THE BRIDGE DIGITAL TALENT ACCELERATOR

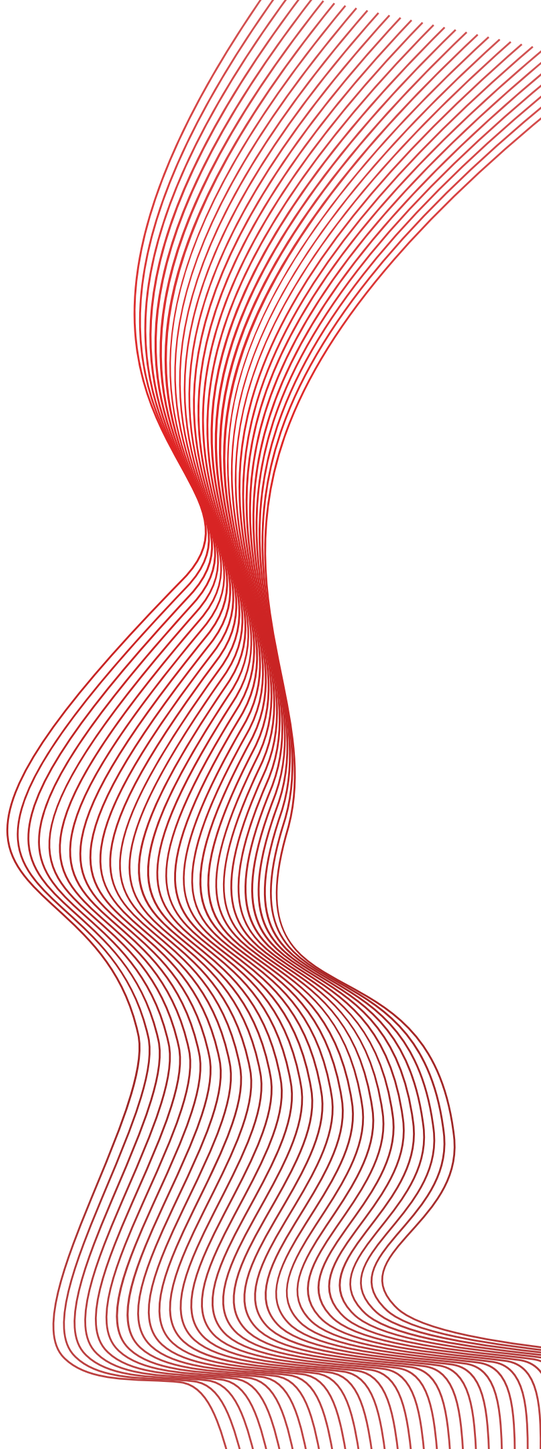
MACHINE LEARNING PROJECT

MODELO DE PREDICCIÓN DE PRECIOS SEGUROS MÉDICOS

PROYECTO FINAL

DATA SCIENCE ONLINE 2025

Presentado por: Deivid Jimenez Sanchez



INTRODUCCIÓN

Este proyecto tiene como objetivo desarrollar un modelo de machine learning capaz de predecir el costo del seguro médico de una persona en función de variables como la edad, el índice de masa corporal (IMC), el número de hijos, el género, la región de residencia y si es fumador o no. Para ello, se ha utilizado un conjunto de datos públicos y se aplicaron diversas técnicas de regresión, incluyendo Regresión Lineal, Random Forest y Gradient Boosting.

Exploración inicial

Se exploran los datos para conocer su estructura, tipos de variables, valores faltantes y estadísticas descriptivas. Se utiliza el método head() de pandas para mostrar las primeras 5 filas del conjunto de datos, de esta manera:

- Obtenemos una visión preliminar de la estructura del dataset.
- Verificamos que los datos se hayan cargado correctamente.
- Identificamos tipos de variables (numéricas, categóricas).
- Detectamos posibles valores faltantes o inconsistencias desde el principio.

8]:
```
insurance_data.head()
```

8]:

	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

Limpieza de datos

Se revisa si hay valores nulos o registros duplicados para asegurar la calidad del dataset.

Se define una función llamada describe_df() que analiza las principales características estructurales de un DataFrame. Este paso es clave para la limpieza exploratoria, ya que permite identificar rápidamente:

- Si hay valores nulos.
- La cardinalidad de las variables (es decir, cuántos valores únicos tiene cada una).
- El tipo de dato (número, string, etc.).

