

## Segundo cuatrimestre:

En el marco del módulo tercer cuatrimestre de la Tecnicatura Superior en Ciencia de Datos e Inteligencia Artificial, realizamos un análisis descriptivo y visualización de ventas de PepsiCo en Argentina. El trabajo nos permitió aplicar fundamentos de análisis estadístico y de procesamiento de datos en un entorno colaborativo, integrando la práctica con las competencias del rol de Analista de Datos (módulo que unifica materia de Ciencia de Datos I y Estadística I).

En el aspecto estadístico, se observa la aplicación de técnicas descriptivas básicas, utilizadas para resumir y comprender la información inicial del conjunto de datos. Esto incluyó el cálculo de medidas de tendencia central (media, mediana, moda) y de dispersión (rango, varianza, desviación estándar), herramientas esenciales para identificar patrones, valores extremos y la distribución general de las variables. Estas operaciones permitieron caracterizar el comportamiento de los datos y sentar las bases para futuros análisis inferenciales. Además, empleamos representaciones gráficas —como histogramas y diagramas de dispersión— para visualizar relaciones entre variables y detectar posibles correlaciones.

También aplicamos procedimientos de limpieza y normalización para la detección de valores nulos, la eliminación de duplicados y la unificación de formatos. Estos pasos son fundamentales dentro del flujo de trabajo de analítica, ya que garantizan la calidad y coherencia del dataset. También se utilizaron métodos de análisis exploratorio (EDA), centrados en describir las relaciones y variaciones de los datos sin modelado predictivo, etapa crucial para entender el contexto antes de aplicar modelos más complejos. Estos mismos pasos aplicamos para el análisis estadístico de diabetes, proyecto del segundo cuatrimestre.

El proyecto, por ser los primeros pasos y ejecuciones, se mantuvo dentro de un enfoque exploratorio y descriptivo, sin llegar a técnicas avanzadas de regresión o clasificación. Es importante aclarar que no fue continuado en el segundo cuatrimestre, dado que el equipo inició un nuevo proyecto orientado a un análisis más profundo y al uso de técnicas predictivas. Aun así, esta primera experiencia resultó clave para afianzar las competencias en estadística aplicada, manejo de datos y trabajo colaborativo, estableciendo una base sólida para los proyectos posteriores en la formación.

---

# DIABETES

## Análisis Estadístico: Elección y Aplicación de técnicas

**Integrantes:**

- Guillén Jonathan
- Majzum Maia
- Oviedo Francisco
- Pich Valentina
- Palomeque Jonathan Manuel
- Eglimar Ramírez



## 1. Descripción del Dataset

El conjunto de datos de la diabetes es una recopilación de datos médicos y demográficos de los pacientes, junto con su estado de diabetes (positivo o negativo). Los datos incluyen características como edad, sexo, índice de masa corporal (IMC), hipertensión, enfermedades cardíacas, antecedentes de tabaquismo, nivel de HbA1c y nivel de glucosa en sangre. Este conjunto de datos se puede utilizar para crear modelos de aprendizaje automático para predecir la diabetes en pacientes en función de su historial médico y su información demográfica. Esto puede resultar útil para los profesionales de la salud a la hora de identificar pacientes que pueden estar en riesgo de desarrollar diabetes y desarrollar planes de tratamiento personalizados. Además, los investigadores pueden utilizar el conjunto de datos para explorar las relaciones entre diversos factores médicos y demográficos y la probabilidad de desarrollar diabetes.

El dataset utilizado esta disponible de forma pública en el portal web de Kaggle, la fuente es la siguiente: <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>

---

## Objetivos y Aclaraciones

El presente trabajo, parte de un dataset cuyo propósito principal es la clasificación binaria: predecir si una persona presenta o no diabetes en función de diferentes factores de riesgo. Variables como la edad, la hipertensión, los niveles de glucosa en sangre, el hábito de fumar o la presencia de cardiopatía se utilizan habitualmente para construir modelos de predicción orientados a esta tarea.

