0.358) señala una variabilidad moderada y controlada.

Además, el coeficiente de determinación R^2 para la regresión lineal entre *age* y *charges* fue de 0.0894, lo cual indica que, aunque la relación es clara, la edad por sí sola explica menos del 10% de la variabilidad en los gastos médicos.

En el caso del índice de masa corporal (IMC), no se observa una relación lineal fuerte con los gastos. Aunque se ve una concentración de gastos altos en aquellos con IMC elevado, la variabilidad es considerable en todos los niveles. Esto coincide con su baja correlación con los gastos en el mapa de calor, y su coeficiente de variación de 0.199 muestra que es una variable bastante homogénea, lo cual puede limitar su utilidad como predictor individual.

Esto también se refleja en su R^2 de apenas 0.0393, lo que indica que el IMC explica solo un 3.9% de la variación en los costos.

La variable que representa el número de hijos tiene una conexión muy débil con los gastos médicos. La curva permanece casi nivelada y no hay diferencias significativas entre aquellos que tienen más o menos hijos. Este hallazgo se refleja en su correlación cercana a cero en el mapa de calor, mientras que su coeficiente de variación (1.101) indica una alta dispersión. En otras palabras, aunque presenta mucha variabilidad respecto a su media, no ayuda mucho a explicar los costos.

Esto se confirma con su R^2 de apenas 0.0046, el más bajo entre las variables numéricas analizadas.

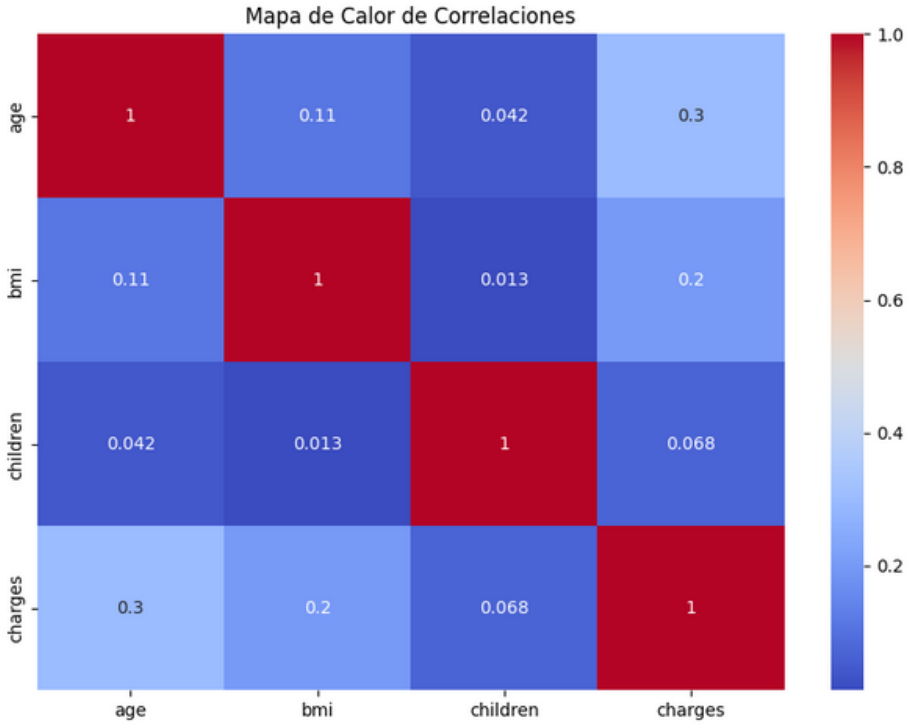

Figura 4: Mapa de Calor de Correlaciones de las Variables Numéricas

7.2.2. Variables Categóricas

El análisis exploratorio de las variables categóricas se realizó por medio del estudio de la frecuencia de cada una. Los resultados nos indican que, en términos de género, la mayor cantidad de los registros son masculinos. En cuanto a la región,

la mayoría son del sureste (*southeast*). Y gran parte de los que participaron en el estudio, no fuman, o al menos esa fue la respuesta que dieron.

a. Relación con la Variable Objetivo

Por medio de boxplots que vemos en la Figura 5, se exploró la relación entre charges y las variables categóricas. En el caso de la variable **smoker**, se puede ver que las personas que fuman gastan mucho más en salud que aquellos que no fuman.