COL_N	age	sex	bmi	children	smoker	region	charges
DATA_TYPE	int64	object	float64	int64	object	object	float64
MISSINGS (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
UNIQUE_VALUES	47	2	548	6	2	4	1337
CARDIN (%)	3.51	0.15	40.96	0.45	0.15	0.3	99.93

Reciclar funciones ya creadas en los Team Challenge

Para simplificar código y mejorar la estructura, vamos a usar algunas funciones creadas en clase y diseñadas para este tipo de problemas...

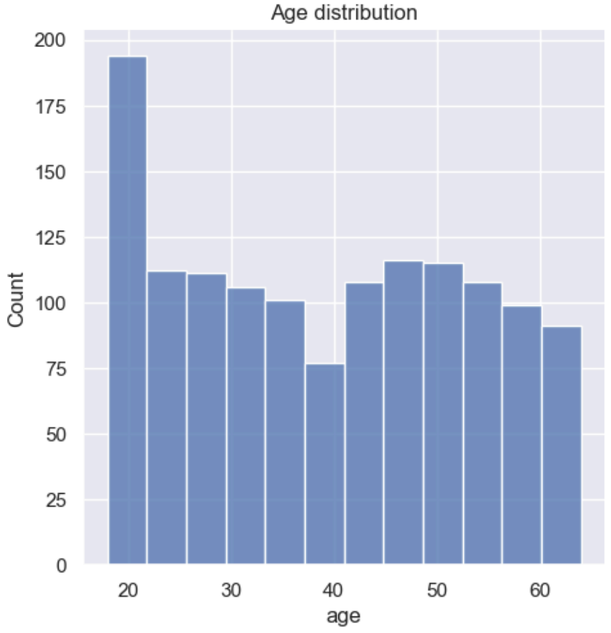
Se define la función tipifica_variables() para clasificar automáticamente las variables del DataFrame según su naturaleza: numérica continua, numérica discreta, categórica o binaria.

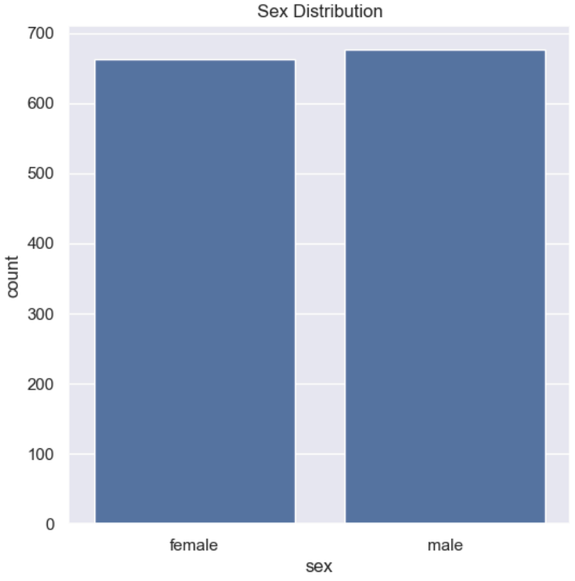
Este proceso es fundamental para preparar el preprocesamiento, ya que permite aplicar correctamente transformaciones como escalado o codificación categórica.

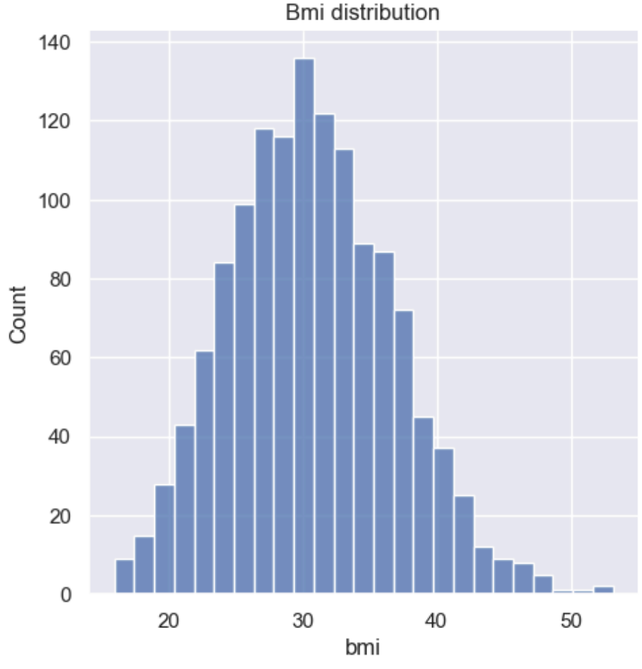
Se basa en dos criterios clave:

