Pronóstico de Costos Médicos con Regresión Lineal

Jorge Armando Jurado Peralta
Heidy Alexandra Moreno
Óscar Andrés Guzmán

28 de junio de 2025

1. Introducción

En la actualidad, la predicción de costes médicos se ha convertido en una tarea fundamental para las aseguradoras de salud, ya que permite estimar de manera anticipada los gastos asociados a sus clientes y, con ello, optimizar la asignación de recursos y definir primas más ajustadas. Un enfoque común y eficaz para abordar este tipo de problema es el uso de modelos estadísticos y de aprendizaje automático, como la **regresión lineal**.

Este estudio se basa en un conjunto de datos que contiene información personal de individuos, incluyendo variables como la edad, el sexo, el índice de masa corporal (IMC), el número de hijos, el hábito de fumar y la región de residencia. La variable objetivo es el coste médico anual que representa el gasto que cada individuo genera para una aseguradora.

El objetivo principal de este trabajo es construir un modelo de regresión lineal capaz de predecir con precisión el coste médico individual, utilizando las variables mencionadas como predictores. A través del análisis exploratorio de datos, la evaluación de la correlación entre variables y la aplicación de técnicas estadísticas, se busca comprender la relación existente entre las características personales y los costes médicos asociados, proporcionando una herramienta útil para la toma de decisiones en el ámbito de los seguros de salud.

2. Fundamentos teóricos

2.1. ¿Qué es la regresión lineal?

La regresión lineal es una técnica de modelado estadístico que permite predecir el valor de una variable dependiente (o de respuesta) basada en el valor de una o más variables

independientes (o predictoras), asumiendo que la relación entre ellas es lineal. Es una forma común de analizar y entender la relación entre dos o más variables.

2.2. Aplicaciones

La regresión lineal se utiliza en una amplia variedad de campos, incluyendo:

- Ciencia: Para estudiar la relación entre variables en experimentos y observaciones.
- Economía: Para predecir el comportamiento de variables económicas, como la demanda, el crecimiento económico, etc.
- Ingeniería: Para modelar y analizar sistemas y procesos.
- Salud: Para analizar la relación entre factores de riesgo y enfermedades.
- Finanzas: Para modelar el comportamiento de inversiones y predecir riesgos.

2.3. Supuestos del modelo lineal

1. Linealidad: La relación entre la variable dependiente (Y) y las variables independientes (X) debe ser lineal. Esto significa que la relación puede representarse con una recta.
2. Independencia: Los errores (o residuos) deben ser independientes entre sí, es decir, no deben estar correlacionados. Esto implica que el error en una observación no debe depender del error en otra observación.
3. Homocedasticidad: Los errores deben tener una varianza constante para todas las observaciones. Esto significa que la dispersión de los errores debe ser la misma en todos los niveles de las variables independientes.
4. Normalidad: Los errores deben tener una distribución normal. Esto implica que la distribución de los errores debe ser simétrica y tener forma de campana.

2.4. Métricas de evaluación

Se utilizan métricas como el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2) para evaluar la calidad del modelo.

Cuadro 1: Métricas comunes para evaluación de modelos de regresión

Métrica	**Descripción**
MAE (Mean Absolute Error)	Promedio de los valores absolutos de los errores. Mide la magnitud promedio de los errores sin considerar su dirección.
MSE (Mean Squared Error)	Promedio de los errores al cuadrado. Penaliza más fuertemente los errores grandes. Útil para modelos donde los errores grandes son especialmente indeseables.
R^2 (Coeficiente de Determinación)	Mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo. Su valor está entre 0 y 1.

3. Descripción del conjunto de datos

3.1. Fuente de datos

- **Fuente:** Kaggle – Medical Cost Personal Dataset
- **Descripción del objetivo:** Predicción de costos médicos en dólares
- **Características:**
  - **Edad:** Edad del paciente
  - **Sexo:** Género del paciente (masculino/femenino)
  - **IMC:** Índice de masa corporal
  - **Hijos:** Número de dependientes a cargo
  - **Fumador:** Condición de fumador (sí/no)
  - **Región:** Región geográfica de residencia
  - **Cargos:** Costo médico total (variable objetivo)

4. Desarrollo del modelo

4.1. Exploración y visualización

4.2. Trazando los datos

Se comienza cargando el conjunto de datos e importar diferentes librerias:

- NumPy

- pandas
- Matplotlib
- Seaborn
- scikit-learn
- XGBoost

1. Importación de Librerias

Comenzamos importando todas las bibliotecas necesarias de Python para la manipulación de datos, visualización y aprendizaje automático.