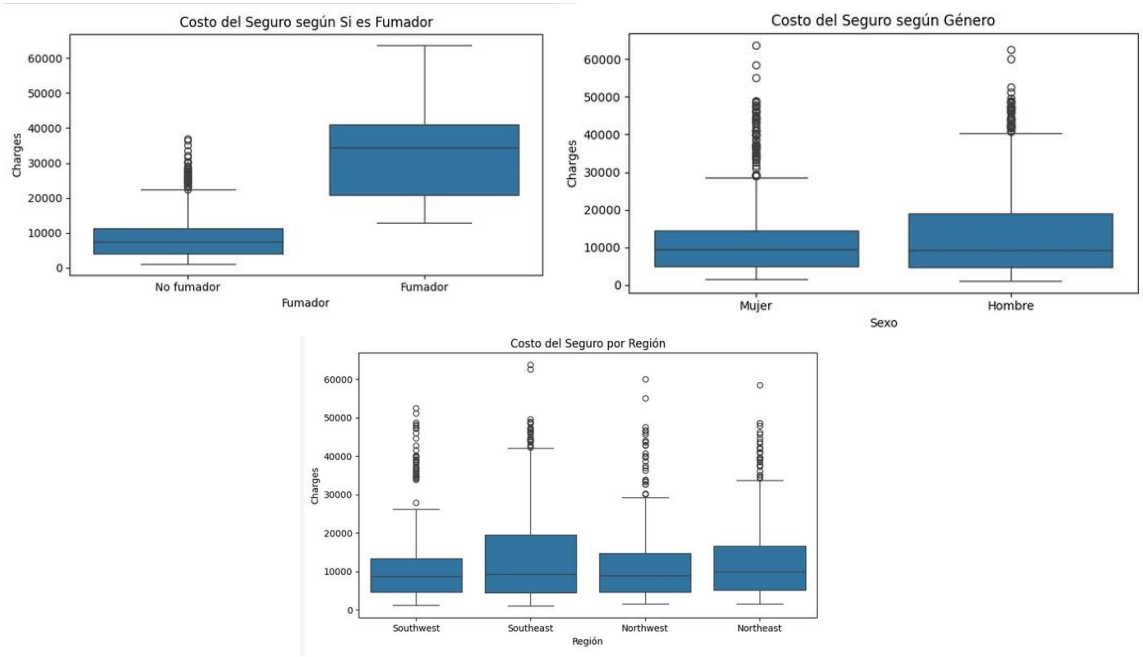

Figura 5: Boxplots de la distribución de la variable objetivo para las diferentes variables categóricas.

Esto lo confirma el hecho de que solo tenemos *outliers* para los no fumadores, en comparación con los fumadores. Ya que esto indica que en para los no fumadores, los gastos altos son poco comunes, mientras que, en los fumadores, aunque los gastos son altos, se consideran normales dentro del grupo. Esto muestra que **fumar tiene un fuerte impacto en el costo del seguro**.

Este impacto también se ve reflejado en el coeficiente de determinación R^2 = 0.6197, el cual es el más alto entre todas las variables analizadas, lo que sugiere que fumar explica más del 60% de la variación en los costos médicos. También, su correlación con los gastos es de 0.7872, lo cual refuerza la fuerte relación positiva entre ambos.

En contraste, para la variable sex, las distribuciones de hombres y mujeres son bastante similares, tanto en términos de dispersión como en la mediana, aunque se presentan valores extremos en ambos casos. Esto indica que **el género no parece tener un efecto importante en los costos**, dado que los patrones de gasto son parecidos en los dos grupos.

Esto podemos confirmarlo con su R^2 de solo 0.0034 y una correlación de apenas 0.0580 con los gastos, lo que indica una relación prácticamente nula.

Con relación a la región, se pueden notar diferencias marcadas entre las distintas áreas. El sureste muestra una mediana algo más elevada, junto con una mayor dispersión entre la mediana y el tercer cuartil, lo que señala una variación significativa en los valores más altos.

Por otro lado, las regiones del norte, como el noreste y el noroeste, muestran una mayor agrupación de valores fuera de lo normal, lo que implica la existencia de casos extremos. Esto indica que, **aunque la zona podría tener cierta influencia en los costos médicos, su impacto no es constante ni definitivo por sí mismo**.

Esto se evidencia en los bajos valores de R^2: 0.0015 para el noroeste, 0.0054 para el sureste y 0.0019 para el suroeste. Las correlaciones también son muy débiles: -0.0387, 0.0736 y -0.0436 respectivamente.

Es importante mencionar que, para poder trabajar estas variables, primero fue necesario transformarlas a valores booleanos.

7.3. PREPROCESAMIENTO Y LIMPIEZA

Luego de la exploración inicial, se pudo concluir que la data presentaba cierto número de valores atípicos. Se evaluó la posibilidad de eliminar estos *outliers* del *dataset*. En total se identificaron 39 casos de valores atípicos en la variable *charges* y 9 en la variable *bmi*, como podemos ver en la Figura 6. Pero, como estos representaban apenas el 3.6% de los datos, se llegó a la conclusión de que no afectaban significativamente el estudio.

Además, si eliminamos estos valores, nuestro modelo aprendería únicamente sobre los casos "normales", y tendría dificultades para predecir con precisión los costos para clientes con características que hagan subir los costos. Provocaría que el modelo subestime de manera constante esos casos reales.

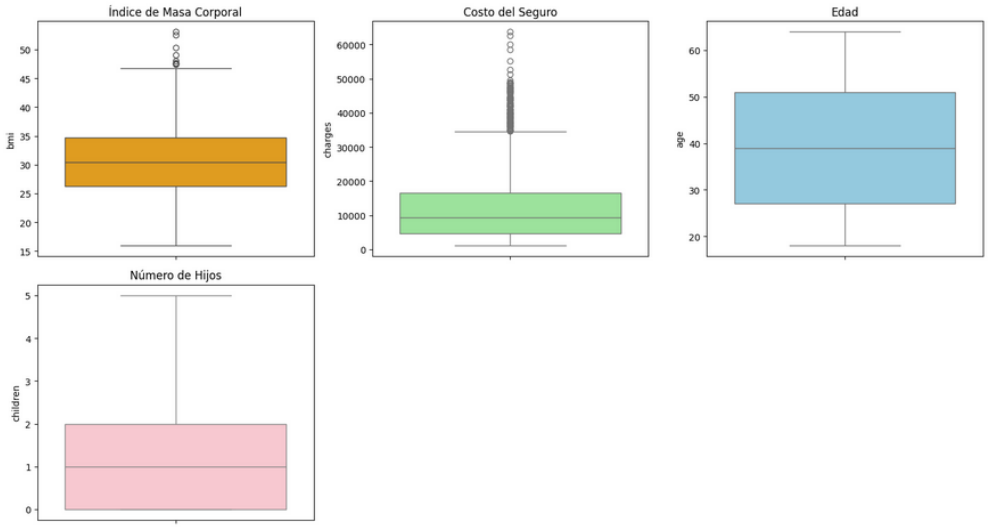

Figura 6: Boxplots de las variables numéricas

Sin embargo, sí se realizó una **limpieza inicial del dataset**, donde se eliminó el único valor duplicado que se identificó. Por ende, el modelo se desarrolló a partir de los 1,337 registros restantes. Además, cabe resaltar que no se eliminó ninguna de las columnas, ya que no se podía descartar la posibilidad de que cada una tuviera un impacto significativo en la variable objetivo, a pesar de lo que concluían estudios previos. Y, al no contar con valores nulos, no fue necesario realizar una imputación de los datos.