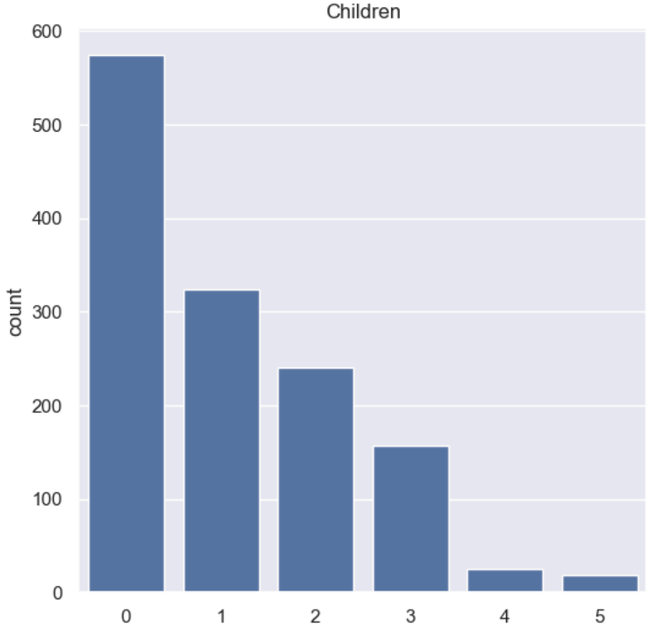
- Cantidad de valores únicos (UNIQUE_VALUES)
- Porcentaje de cardinalidad (CARDIN (%)):

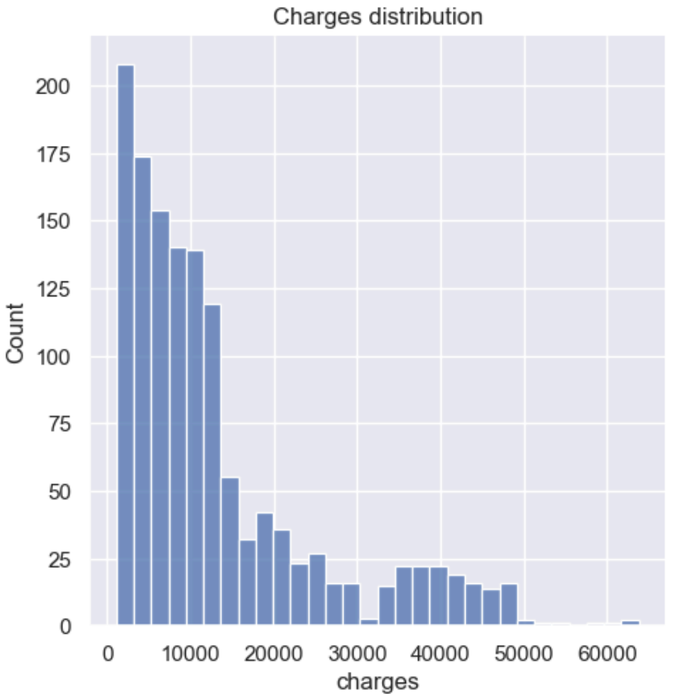
Esta funcion Automatiza la clasificación de variables, lo cual facilita la creación de pipelines de procesamiento y Evita errores comunes al tratar variables numéricas como categóricas, o viceversa.