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Carga del Dataset

Cargamos el conjunto de datos de seguros desde el directorio /data mediante la biblioteca pandas, para su posterior análisis y procesamiento.

In [ ]:
```
url = 'https://github.com/Geo1486/Costos_Medicos_ML/blob/main/Data/Medical_Cost.csv'
df = pd.read_csv(url, sep=",", header=None)
df.head()
```

```
Dataset loaded successfully.
```

Out[ ]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figura 1: Importación

4. Exploración preliminar de los datos

Exploramos las correlaciones entre las variables numéricas.

In [5]:

```
# Plot distribution of charges
plt.figure(figsize=(8, 5))
sns.histplot(df['charges'], kde=True, bins=30)
plt.title("Distribution of Medical Charges")
plt.show()

# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include=[np.number])

# Plot correlation matrix only for numeric data
plt.figure(figsize=(8, 6))
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Matrix (Numeric Features Only)")
plt.show()
```

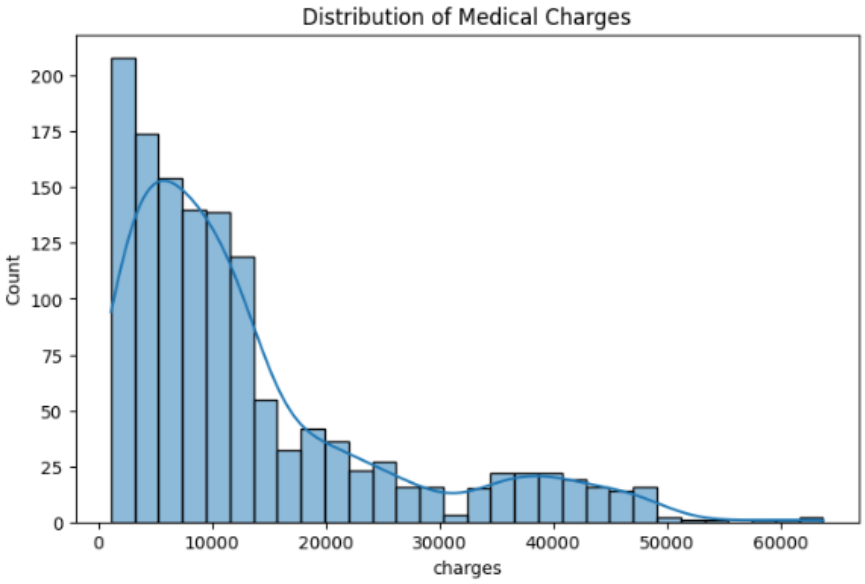

Figura 2: Exploración

8. Visualización de predicciones vs. valores reales

Representamos gráficamente la comparación entre los cargos médicos predichos y los reales del conjunto de prueba, con el fin de evaluar visualmente la precisión del modelo.

In [10]:
```
# Visualize predictions vs actual values
plt.figure(figsize=(8, 6))
sns.scatterplot(x=y_test, y=y_pred)
plt.xlabel("Actual Charges")
plt.ylabel("Predicted Charges")
plt.title("Actual vs Predicted Medical Charges")
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')  # perfect prediction line
plt.show()
```