7.3.1. Transformación de variables categóricas a variables dummy

Las variables categóricas se transformaron en binarias, para que los modelos de pudieran comprenderlas bien, ya que necesitan datos en formato numérico. Este procedimiento consistió en crear nuevas columnas que tuvieran valores de 0 o 1, los cuales indican si cada categoría está presente o no. En la Tabla 2, se explica cuáles fueron las columnas que se crearon.

Columna	Descripción
sex_male	El valor 1 indica género masculino, el 0, femenino
smoker_yes	El valor 1 indica que es fumador, el 0, que no
region_northwest	El valor 1 indica que es de la región noroeste, el 0, que es de otra región
region_northeast	El valor 1 indica que es de la región noreste, el 0, que es de otra región
region_southeast	El valor 1 indica que es de la región sureste, el 0, que es de otra región

Tabla 2: Variables Dummies creadas a partir de las variables categóricas

Para evitar multicolinealidad, la región suroeste se usó como categoría base, por eso no se le creó su propia columna. Por lo tanto, si los valores de todas las demás regiones están en cero, podemos concluir que el individuo es de la región suroeste.

7.5. MODELADO

7.5.1. SELECCIÓN DE VARIABLES

Como se explicó en la sección de Preprocesamiento, se tomó la decisión de tomar en consideración todas las variables para el desarrollo de los modelos predictivos, ya que todas tienen cierto impacto sobre el costo del seguro médico. Se trabajó con el dataset ya limpio y codificado, y se exploraron diferentes enfoques para mejorar el rendimiento del modelo.

Actualmente, nuestro "*Dataset codificado*" contiene nueve variables independientes y una variable objetivo. Solo a variable región se descartará, debido a que las columnas que se crearon cumplirán su función dentro del modelo.

7.5.2. SELECCIÓN DE MODELOS

Debido a la naturaleza del dataset, y basándonos en las características obtenidas a partir del análisis descriptivo de los datos, se eligieron cinco modelos para trabajar los datos, con a intención de evaluar su desempeño y así determinar cuál ofrece la mejor precisión en la predicción.

a. Regresión Lineal

Una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal (AWS, 2024).

b. XGBoost

Es un modelo mejorado de árboles de decisión. Es muy útil en situaciones donde las variables tienen relaciones complicadas. Tiene un buen rendimiento incluso con valores atípicos o distribuciones irregulares. Se destaca por su rapidez, efectividad y habilidad para manejar grandes cantidades de datos (Kavlakoglu & Russi, 2024).

c. Ridge

Es una técnica de regularización estadística. Corrige el sobreajuste de los datos de entrenamiento en los modelos de machine learning. Es uno de los varios tipos de regularización para modelos de regresión lineal (Murel & Kavlakoglu, 2023).

d. Random Forest

IBM (2023) lo define como un modelo que combina varios árboles de decisión para lograr un resultado que sea más exacto y fuerte. Gracias a su método de agrupación, es menos sensible a datos extremos y a la información ruidosa, ya que cada árbol examina una sección diferente de los datos y, al promediar sus salidas, se reducen los errores individuales. Además, su habilidad para evaluar variables de forma aleatoria le permite reconocer relaciones complejas entre características que un solo árbol o un modelo lineal no podrían detectar.

Al explorar diferentes combinaciones de variables, el modelo puede resaltar interacciones no lineales o conexiones sutiles que no se notarían en un análisis tradicional. Esto lo convierte en una herramienta valiosa en conjuntos de datos con muchas variables o donde las relaciones no son claras a simple vista.

e. Gradient Boosting

Está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar y corregir los errores de los árboles anteriores. La predicción de una nueva observación se obtiene combinando las predicciones de todos los árboles individuales que forman el modelo (Amat, 2020). En este proyecto, fue el modelo con mejor rendimiento.

7.5.3. DIVISIÓN DEL CONJUNTO DE DATOS

Inicialmente se dividió el dataset en datos de entrenamiento y prueba aplicando CART, pero los resultados se vieron afectados significativamente al momento de evaluar los modelos. No solo se obtuvieron valores anormales, sino que además, los valores del coeficiente de determinación superaban los rangos teóricos, incluso algunos daban negativo. Ejemplo, un R^2 de -10.13 o de -7.33.

Ante esta situación, se decidió optar por una división estándar del conjunto de datos utilizando la técnica de *train-test split* con una proporción de 80/20. Este

cambio nos ayudó a obtener valores más coherentes para el coeficiente de determinación.

7.5.4. TRANSFORMACIONES DE LA VARIABLE OBJETIVO

Al comenzar a analizar los datos, observamos que la variable objetivo tenía un rango de valores muy amplio y variado, que iban desde 1,121.87 hasta 63,770.43, como pudimos observar en la Tabla 1. Esto podía afectar las predicciones de los modelos.

Durante las pruebas iniciales con los modelos, se calcularon las siguientes métricas, para evaluar los modelos: **MAE** (Error Absoluto Medio), **MSE** (Error Cuadrático Medio), **RMSE** (Raíz del MSE) y **MAPE** (Error Porcentual Absoluto Medio).

Aunque se lograron buenos valores de R^2, el análisis de los errores absolutos reveló lo contrario. Los errores eran tan altos que demostraban que el modelo se equivocaba en gran medida al momento de predecir el costo. En la Tabla 3, se presenta un breve resumen de los valores obtenidos, con los diferentes modelos.