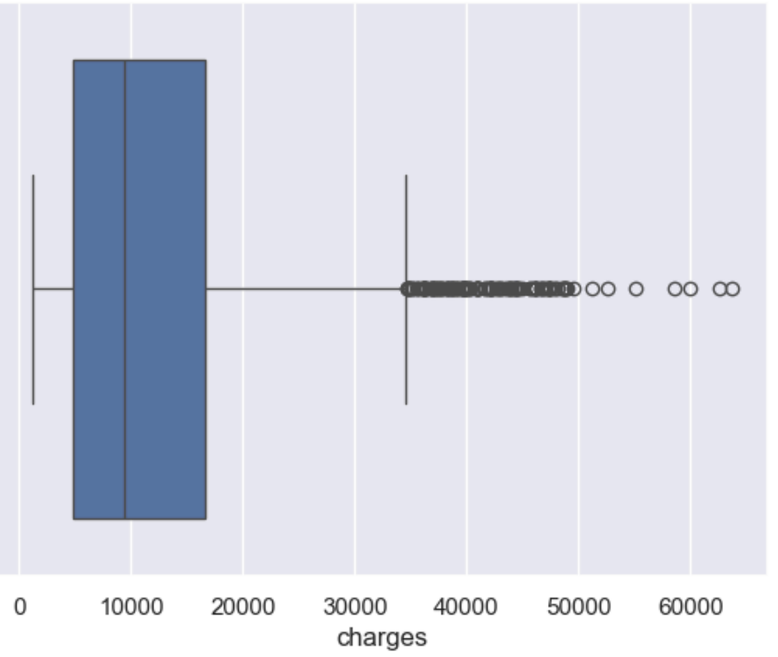

Age distribution
Count
age
0
25
50
75
100
125
150
175
200
20
30
40
50
60

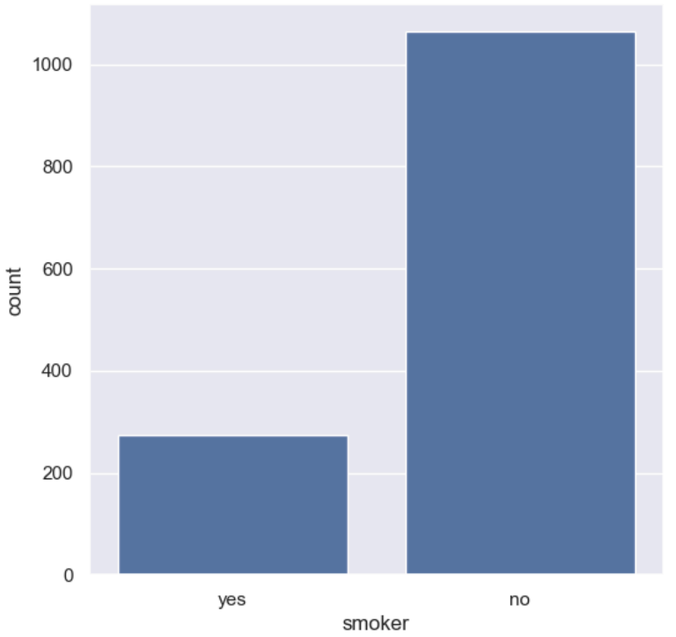

Sex Distribution
count
sex
0
100
200
300
400
500
600
700
female
male

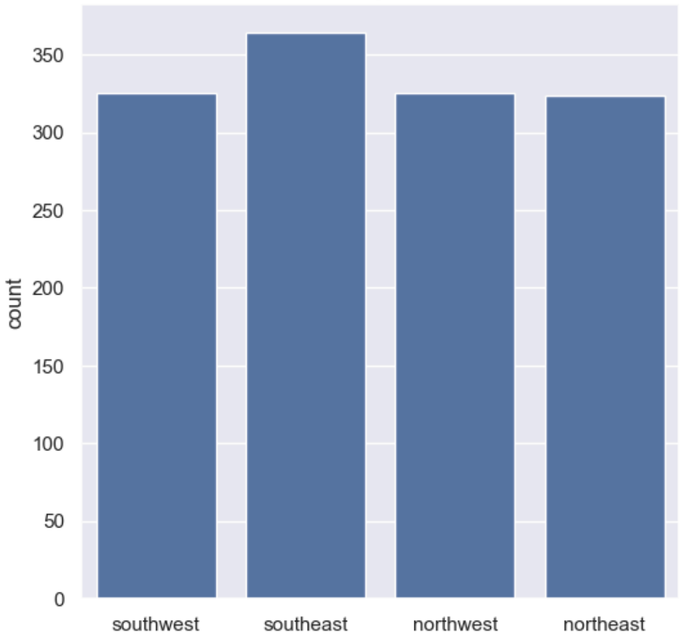

Bmi distribution
Count
bmi
0
20
40
60
80
100
120
140
20
30
40
50


Children
count
0
100
200
300
400
500
600
0
1
2
3
4
