Facultad de Ingeniería Mecánica y Eléctrica

Universidad Autónoma de Nuevo León

Programación para Inteligencia Artificial

Grupo: 001 | Equipo: 09

Implementación de un Modelo Supervisado en Python

Actividad Fundamental 6

Fecha: 21/11/2025

Profesor: Dr. Erick de Jesús Ordaz Rivas

Integrantes

Nombre	Matrícula	Hora clase
Orlando Alvarado Vargas	2226968	V1
Diego Alonso Carrillo Castillo	2144556	V1

Índice

1. Introducción

El objetivo de este proyecto es desarrollar un modelo de Machine Learning capaz de predecir el costo médico individual basado en características demográficas y de salud de los pacientes. Para ello se emplea un enfoque supervisado utilizando el algoritmo de Regresión Lineal, debido a que la variable objetivo (costo del seguro médico) es continua.

El dataset empleado contiene información de múltiples personas aseguradas, incluyendo variables como edad, índice de masa corporal, número de hijos, género, hábito de fumar y región geográfica. Mediante un proceso de preprocesamiento, normalización, codificación categórica e implementación del modelo, se evalúa el rendimiento predictivo mediante métricas estándar para regresión.

2. Selección del dataset

2.1. Nombre del dataset y fuente (URL)

Insurance Cost Dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance

2.2. Número de registros y variables.

1338 registros y 7 variables

2.3. Descripción de las variables principales y tipo de problema (regresión o clasificación)

El dataset tiene 7 columnas: Edad, Sexo, Índice de Masa Corporal (IMC), número de hijos, si la persona fuma, región y costos médicos. Lo que se quiere predecir son los costos médicos mediante las otras variables. Al ser una predicción numérica, se necesita de una regresión lineal para este modelo.

3. Procesamiento y normalización

3.1. Identificación de valores nulos o atípicos

No se detectaron valores nulos en el dataset

3.2. Transformación de variables categóricas (si aplica).

Se tuvieron que transformar las columnas de sexo, si la persona fuma y región. Esto se realizó con el método de One-hot-encoding mediante la función pd.get_dummies() de pandas

3.3. Normalizacion o Estandarización

Se aplicó MinMaxScaler para escalar todas las columnas al rango [0, 1]:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df_normalizado = pd.DataFrame(scaler.fit_transform(ndatos), columns=ndatos.
    columns)
```

Listing 1: Aplicación MinMaxScaler

3.4. Fragmento del dataframe antes y después del proceso

Antes del preprocesamiento:

age	sex	bmi	children	smoker	region	charges
19	female	27.90	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.55

Después del preprocesamiento:

age	bmi	children	charges	female	male	no	yes	northeast	northwest	...
19	27.9	0	16884.92	1	0	0	1	0	0	...
18	33.7	1	1725.55	0	1	1	0	0	0	...

Después de la normalización:

age	bmi	children	charges	female	male	no	yes	northeast	...
0.00	0.42	0.00	0.45	1.0	0.0	0.0	1.0	0.0	...
0.05	0.52	0.02	0.04	0.0	1.0	1.0	0.0	0.0	...

4. Implementación del modelo

4.1. Descripción del algoritmo seleccionado

Se utilizó Regresión Lineal, la cual modela la relación entre variables mediante una ecuación lineal. Es una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido.

4.2. Breve justificación del porqué se eligió ese modelo.

Se eligió la regresión lineal porque la tarea consiste en predecir un valor numérico continuo, como lo es el costo anual del seguro médico. Este modelo es adecuado cuando se busca comprender cómo influyen distintas variables sobre un resultado cuantitativo. Además, la regresión lineal permite interpretar fácilmente la importancia y el impacto de cada característica del dataset, ya que cuenta con variables que se comportan relativamente lineal respecto a la varible objetivo, lo cual es útil para entender qué factores elevan o reducen el costo médico.

4.3. Código implementado (fragmentos más relevantes).