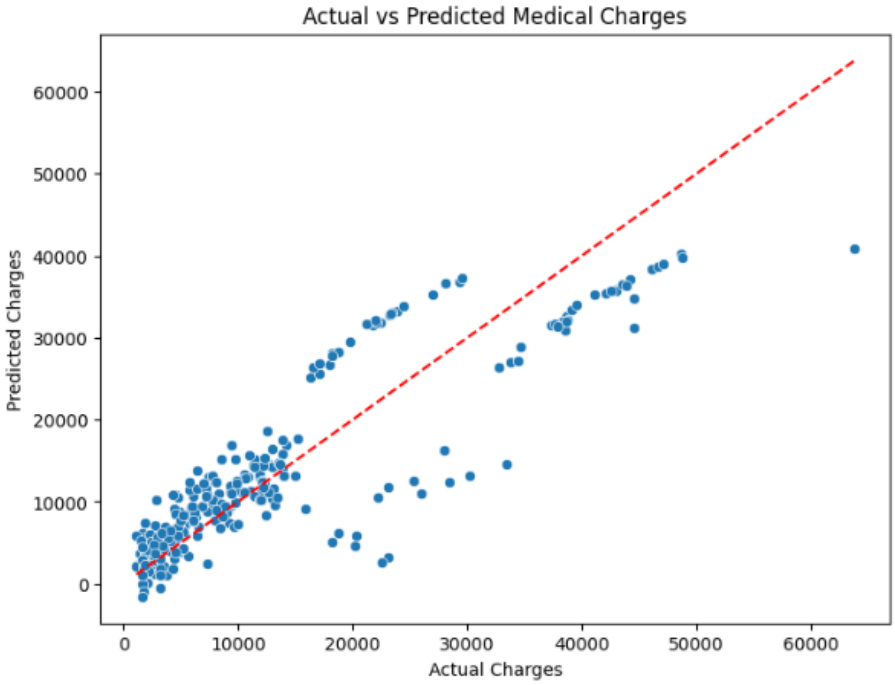

Figura 3: Predicciones vs Valores reales

Hallazgos Clave

1. **Correlaciones**

   - *Análisis visual:* El mapa de calor muestra correlaciones entre variables numéricas.

   Relaciones esperadas:

   - **Edad y costos:** Positiva (a mayor edad, mayores costos).
   - **BMI y costos:** Positiva (mayor BMI puede asociarse a mayores costos).
   - **Fumadores:** Probablemente tengan costos significativamente mayores.

2. **Distribuciones**

- **Edad:** Rango de 18 a 64 años (según datos de muestra).
- **BMI:** Valores típicos entre 27.9 y 33.77 en los primeros registros.
- **Costos:** Amplio rango desde $1,725 hasta $16,884 en los primeros registros.

3. **Calidad de Datos**
   - No hay valores nulos.
   - Tipos de datos adecuados para cada columna.
   - Datos limpios y listos para análisis más profundo.

5. Regresión logística

5.1. Anuncios en redes sociales

Objetivo del análisis

Predecir la probabilidad de que un usuario realice una compra tras visualizar un anuncio en una red social, utilizando como base sus características demográficas, como edad, género y salario estimado. Se trata de un problema de clasificación binaria, donde el resultado puede ser **compra** o **no compra.**

Análisis exploratorio sugerido

- **Distribución de variables:**
  - Histograma de edad y salario.
  - Conteo por género.
  - Proporción de compras (`Purchased`).
- **Correlaciones:**
  - ¿Existe relación entre edad y probabilidad de compra?
  - ¿El salario influye en la decisión?
- **Visualización:**
  - Gráficos de dispersión (Age vs Salary) coloreados por `Purchased`.
  - Boxplots para comparar distribución de salario entre quienes compran y quienes no.

Modelado

Puedes aplicar modelos como:

- Regresión logística
- K-Nearest Neighbors (KNN)
- Árboles de decisión o Random Forest
- SVM (Support Vector Machine)

Con técnicas como:

- División en conjunto de entrenamiento y prueba
- Escalado de variables (por ejemplo, con `StandardScaler`)
- Validación cruzada

Métricas de evaluación

- Accuracy
- Matriz de confusión
- Precision, Recall, F1-score
- Curva ROC y AUC

1. Objetivo del análisis

Predecir la probabilidad de que un usuario realice una compra tras visualizar un anuncio en una red social, utilizando como base sus características demográficas, como edad, género y salario estimado. Se trata de un problema de **clasificación binaria**, donde el resultado puede ser "compra" (1) o "no compra" (0).

2. Análisis exploratorio

Se realizó una exploración inicial de los datos con los siguientes hallazgos:

- **Distribución del objetivo (`Purchased`):** El conjunto de datos muestra una distribución relativamente equilibrada entre usuarios que compran y los que no.
- **Edad:** Los usuarios que compran suelen tener una edad promedio mayor que los que no compran.

- **Salario estimado:** Existe una relación positiva entre el salario estimado y la probabilidad de compra. Los usuarios con mayores ingresos presentan mayor tasa de conversión.
- **Género:** La variable género se distribuye de forma equitativa, aunque se observan diferencias en patrones de compra entre hombres y mujeres, especialmente al analizar en conjunto con edad y salario.
- **Visualizaciones clave:**
    - Diagramas de caja (boxplots) para edad y salario según decisión de compra.
    - Gráficos de dispersión: edad vs. salario.
    - Gráficos de barras segmentados por género y clase.