Modelo	MAE	MSE	RMSE	R^2	MAPE
Regresión Lineal	4,177.05	35,478,020.68	5,956.34	0.80690	41.40%
XGBOOST	2,922.40	24,786,419.95	4,978.60	0.86510	42.11%
Ridge	4,194.01	35,656,881.00	5,971.34	0.806	41.64%
Random Forest	2,668.82	22,240,867.62	4,716.02	0.879	36.86%
Gradient Boosting	2,484.94	17,993,073.50	4,241.82	0.9021	30.69%

Tabla 3: Evaluación de los Modelos. Variable objetivo charges

Los resultados muestran que los modelos tienen dificultades para predecir sobre todo valores bajos de costo de seguro.

Por ejemplo, un MAE de $4,000 indica que, en promedio, el modelo se equivoca por esa cantidad cuando hace una predicción.

Si se compara con casos donde el valor real del seguro es de $1,200, vemos que el error puede ser incluso mayor que el valor real, lo que representa un problema importante. Esto sugiere que los modelos funcionan mejor cuando el costo es alto, pero no son precisos cuando el valor es bajo, lo que limita su uso en escenarios reales.

En general, todos los modelos evaluados presentaron errores altos. El que mostró el mejor desempeño fue Gradient Boosting, a pesar de que sus resultados también fueron elevados.

Esta situación nos llevó a evaluar diferentes posibilidades, llegando a la siguiente solución: **la transformación de variable la variable objetivo**. De esta manera, esperamos poder: reducir el impacto de los valores atípicos, obtener una distribución más equilibrada y mejorar la precisión de los modelos. Las transformaciones aplicadas fueron las siguientes:

a. **Logaritmo Base 10 ($\log y$)**: Reduce el impacto impacto de números elevados y ajusta distribuciones que están sesgadas hacia la derecha.
b. **Logaritmo Natural ($\ln y$)**: Facilita la interpretación de cambios en forma de porcentajes, lo que ayuda en el estudio del crecimiento exponencial.
c. **Raíz Cuadrada ($\sqrt{y}$):** Reduce la variabilidad y es útil cuando los datos son conteos o tienen menor sesgo. Ayuda a estabilizar la varianza y mejorar la simetría de la distribución
d. **Inverso ($1/y$):** Transforma números grandes en pequeños, lo cual es útil para castigar valores extremos altos. No obstante, puede ser bastante drástico y se debe utilizar con cuidado.

Cada transformación se aplicó en una nueva versión del dataset original, trabajada en Excel.

8. RESULTADOS

8.1. EVALUACIÓN DE LOS MODELOS

8.1.1. TRANFORMACIÓN ln(y)

Si bien todos los modelos se entrenaron y se evaluaron con las mismas versiones de la variable objetivo, y los mismos modelos, se lograron los mejores resultados al usar la transformación: $\ln y$ junto con el modelo de **Gradient Boosting**.

Esta combinación permitió obtener el mejor balance entre el error y el coeficiente de determinación (R^2), lo que sugiere un excelente ajuste entre el modelo y los datos.

Y los errores absolutos y relativos (MAE, MSE, RMSE y MAPE) se mantuvieron en niveles muy bajos, sobre todo en proporción al rango de valores del dataset.

count	*mean*	*std*	*Min*	*25%*	*50%*	*75%*	*max*
1,337	*9.099928*	*0.918699*	*7.022756*	*8.46513*	*9.146992*	*9.720629*	*11.063045*

Tabla 4: Estadísticas descriptivas de la variable ln(y)

MODELO	*MAE*	*MSE*	*RMSE*	*R^2*	*MAPE*
Regresión Lineal	0.26	0.16	0.40	0.8295	2.89%
XGBoost	0.23	0.18	0.42	0.8098	2.55%
Ridge	0.26	0.16	0.40	0.8291	2.90%
Random Forest	0.20	0.15	0.38	0.8418	2.28%
Gradient Boosting	0.19	0.11	0.33	0.8819	2.10%

Tabla 5: Evaluación de los modelos predictivos para la variable ln(y)

Como podemos observar en la Tabla 4, la variable $\ln y$ tuvo un intervalo de aproximadamente 7.02 a 11.06.

En este sentido, los errores del modelo de Gradient Boosting, como un MAE de 0.19 o un RMSE de 0.33, son bastante bajos y manejables, mostrando pequeñas variaciones dentro del rango natural de la variable.

Igualmente, el MAPE siempre permaneció por debajo del 3%, alcanzando su valor más bajo con el modelo de Gradient Boosting, lo que sugiere que, en promedio, las predicciones del modelo se desviaron menos del 3% del valor real. Esto constituye un resultado muy positivo, lo cual respalda su selección como la mejor configuración de modelo y transformación.

Además, dado que esta transformación eliminó los valores atípicos, los errores se mantuvieron estables sin verse influenciados por extremos, lo que reforzó la solidez del modelo, como se refleja en el diagrama de caja de la Figura 7.

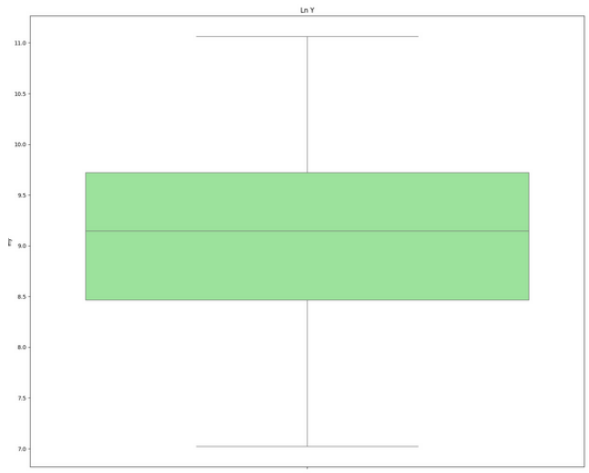

Figura 7: Boxplot para la variable de ln(y)

8.1.2. TRANFORMACIÓN log(y)

Por otro lado, los modelos que fueron entrenados con la transformación $\log y$ también mostraron grandes mejorías en comparación con el uso de la variable objetivo original. Esta transformación facilitó la estabilización de la distribución, disminuyendo la varianza y aumentando la precisión de las predicciones. Y eliminando también los valores atípicos.