```
X = df_scaled.drop("charges", axis=1)
```

La instrucción donde se elimina la columna charges para obtener X es importante porque define las variables que el modelo va a utilizar para aprender. Esto evita que la columna que queremos predecir quede dentro de las variables de entrada, lo cual provocaría que el modelo aprenda información que no debería

```
y = df_scaled["charges"]
```

Cuando se selecciona la columna charges como y, se establece formalmente cuál es la variable que se desea predecir. Esto es fundamental porque el modelo necesita un objetivo claro para poder ajustarse durante el entrenamiento

```
X_train, X_test, y_train, y_test = train_test_split(...)
```

Esta parte del código es crucial porque divide el dataset en dos partes: una para entrenar al modelo y otra para evaluar su desempeño. Esto permite saber qué tan bien generaliza el modelo a datos nuevos.

```
model = LinearRegression()
```

es importante porque define el tipo de modelo que se va a utilizar. En este caso, la regresión lineal es ideal porque el objetivo del proyecto es predecir un valor continuo.

```
model.fit(X_train, y_train)
```

Es la más importante del proceso, porque es donde el modelo realmente aprende. Durante este paso, el algoritmo encuentra la relación matemática entre las variables independientes y el valor que se quiere predecir.

4.4. Parámetros principales del modelo (`LinearRegression()` y breve explicación.

fit_intercept=True: Indica que el modelo debe calcular la ordenada al origen (la "b" de y = mx + b).

positive=False: Indica si los coeficientes deben ser únicamente positivos.

copy_X=True: Copia la matriz de entrada para evitar modificar los datos originales.

n_jobs=None: Número de procesadores usados en el cálculo.

5. Evaluación del modelo

5.1. División del dataset

- 70 % entrenamiento
- 15 % validación
- 15 % prueba

5.2. Métricas adecuadas al tipo de modelo