Gráficas

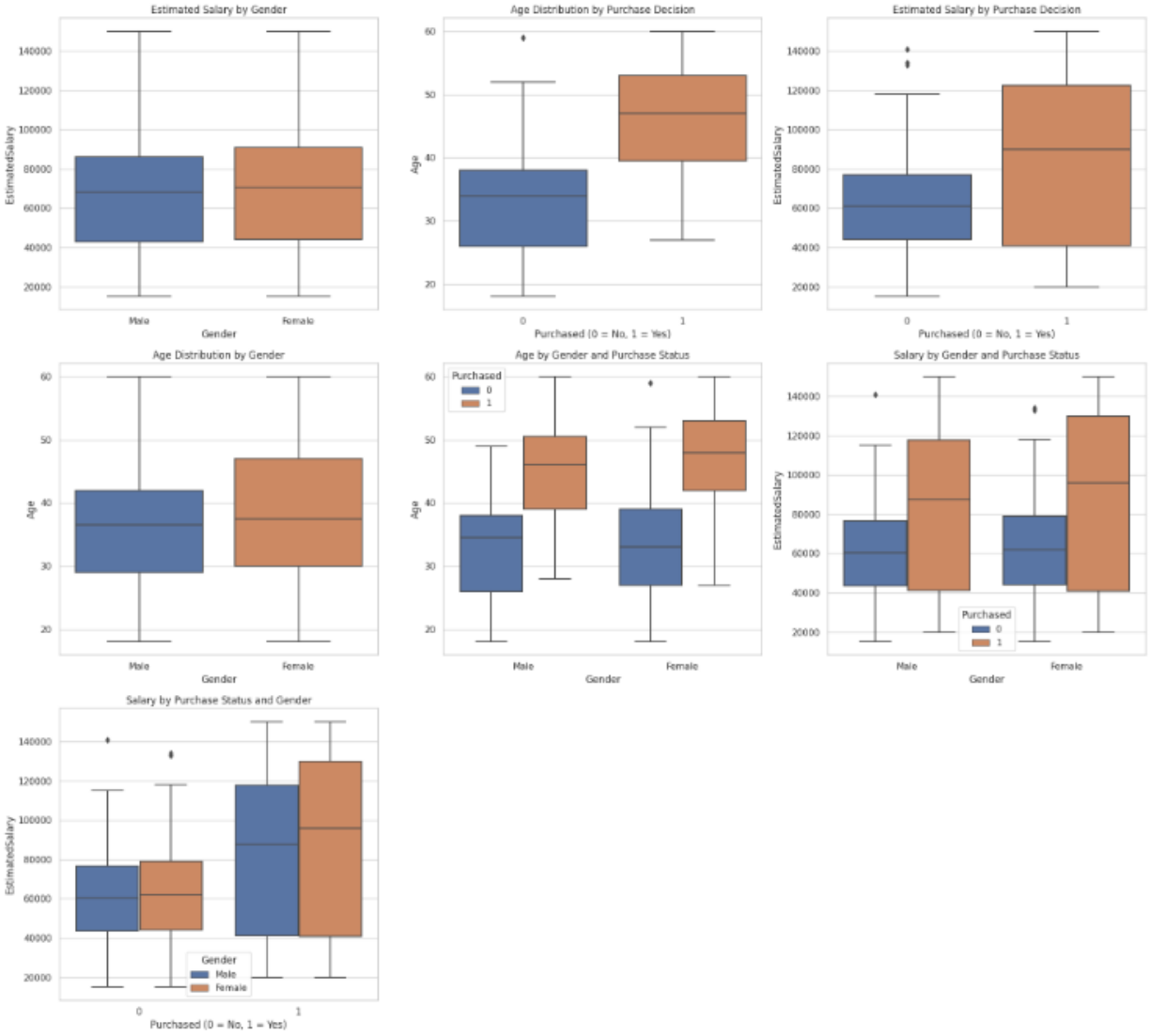

Figura 4: Diagramas de caja (boxplots) para edad y salario según decisión de compra