5

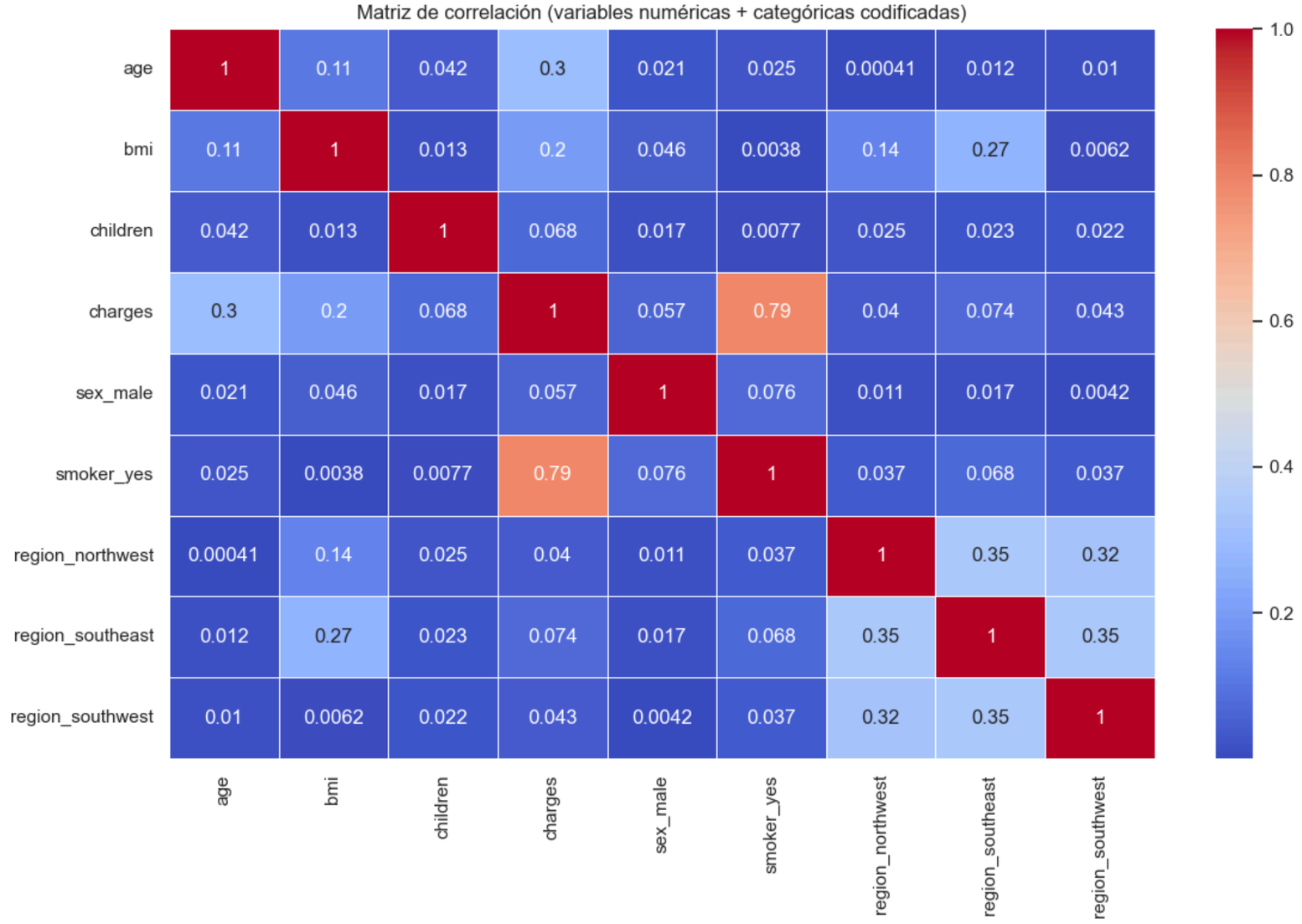

Charges distribution
Count
charges
0
25
50
75
100
125
150
175
200
0
10000
20000
30000
40000
50000
60000


0
10000
20000
30000
40000
50000
60000
charges


Smoker
count
smoker
0
200
400
600
800
1000
yes
no


Region
count
0
50
100
150
200
250
300
350
southwest
southeast
northwest
northeast

MATRIZ DE CORRELACION



os objetivos de la matriz:

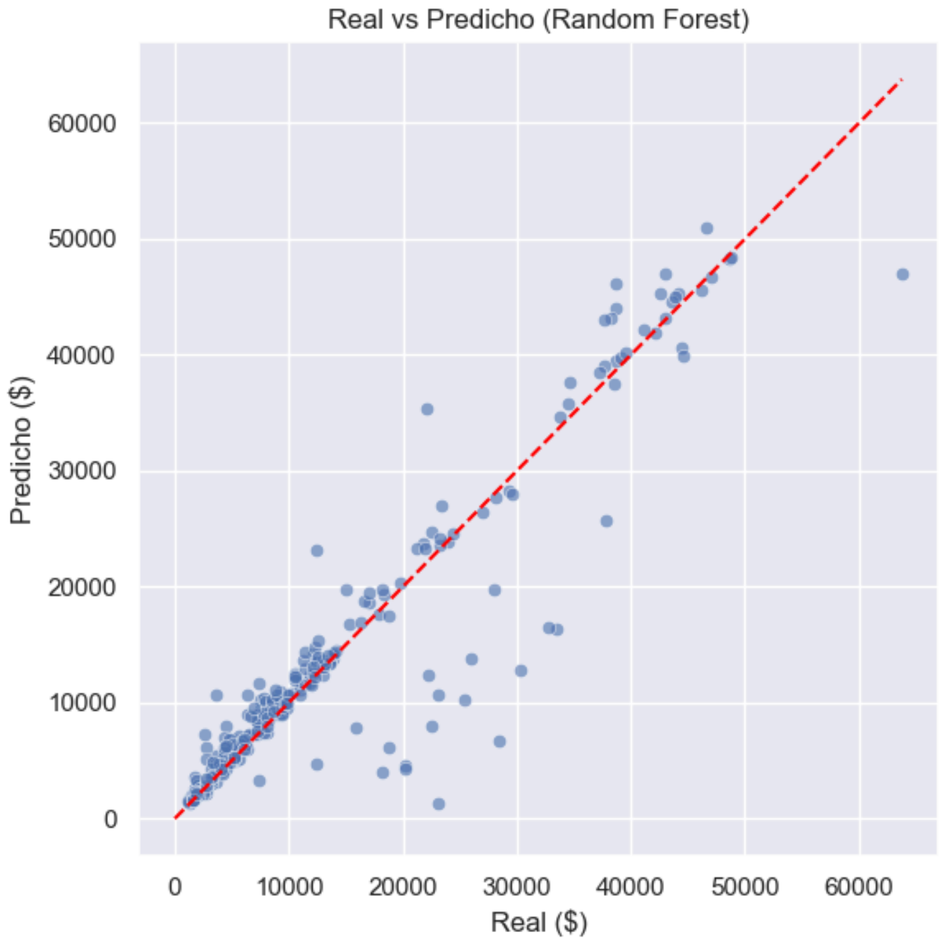
REGRESION LINEAL

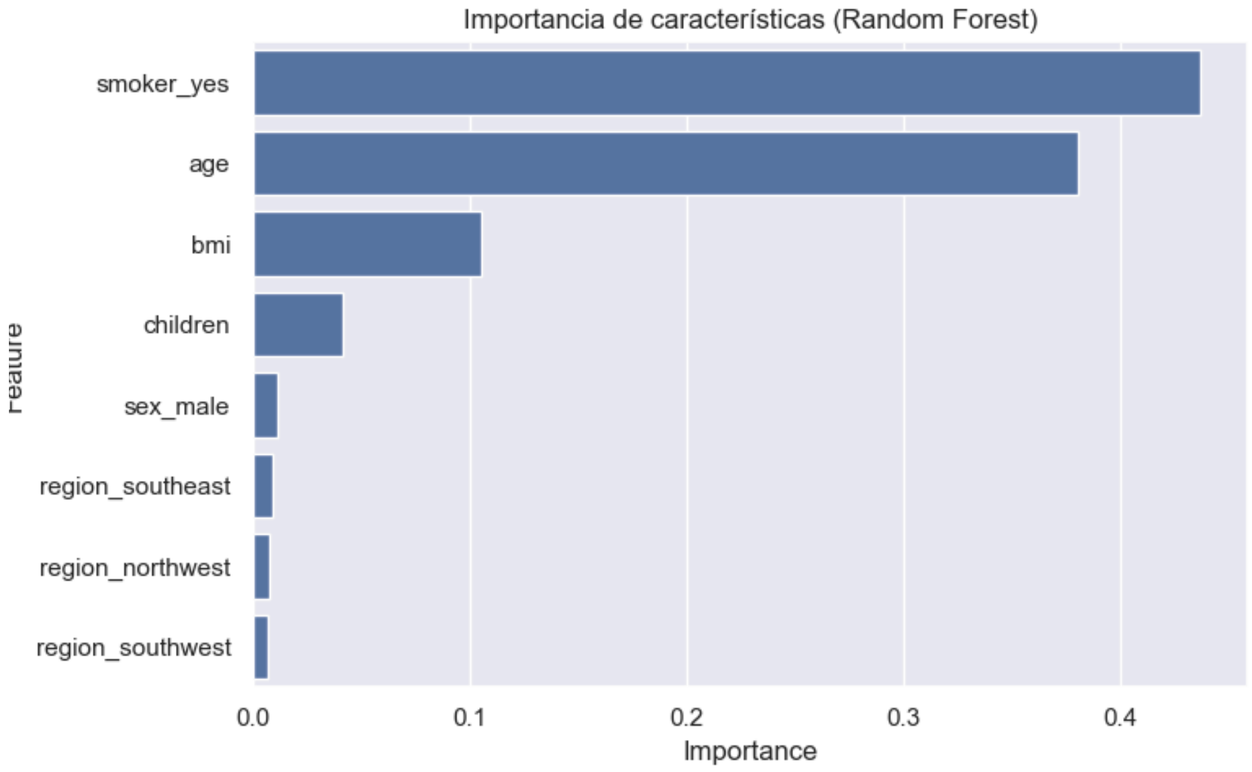
```
MAE (log): 0.2697
RMSE (log): 0.4189
R² (log): 0.8047
```