Así que podemos decir que las **transformaciones logarítmicas** ayudan a aumentar notablemente el desempeño de los modelos al ajustar mejor las dimensiones de los datos a los algoritmos de aprendizaje. Al reducir valores extremos y minimizar la asimetría, estas modificaciones crean distribuciones más equilibradas que permiten hacer pronósticos más exactos y coherentes a través de todo el espectro de valores. Al menos para este conjunto de datos.

count	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
1337	3.952048	0.398986	3.049944	3.676359	3.972488	4.221615	4.804619

Tabla 6: Estadísticas descriptivas de la variable log(y)

Modelo	MAE	MSE	RMSE	R^2	MAPE
Regresión Lineal	0.11	0.03	0.17	0.8295	2.89%
XGBoost	0.10	0.03	0.19	0.8038	2.50%
Ridge	0.11	0.03	0.17	0.8291	2.90%
Random Forest	0.09	0.03	0.16	0.8517	2.20%
Gradient Boosting	0.08	0.02	0.15	0.8783	2.15%

Tabla 7: Evaluación de los modelos predictivos para la variable log(y)

8.1.3. TRANFORMACIÓN 1/y

En la Tabla 9, podemos ver que los modelos que usaron la transformación 1/Y lograron cifras muy cercanas a cero, en todas las métricas de error.

count	mean	std	min	25%	50%	75%	max
1337	0.000169	0.000167	0.000016	0.00006	0.000107	0.000211	0.000891

Tabla 8: Estadísticas descriptivas de la variable 1/y

Modelo	MAE	MSE	RMSE	R^2	MAPE
Regresión Lineal	0	0	0	0.66090	94.14%
XGBoost	0	0	0	0.72680	77.55%
Ridge	0	0	0	0.6608	93.88%
Random Forest	0	0	0	0.8695	23.47%
Gradient Boosting	0	0	0	0.8814	30.69%

Tabla 9: Estadísticas descriptivas de la variable 1/y

Aunque esto a simple vista pueda parecer bueno, la verdad es que se debe principalmente a que esta transformación reduce los valores originales a un rango muy pequeño, como vemos en las estadísticas de la Tabla 8. La verdadera medida que tenemos para medir qué tanto se equivocaban los modelos es el MAPE, el cual como podemos ver fue muy alto para todos los modelos. Entonces, podemos concluir que esta no es una buena transformación para estos modelos, ni para este conjunto de datos.

8.1.4. TRANFORMACIÓN RAIZ DE Y

Si observamos el rango de datos obtenido luego de calcular la raíz de la variable objetivo (**raíz de y**), vemos que los errores de todos los modelos son muy altos, sobre todo para las predicciones de costos más bajos, pues la diferencia sería muy alta. Esto es algo similar a lo que ocurrió con la variable original, solo que en una menor magnitud. Por eso, podemos decir que no es una buena trasnformación para trabajar estos modelos,ni este conjunto de datos.

count	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
1,337	104.88173	47.756152	33.494386	68.893715	96.882203	129.06478	252.528074

Tabla 10: Estadísticas descriptivas de la variable 1/y

Modelo	*MAE*	*MSE*	*RMSE*	*R^2*	*MAPE*
Regresión Lineal	14.5	415.88	20.39	0.8436	14.49%
XGBoost	12.05	449.95	21.21	0.8308	13.70%
Ridge	14.58	417.67	20.44	0.8429	14.58%
Random Forest	10.79	375.94	19.39	0.8586	12.29%
Gradient Boosting	9.98	281.47	16.78	0.8942	10.93%

Tabla 11: Estadísticas descriptivas de la variable 1/y

Esto confirma que la mejor transformación para trabajar este conjunto de datos es la transformación logarítmica, particularmente de **logaritmo natural**, pues arrojó los mejores resultados. Y, el mejor modelo predictivo para este caso sería el **Gradient Boosting**.

Es por eso por lo que ahora enfocaremos nuestro análisis en la variable objetivo: **logaritmo natural del valor de los costos de salud**. Y realizaremos las predicciones, utilizando el modelo de **Gradient Boosting**, evaluando su desempeño.

No obstante, aunque los errores sobre $\ln y$ sean mínimos, incluso al compararlas con el rango de valores de la variable objetivo, es importante tener en cuenta que al regresar las predicciones a la forma original del objetivo (dinero), pueden surgir cambios. Por eso, debe analizarse con cuidado ese proceso, para asegurar que el modelo permanezca efectivo en la práctica.

8.2. GRÁFICAS

8.2.1. GRÁFICA: Valores Reales vs Predichos - GRADIENT BOOSTING – Variable ln(y)

En la Figura 8 tenemos una comparación entre los valores reales y los valores predichos por el modelo de *Gradient Boosting*, para la variable de ln(y).

Podemos observar que la mayoría de los puntos azules están bastante próximos a la línea roja, que simboliza el punto ideal en el que el valor previsto coincide exactamente con el valor real.

Esto sugiere que **el modelo es muy efectivo en sus predicciones**, dado que muchas de ellas están muy cerca de lo que realmente sucedió.

Sin embargo, en los extremos, vemos que tanto para los valores bajos como para los altos, existen algunos puntos que se encuentran un poco más distantes de la línea. Esto indica que el modelo comete errores un poco más significativos cuando se enfrenta a valores extremadamente pequeños o grandes, aunque en general sigue teniendo un buen desempeño.