```
mse = mean_squared_error(Y_test, Y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(Y_test, Y_pred)
```

Listing 2: Regresión Lineal

- MSE: Error cuadrático medio
- RMSE: Raíz del MSE
- R^2 Score: Qué tanto el modelo explica los datos

5.3. Tabla o gráfica de resultados obtenidos

La primera imagen es un mapa de calor (heatmap) que muestra la correlación entre todas las variables del dataset.

Los valores van desde -1 (correlación negativa fuerte) hasta 1 (correlación positiva fuerte).

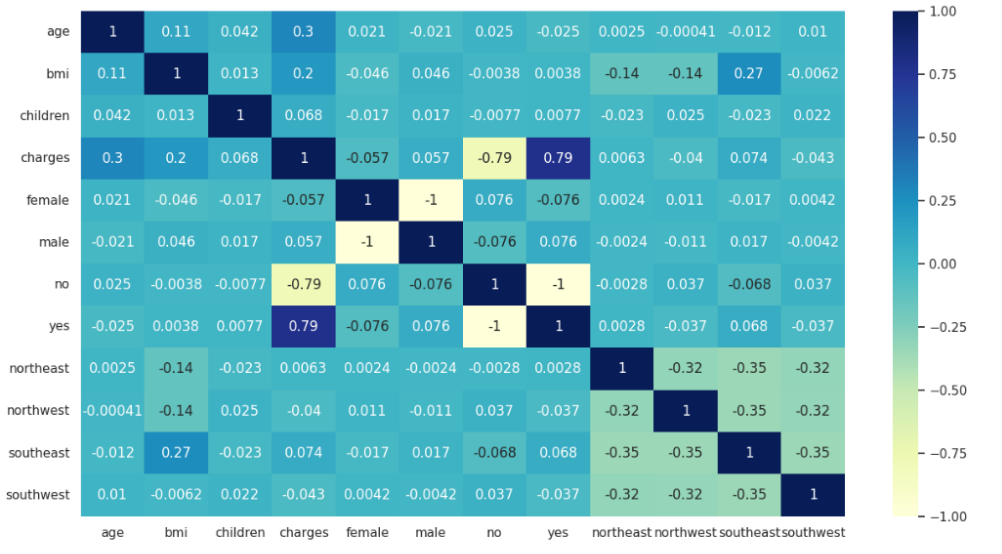


Figura 1: Ejemplo 1 – Heat Map de correlación.

La segunda imagen es una gráfica que compara los valores reales del costo médico (eje X) contra los valores predichos por el modelo (eje Y).

La línea diagonal punteada representa una predicción perfecta.

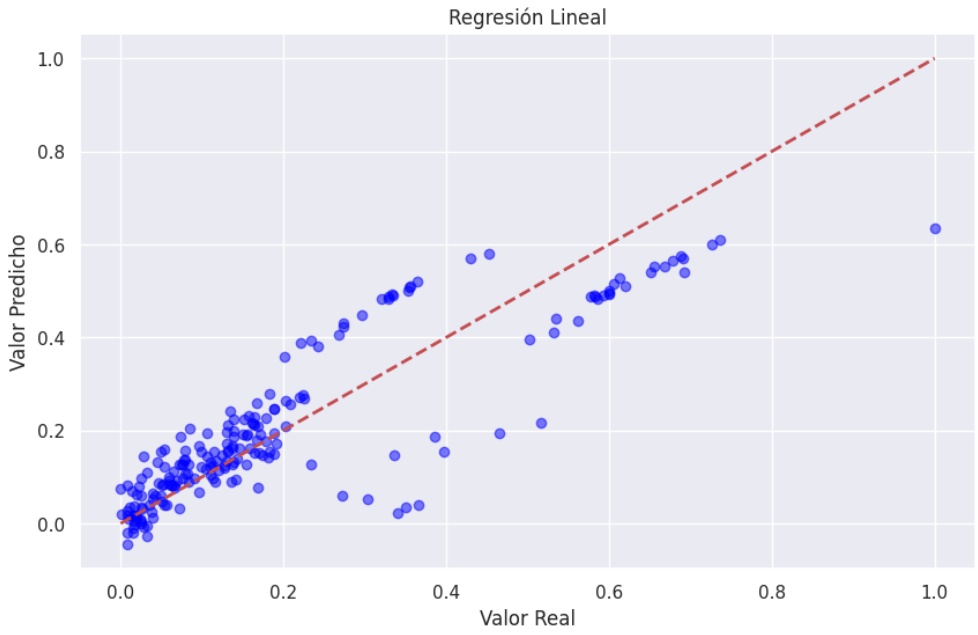


Figura 2: Ejemplo 1 – Gráfica De Regresión Lineal.

6. Interpretación y conclusiones

Después de entrenar el modelo, se puede ver que algunas variables influyen mucho más en el costo médico que otras. Principalmente destacan la edad, el BMI y, sobre todo, si la persona fuma. Esto tiene sentido porque estos factores suelen aumentar el riesgo de problemas de salud, y eso hace que los costos suban.

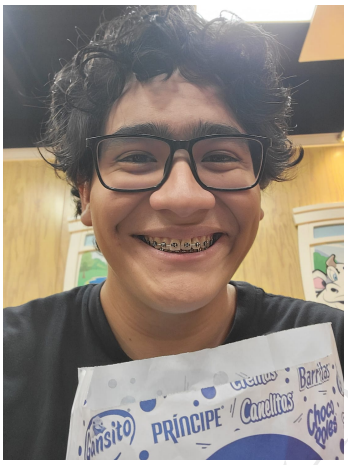
En cuanto al desempeño, el modelo obtuvo un R^2 aceptable, lo que significa que sí logra explicar una buena parte de los costos, aunque claramente no es perfecto. Esto se debe a que los gastos médicos pueden variar muchísimo entre personas y la regresión lineal no alcanza a capturar relaciones más complejas entre las variables.

Aun así, el modelo funciona bien como una primera aproximación. Para mejorar, se podrían probar modelos más avanzados como Random Forest, agregar nuevas variables o incluso probar combinaciones entre las ya existentes. Con esto probablemente aumentaría la precisión.

7. Enlace Al Repositorio Git Hub

https://github.com/NotReysi/PIA-AF6

Fotogafía de los participantes



Orlando Alvarado Vargas

Diego Alonso Carrillo Castillo

Referencias

[1] Choi, M. (2019). Medical cost personal dataset. Recuperado de: https://www.kaggle.com/datasets/mirichoi0218/insurance. Consulta: 21/11/2025.

[2] Developers, S.-L. (2024). Linear regression documentation. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Consulta: 21/11/2025.

[3] Services, A. W. (2024). ¿qué es la regresión lineal? Recuperado de: https://aws.amazon.com/es/what-is/linear-regression/. Consulta: 21/11/2025.

[4] Sánchez, F. (2022). Regresión lineal para machine learning. Recuperado de: https://apiumhub.com/es/tech-blog-barcelona/regresion-lineal-para-machine-learning/. Consulta: 21/11/2025.