- MAE (log): 0.2697 Error absoluto medio en escala logarítmica.
- RMSE (log): 0.4189 Raíz del error cuadrático medio, también en logaritmos.
- R^2 (log): 0.8047 El modelo explica el 80.47% de la varianza de los datos, lo cual es bastante bueno para un modelo lineal.

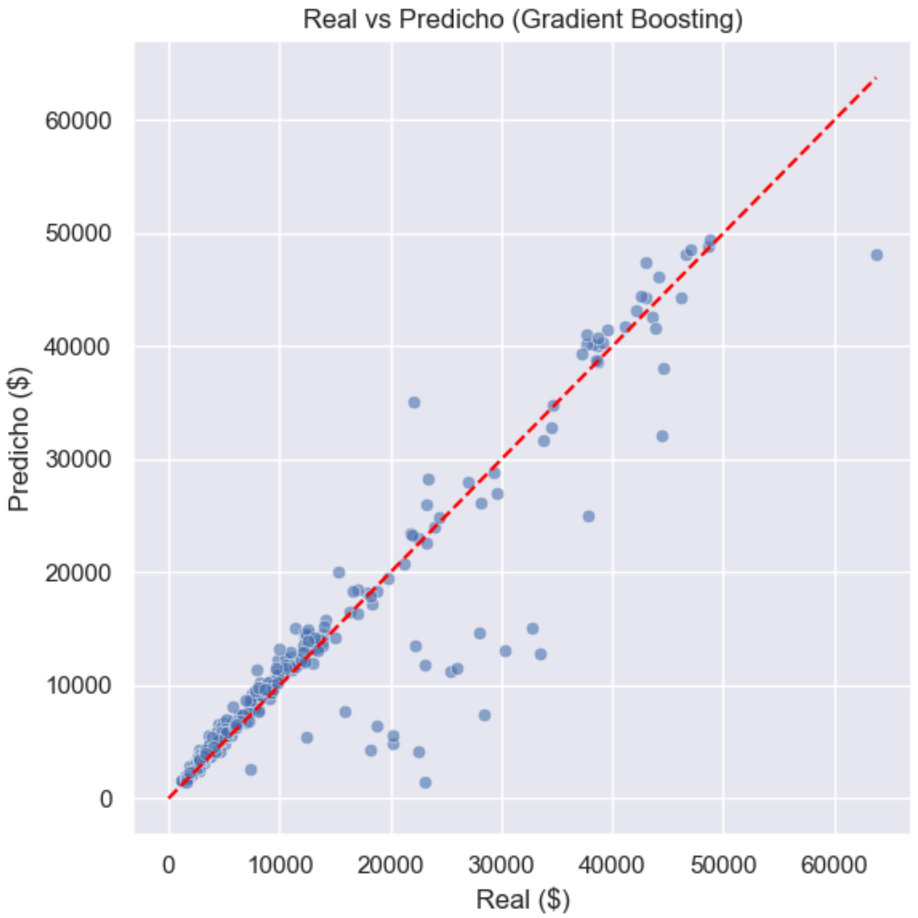
- Buen ajuste general: $R^2 > 0.8$ sugiere que el modelo logra capturar gran parte de la estructura de los datos.
- Error moderado: Las métricas MAE y RMSE indican que las predicciones son razonablemente cercanas a los valores reales (en la escala logarítmica).