Sin embargo, para efectos de la consigna académica, se tomarán algunas decisiones metodológicas libres que permiten aplicar y ejemplificar distintas técnicas estadísticas, tales como la correlación y la regresión lineal simple. Esto implica que, además de explorar ciertas asociaciones relacionadas con la diabetes, se analizarán también relaciones entre otras variables numéricas del dataset que no necesariamente forman parte directa del problema de clasificación, pero que sirven para ilustrar el uso correcto de dichas técnicas.

De este modo, los objetivos específicos son:

- Mostrar criterios de selección de técnicas estadísticas (paramétricas y no paramétricas) según el tipo de variable objetivo y de predictores.
- Aplicar pruebas de correlación y modelos de regresión lineal simple en pares de variables numéricas, justificando su elección en base a la exploración de datos.
- Presentar resultados clave sobre la diabetes, destacando la importancia de algunos factores de riesgo y su potencial utilidad en el ámbito clínico.

## 2. Librerías e Importación de Datos

```
[ ] 
# Importar librerias

# Tratamiento de datos
# =====
=====

import numpy as np
import pandas as pd

# Gráficos
# =====
=====

import matplotlib.pyplot as plt
import seaborn as sns

# Preprocesado y análisis
```

```
# =====
=====
from scipy import stats
import statsmodels.api as sm
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.diagnostic import het_breuschpagan
from scipy.stats import jarque_bera
from scipy.stats import pearsonr, spearmanr
from sklearn.model_selection import train_test_split
from sklearn.metrics import root_mean_squared_error

# Configuración warnings
# =====
=====
import warnings
warnings.filterwarnings("ignore")

[ ]
from google.colab import drive
drive.mount('/content/drive')
Mounted at /content/drive

[ ]
# Cargamos el archivo 'diabetes_dataset.csv' en un dataframe
df = pd.read_csv('/content/drive/MyDrive/ISPC CIENCIA DE DATOS/diabetes_dataset.csv')
df.head()
```

### 3. Análisis exploratorio (EDA)

#### Tamaño de los Datos

```
[ ]
df.shape
(100000, 9)
```

**Comentario:** El dataset consta de 100000 filas y 9 columnas

#### Columnas

```
[ ]
```

```
df.columns.values
array(['gender', 'age', 'hypertension', 'heart_disease',
       'smoking_history', 'bmi', 'HbA1c_level', 'blood_glucose_level',
       'diabetes'], dtype=object)
```

---

### Comentarios:

- El género (gender) se refiere al sexo biológico del individuo, que puede tener un impacto en su susceptibilidad a la diabetes. Hay tres categorías: masculino, femenino y otras.
- La edad (age) es un factor importante ya que la diabetes se diagnostica con mayor frecuencia en adultos mayores. La edad oscila entre 0 y 80 años en nuestro conjunto de datos.
- La hipertensión (hypertension) es una afección médica en la que la presión arterial en las arterias está elevada persistentemente. Tiene valores 0 o 1 donde 0 indica que no tiene hipertensión y 1 significa que tiene hipertensión.
- La enfermedad cardíaca (heart\_disease) es otra condición médica que se asocia con un mayor riesgo de desarrollar diabetes. Tiene valores 0 o 1 donde 0 indica que no tienen enfermedad cardíaca y 1 significa que tienen enfermedad cardíaca.
- El historial de tabaquismo (smoking\_history) también se considera un factor de riesgo para la diabetes y puede exacerbar las complicaciones asociadas con la diabetes. En nuestro conjunto de datos tenemos 6 categorías: no actualmente, anteriormente, sin información, actualmente, nunca y jamás.
- El IMC (índice de masa corporal) (bmi) es una medida de la grasa corporal basada en el peso y la altura. Los valores más altos de IMC están relacionados con un mayor riesgo de diabetes. El rango de IMC en el conjunto de datos es de 10.16 a 71.55. Un IMC inferior a 18.5 indica bajo peso, entre 18.5 y 24.9 es normal, entre 25 y 29.9 indica sobrepeso y 30 o más indica obesidad.
- El nivel de HbA1c (hemoglobina glicosilada) (HbA1c\_level) es una prueba de sangre que mide el nivel promedio de glucosa en la sangre durante los últimos 2 o 3 meses. se reporta en porcentaje (%). De acuerdo con los criterios clínicos, un valor de 6.5% o más es consistente con un diagnóstico de diabetes, mientras que valores entre 5.7% y 6.4% se asocian a prediabetes.
- El nivel de glucosa en sangre (blood\_glucose\_level) se mide en miligramos por decilitro (mg/dL) e indica la cantidad de glucosa presente en el torrente sanguíneo en un momento específico. En condiciones clínicas habituales, valores en ayunas menores a 100 mg/dL se consideran normales, entre 100 y 125 mg/dL corresponden a