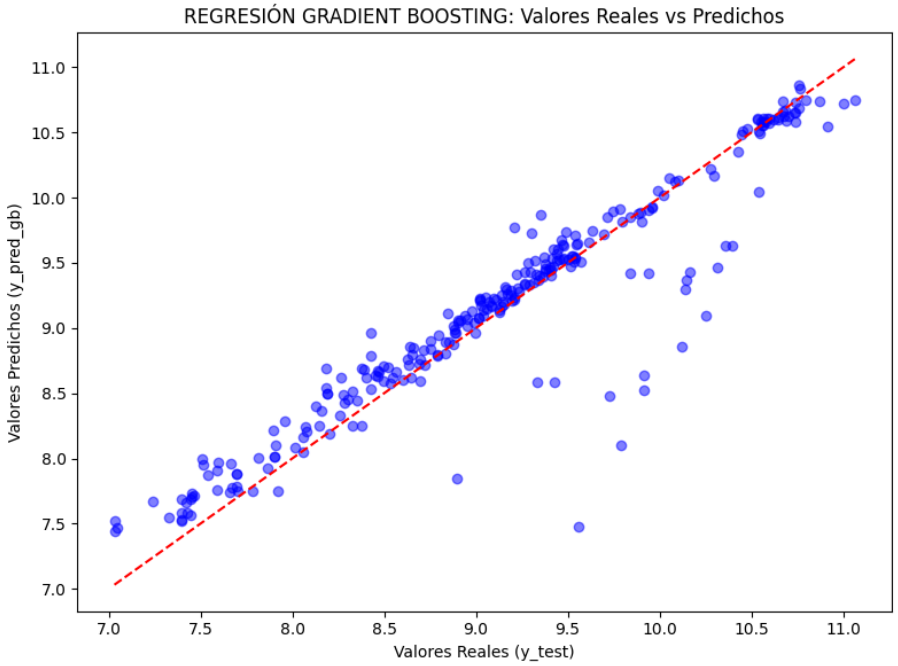

Figura 8: GRÁFICA: Valores Reales vs Predichos - GRADIENT BOOSTING – Variable ln(y)

8.2.2. GRÁFICA: Gráfico de Residuos - GRADIENT BOOSTING – Variable ln(y)

En la Figura 9 se observa la distribución de los residuos del modelo, es decir, la diferencia entre los valores reales y los valores que predijo el modelo. En primer lugar, esta distribución está centrada en cero, lo cual es una buena señal. Esto significa que el modelo casi siempre acierta en sus predicciones o se equivoca por poco.

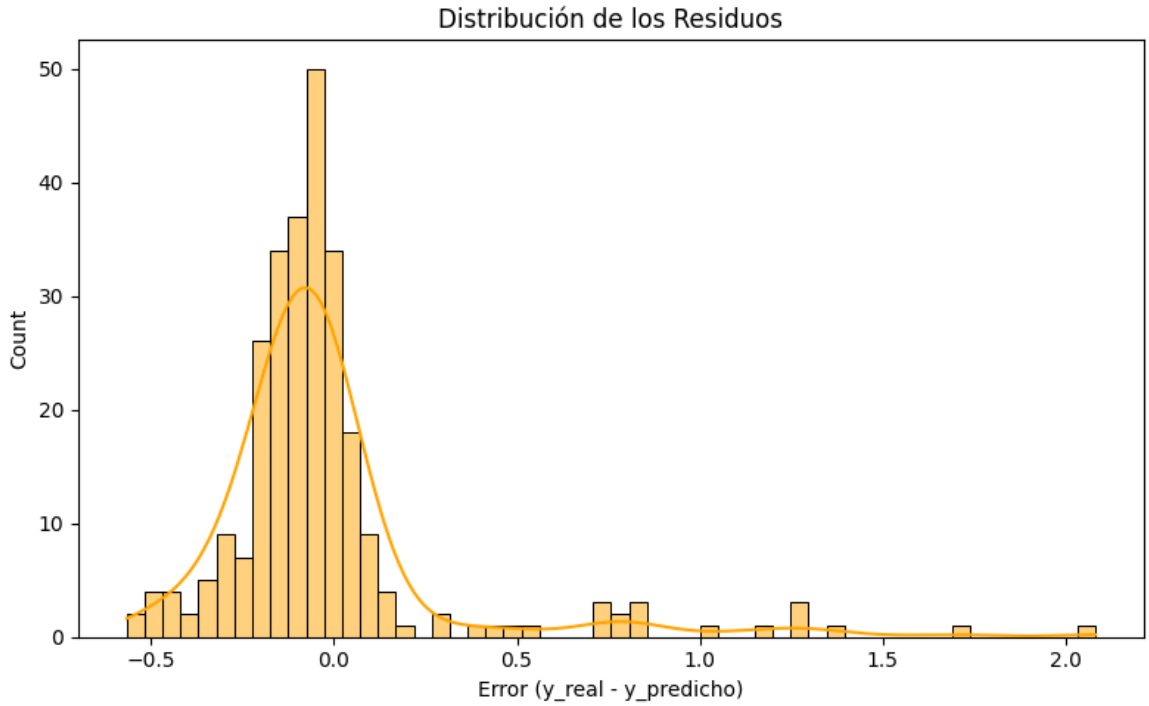

Figura 9: GRÁFICA: Distribución de Residuos - GRADIENT BOOSTING – Variable ln(y)

Además, la forma del gráfico nos sugiere que los residuos siguen una distribución normal y que solo unos pocos son significativos.

Aunque vemos algunos errores un poco más grandes hacia la derecha (errores positivos), estos son escasos. Esto implica que el modelo no presenta errores que puedan signficar un problema para el estudio.

En términos generales, podemos decir que el modelo de Gradient Boosting funciona de manera efectiva, realiza sus predicciones con precisión, y no presenta indicios de que esté fallando o prediciendo de manera inusual o repetitiva. Solo presenta algunos errores, pero son mínimos.