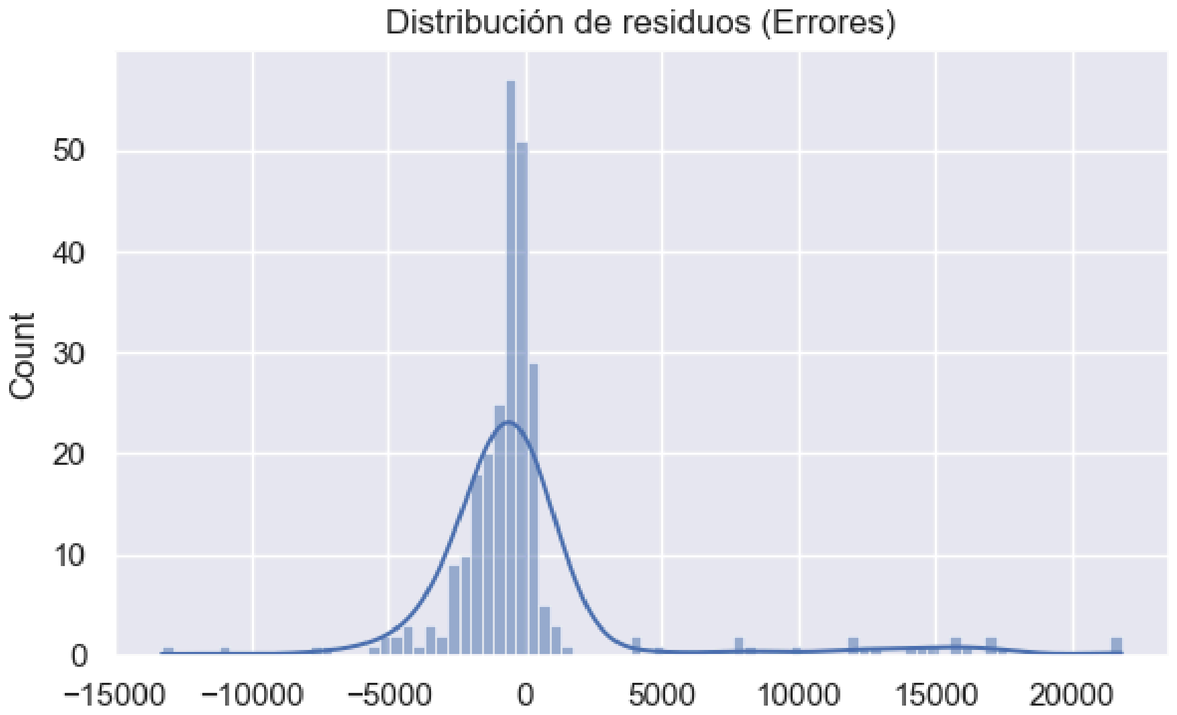
RANDOM FOREST





GRADIENT BOOSTING





COMPARACION ENTRE MODELOS

Modelo	MAE	RMSE	R^2
Regresión Lineal	3888.44	7814.06	0.6067
Random Forest	1951.56	3693.63	0.8482
Gradient Boosting	2029.37	4384.07	**0.8762**

CONCLUSIONES

- El análisis predictivo mediante machine learning permite estimar con buena precisión el costo del seguro médico, aportando una herramienta útil tanto para aseguradoras como para estudios de riesgo y salud.
- Las variables más influyentes en el costo fueron el IMC, el hábito de fumar y la edad, lo que valida la lógica del modelo y su alineación con factores reales de riesgo médico.
- Aunque la Regresión Lineal ofreció una base comprensible y rápida de implementar, su capacidad predictiva fue limitada frente a métodos más avanzados.
- En contraste, los modelos Random Forest y Gradient Boosting demostraron mejor desempeño, siendo este último el más preciso.
- El uso de pipelines de preprocesamiento y la modularización del flujo de trabajo mejoraron la reproducibilidad del proyecto, haciendo que el modelo final pueda reutilizarse fácilmente en entornos reales o integrarse en una aplicación.
- Finalmente, este proyecto demuestra que, con un conjunto de datos bien estructurado y técnicas adecuadas, es posible construir soluciones inteligentes que optimicen procesos de toma de decisiones en sectores como la salud y los seguros.

THE BRIDGE DIGITAL TALENT ACCELERATOR

¡GRACIAS!