prediabetes y niveles iguales o superiores a 126 mg/dL sugieren diabetes. Asimismo, una medición aleatoria de 200 mg/dL o más también es indicativa de la enfermedad.

- La diabetes (diabetes) es la variable objetivo que se predice, donde los valores de 1 indican la presencia de diabetes y 0 indican la ausencia de diabetes.
- 

## Tipo de Datos

```
[ ] # Mostrar la información del dataframe
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  --  
0   gender          100000 non-null   object 
1   age              100000 non-null   float64
2   hypertension     100000 non-null   int64  
3   heart_disease    100000 non-null   int64  
4   smoking_history  100000 non-null   object 
5   bmi              100000 non-null   float64
6   HbA1c_level     100000 non-null   float64
7   blood_glucose_level 100000 non-null   int64  
8   diabetes         100000 non-null   int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

---

### Comentarios:

- La data incluye variables numéricas ['age', 'bmi', 'HbA1c\_level', 'blood\_glucose\_level'] y categóricas ['gender', 'hypertension', 'heart\_disease', 'smoking\_history', 'diabetes']
  - Las variables categóricas ['hypertension', 'heart\_disease', 'diabetes'], se encuentran codificadas, por lo que las toma como int64.
  - Con el conteo de no nulos, se observa que no hay valores faltantes en el conjunto de datos.
- 

## Registros Duplicados

```
[ ] # cantidad de filas duplicadas
df.duplicated().sum()
np.int64(3854)
```

```
[ ]  
# Mostrar filas duplicadas  
df[df.duplicated()].head()
```

---

```
[ ]  
# Eliminar Duplicados  
df = df.drop_duplicates()  
  
# Validar que se eliminaron  
df.duplicated().sum()  
np.int64(0)
```

---

```
[ ]  
df.shape  
(96146, 9)
```

---

### Comentario:

Se encontraron 3854 registros duplicados, los cuales fueron eliminados, para finalmente obtener un dataset con 96146 filas y 9 columnas.

---

## Explorando Variables Categóricas

---

```
[ ]  
cat_vars = ['gender', 'hypertension', 'heart_disease', 'smoking_history', 'diabetes']  
  
# Diccionario con frecuencias absolutas y relativas  
freqs = {}  
for col in cat_vars:  
    abs_freq = df[col].value_counts(dropna=False)  
    rel_freq = df[col].value_counts(normalize=True, dropna=False) * 10  
    0  
    freqs[col] = pd.DataFrame({'Frecuencia': abs_freq, 'Porcentaje': r  
el_freq.round(2)})  
  
# Mostrar tablas individuales  
for col, tabla in freqs.items():  
    print(f"\n### {col}")  
    display(tabla)
```

```
[ ] # Eliminar registros donde gender == "Other"
df = df[df["gender"].isin(["Female", "Male"])]  
  

#Actualizar conteo
conteo_genero = df['gender'].value_counts()
print(conteo_genero)
gender
Female      56161
Male        39967
Name: count, dtype: int64  
  

[ ] col = "gender"  
  

abs_freq = df[col].value_counts(dropna=False)
rel_freq = df[col].value_counts(normalize=True, dropna=False) * 100  
  

tabla = pd.DataFrame({
    "Frecuencia": abs_freq,
    "Porcentaje": rel_freq.round(2)
})  
  

print(f"\n## {col}")
display(tabla)
```

---

### Comentario:

En la variable 'gender' se detecta una categoría 'other'= 'otro', dado que no suma al análisis porque son pocos registros (<0.02%), los eliminamos.

---

### Visualización de Variables Categóricas

---