9. CONCLUSIONES

El objetivo principal de crear un modelo eficaz y preciso para estimar el costo de los seguros médicos, en función de antecedentes demográficos y de salud, se alcanzó con éxito. Para lograr esto, fue necesario analizar y comparar diversos modelos, eligiendo el que mejor se ajustara a las características de los datos.

Dado que la variable que se busca predecir (charges) tenía una distribución asimétrica con sesgo hacia la derecha, alejándose de lo que sería una distribución normal, se realizó una transformación logarítmica (logaritmo natural) para mejorar el rendimiento del modelo. Este ajuste también dejó claro que había una relación no lineal entre las variables de entrada y el costo del seguro.

De todos los modelos analizados, el Gradient Boosting resultó ser el más eficaz, sin importar la transformación que se aplicó a la variable objetivo. Su habilidad para entender relaciones complejas a través de un proceso de prueba y error en cada árbol de decisión fue fundamental para conseguir predicciones precisas.

10. RECOMENDACIONES Y FUTUROS ESTUDIOS

1. Ingeniería de características y selección de modelos

- Aplicar ingeniería de características para explorar con mayor detalle la relación entre las variables y el objetivo.
- Aunque investigaciones anteriores ya han implementado este enfoque, se recomienda complementarlo con modelos de Ridge o Lasso, dependiendo del número de variables. Esto permitirá evaluar con mayor precisión el impacto directo de cada variable en la variable objetivo.

2. Validación Cruzada

- Si bien en este caso la validación cruzada no fue efectiva para dividir los datos durante el entrenamiento y prueba de los modelos, en futuros estudios, bajo otras condiciones, podría ser una alternativa viable.

3. **Incorporación de Nuevas Variables**
    - Se sugiere enriquecer el modelo con características adicionales, como: Factores de riesgo no considerados en este estudio (ej. alcoholismo, enfermedades hereditarias o congénitas).
    - Profundizar en variables ya analizadas (ej. tabaquismo), pero con mayor granularidad.

4. **Desarrollo de un Modelo en Contexto Panameño**
    - La recomendación principal es recopilar datos locales para construir un modelo adaptado a la realidad panameña, especialmente ante la situación actual del sistema de salud.
    - Realizar un estudio regional. Debido a la variedad geográfica y económica de Panamá, sería útil examinar la información por regiones (como provincias, comarcas o zonas urbanas/rurales). Esto ayudaría a reconocer patrones concretos (como el acceso a servicios de salud y la presencia de factores de riesgo) y a crear estrategias específicas.

11. BIBLIOGRAFÍA

Aflac. (23 de abril de 2025). *Aflac*. Obtenido de 9 de cada 10 estadounidenses han postergado chequeos médicos y exámenes que podrían salvarles la vida: https://newsroom.aflac.com/2025-04-25-9-de-cada-10-estadounidenses-han-postergado-chequeos-medicos-y-examenes-que-podrian-salvarles-la-vida

Amat, J. (Octubre de 2020). *Ciencia de Datos*. Obtenido de Gradient Boosting con Python: https://cienciadedatos.net/documentos/py09_gradient_boosting_python

AWS. (2024). *Amazon Web Services*. Obtenido de ¿Qué es la regresión lineal?: https://aws.amazon.com/es/what-is/linear-regression/

Ballweg, G. (2022). *Action on Smoking Health*. Obtenido de Hidden Costs: The Economic Burden of Cigarette Smoking on U.S. Healthcare Spending: https://ash.org/hidden-costs-healthcare/

Bupa Global. (2025). *Cómo funciona el sistema de salud en Estados Unidos* . Obtenido de BupaSalud: https://www.bupasalud.com.pa/salud/sistema-salud-estados-unidos

Calderón, M. (13 de Febrero de 2025). *Créditos en USA*. Obtenido de ¿Cuánto cuesta un seguro médico en Estados Unidos?: https://www.creditosenusa.com/cuanto-cuesta-un-seguro-medico-en-estados-unidos/

Carlton, E., Scott, J., Moniz, M., & Prescott, H. (27 de Marzo de 2023). *Institute for Healthcare Policy and Innovation*. Obtenido de Even with private insurance, your child's hospitalization could cost $1,300: https://ihpi.umich.edu/news-events/news/even-private-insurance-your-childs-hospitalization-could-cost-1300

Centers for Medicare & Medicaid Services . (24 de Junio de 2025). *Centers for Medicare & Medicaid Services* . Obtenido de NHE Fact Sheet: https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet#:~:text=,with%20average%20spending

CNN. (2 de Octubre de 2023). *CNN: Economía y Dinero*. Obtenido de ¿Por qué las mujeres pagan más que los hombres con su seguro de salud?:

https://cnnespanol.cnn.com/video/eeuu-desigualdad-genero-poliza-pagos-seguro-salud-cnn-dinero-tv/

Djebali, H. (2025). *GitHub*. Obtenido de Kaggle Medical Cost Personal: https://github.com/spideystreet/kaggle-medical-cost-personal

Fei, S. (2023). *Analysis of the Medical Cost Personal Data*. Obtenido de GitHub: https://christopherdavisuci.github.io/UCI-Math-10-S23/Proj/StudentProjects/ShuchengFei.html

Finhabits. (26 de Noviembre de 2024). *Finhabits*. Obtenido de ¿Cuánto cuesta un seguro médico en Estados Unidos?: https://www.finhabits.com/es/cuanto-cuesta-un-seguro-medico-en-estados-unidos/

Henry J. Kaiser Family Foundation. (2012). *Health Care Costs: A Primer.* San Francisco: Henry J. Kaiser Family Foundation.