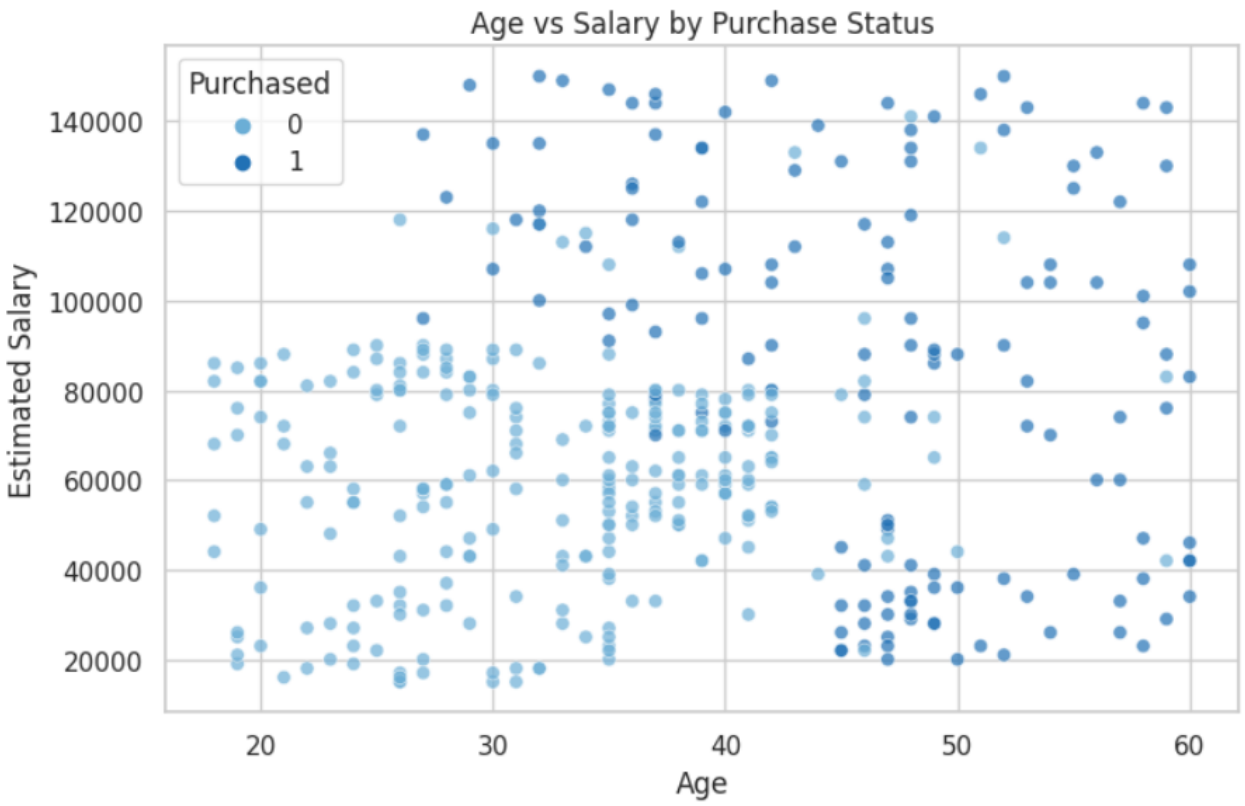

Figura 5: Gráficos de dispersión: edad vs. salario.

Regresión lineal y logística en Machine Learning

A lo largo de este ejercicio hemos explorado dos de los modelos más fundamentales en el aprendizaje automático supervisado: la **regresión lineal** y la **regresión logística**. Aunque ambos parten del mismo principio básico modelar la relación entre variables independientes y una variable objetivo, su aplicación y naturaleza nos muestran caminos distintos pero complementarios.

La **regresión lineal** nos permite entender cómo varía una respuesta continua en función de ciertos factores. Es una herramienta poderosa cuando el objetivo es predecir valores numéricos y explorar relaciones proporcionales entre variables. Es simple, interpretable y, a menudo, una excelente línea base.

Por su parte, la **regresión logística** introduce un giro crucial al permitirnos abordar problemas de *clasificación binaria*, como predecir si un usuario hará clic en un anuncio, si un paciente tiene una enfermedad o si un correo es spam. Nos obliga a pensar en términos de *probabilidades y decisiones*, y nos prepara para escenarios reales en los que las decisiones no son sobre cuánto, sino sobre sí o no.

Ambos modelos nos enseñan que antes de pensar en algoritmos complejos, debemos dominar los fundamentos. Aprender a interpretar coeficientes, evaluar el ajuste del modelo y entender la lógica que subyace en una predicción es más valioso que simplemente obtener una métrica alta.

Para finalizar, estos modelos no solo son el punto de partida del Machine Learning moderno, sino también herramientas esenciales para desarrollar pensamiento crítico, modelar con responsabilidad y tomar decisiones basadas en datos.