Herrera, M. (20 de Febrero de 2025). *Creditos en USA*. Obtenido de Seguros de salud privados en USA: Precios y planes: https://www.creditosenusa.com/seguros-medicos-privados-precios-y-planes-en-usa/

Herrera, M. (20 de Febrero de 2025). *Créditos en USA*. Obtenido de Seguros de salud privados en USA: Precios y planes: https://www.creditosenusa.com/seguros-medicos-privados-precios-y-planes-en-usa/

IBM. (2023). *IBM*. Obtenido de ¿Qué es el bosque aleatorio?: https://www.ibm.com/es-es/think/topics/random-forest

Kavlakoglu, E., & Russi, E. (9 de Mayo de 2024). *IBM*. Obtenido de ¿Qué es XGBoost?: https://www.ibm.com/mx-es/think/topics/xgboost

Lennon, C. (18 de Diciembre de 2018). Employer-Sponsored Health Insurance and the Gender Wage Gap: Evidence from the Employer Mandate. *Revista Económica del Sur*. Obtenido de El seguro médico patrocinado por el empleador y la brecha salarial de género.

McCann, A. (15 de Enero de 2020). *The University of Chicago*. Obtenido de The Real Cost of Smoking by State feat. Dr. Fridberg: https://psychiatry.uchicago.edu/news/real-cost-smoking-state

MSH International. (30 de Diciembre de 2022). *MSH International*. Obtenido de Medical expenses abroad: which countries are the most expensive in the world?: https://www.msh-intl.com/en/medical-expenses-abroad-countries-most-expensive.html

Murel, J., & Kavlakoglu, E. (21 de Noviembre de 2023). *IBM*. Obtenido de ¿Qué es regression de ridge? : https://www.ibm.com/mx-es/think/topics/ridge-regression

NIHCM Foundation. (18 de Junio de 2025). *NIHCM Foundation*. Obtenido de What Is Driving Price Variation in Private Health Insurance?: https://nihcm.org/publications/what-is-driving-price-variation-in-private-health-insurance

Palacio, K. (15 de Marzo de 2025). *Revista Cosas*. Obtenido de ¿Por qué las mujeres pagan hasta 42% más en un seguro médico que los hombres?: https://revistacosas.mx/seguro-medico-mujeres-pagan-mas-costos-brecha-de-genero/

Plenilunia Salud Mujer. (28 de Marzo de 2025). *Plenilunia Salud Mujer*. Obtenido de ¿Por qué las mujeres pagan más por su seguro de salud? Factores detrás de la brecha de género en costos médicos: https://plenilunia.com/estilo-de-vida/cuida-tu-dinero/seguro-de-salud-para-mujeres/102656/

Quantum Pro. (21 de Mayo de 2023). *Quantum Pro*. Obtenido de Tu seguro y tu BMI: https://www.quantumpropr.com/l/tu-seguro-y-tu-bmi/#:~:text=El%20%C3%ADndice%20de%20masa%20corporal%20(IMC)%20es,el%20peso%20de%20una%20persona%20y%20proporciona

Seguros del Pichincha. (s.f.). *Seguros del Pichincha*. Obtenido de ¿Cómo fumar afecta a la prima de su seguro de vida?: https://segurosdelpichincha.com/blogs/impacto-del-tabaquismo-en-la-salud-y-tu-seguro?srsltid=AfmBOoqZedMEtnw9k4kqvnUkjud1ZNwGQTjWYFwOB_xCAzDjizRCNOBa

Stobbe, M. (29 de Abril de 2023). *Los Angeles Times*. Obtenido de La tasa de tabaquismo en EEUU cae al mínimo histórico: https://www.latimes.com/espanol/vida-y-estilo/articulo/2023-04-29/la-tasa-de-tabaquismo-en-eeuu-cae-al-minimo-historico

SWI. (13 de marzo de 2018). *SWI*. Obtenido de La atención sanitaria cuesta el doble en EEUU que en otros países ricos: https://www.swissinfo.ch/spa/la-atenci%C3%B3n-sanitaria-cuesta-el-doble-en-eeuu-que-en-otros-pa%C3%ADses-ricos/43970380

Telesford, I., & Schwartz, H. (24 de Marzo de 2024). *Petterson-KFF Health System Tracker*. Obtenido de How have costs associated with obesity changed over time?: https://www.healthsystemtracker.org/chart-collection/how-have-costs-associated-with-obesity-changed-over-time/

The Texas Heart Institute. (2025). *Calculadora del índice de masa corporal (IMC)*. Obtenido de The Texas Heart Institute: https://www.texasheart.org/heart-health/heart-information-center/topics/calculadora-del-indice-de-masa-corporal-imc/

Vapor Technology Asociation. (18 de Abril de 2025). *Vapor Technology Asociation*. Obtenido de REPORT: Healthcare Costs and GDP Impact of Cigarette Smoking vs. Vaping in the United States (2015–2025): https://vaportechnology.org/healthcare-costs-and-gdp-impact-of-cigarette-smoking/

WorldRemit Ltd. (3 de Febrero de 2023). *WorldRemit*. Obtenido de ¿Cómo funciona el sistema de salud en Estados Unidos?: https://www.worldremit.com/es/blog/life-abroad/como-funciona-el-sistema-de-salud-de-estados-unidos#%C2%BFC%C3%B3mo%20es%20el%20modelo%20de%20sistema%20de%20salu%20de%20Estados%20Unidos?

12. ANEXOS

Elemento	Link (GitHub)
1. **Bases de datos**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/tree/main/Bases%20de%20Datos
2. **Descripción de cada columna de datos**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/blob/main/Descripci%C3%B3n%20de%20cada%20columna%20de%20datos.xlsx
3. **Análisis descriptivo**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/tree/main/ANALISIS%20DESCRIPTIVO
4. **Análisis predictivo**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/tree/main/ANALISIS%20PREDICTIVO
5. **Scripts**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/tree/main/scripts
6. **Códigos**	https://github.com/juanfratm/proyecto_final_juantorres_modelospredictivos/tree/main/codigos