```
[ ] cat_vars = ['gender', 'hypertension', 'heart_disease', 'smoking_history', 'diabetes']  
  

fig, axes = plt.subplots(2, 3, figsize=(16, 10))
axes = axes.flatten()  
  

for i, col in enumerate(cat_vars):
    data = df[col].value_counts(normalize=True).mul(100).reset_index()
    data.columns = [col, "Porcentaje"]
```

```

data = data.sort_values("Porcentaje", ascending=False)

sns.barplot(x=col, y="Porcentaje", data=data, palette="Set2", ax=axes[i])

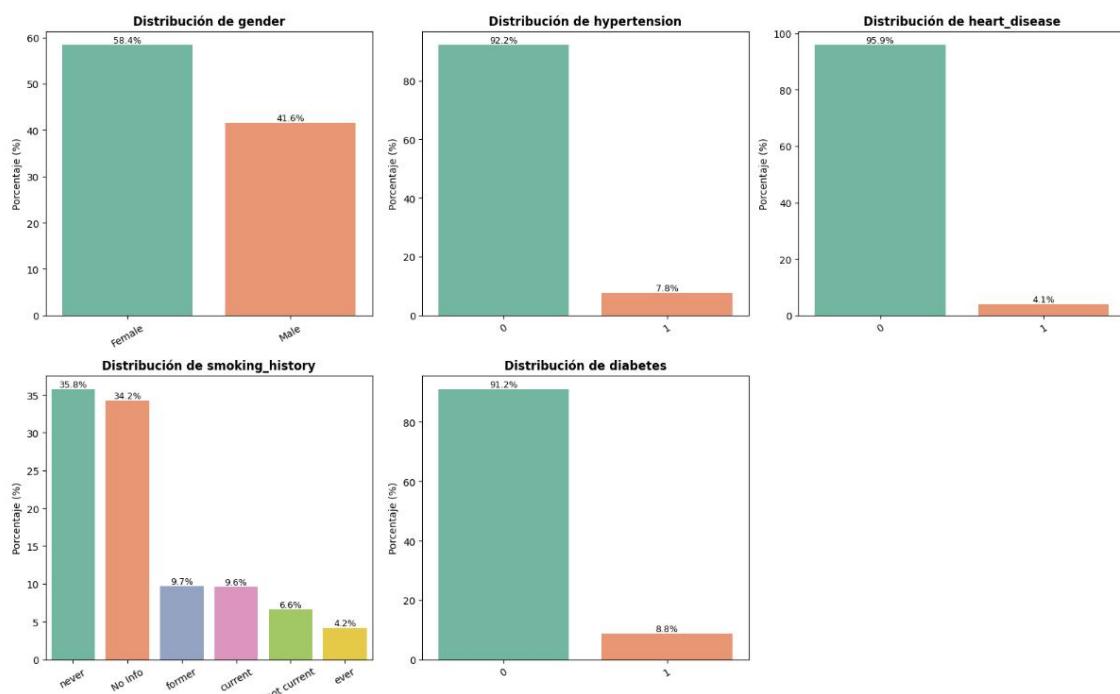
# Etiquetas en cada barra
for p in axes[i].patches:
    axes[i].annotate(f"{p.get_height():.1f}%", (p.get_x() + p.get_width() / 2., p.get_height()),
                      ha="center", va="bottom", fontsize=9)

axes[i].set_title(f"Distribución de {col}", fontsize=12, weight="bold")
axes[i].set_ylabel("Porcentaje (%)")
axes[i].set_xlabel("")
axes[i].tick_params(axis='x', rotation=30)

# Eliminar subplot vacío si sobran casillas
if len(cat_vars) < len(axes):
    fig.delaxes(axes[-1])

plt.tight_layout()
plt.show()

```



## Interpretación de variables categóricas

- Género: La muestra incluye 56,161 mujeres (58.4%) y 39,967 hombres (41.6%). Existe un ligero predominio femenino, aunque no supone un desbalance crítico y permite comparaciones válidas entre ambos grupos.
  - Hipertensión: El 92.2% de los registros corresponde a personas sin hipertensión y solo el 7.8% a personas con esta condición. Aunque es un grupo minoritario, la hipertensión sigue siendo un factor de riesgo relevante en el desarrollo de diabetes.
  - Enfermedad cardíaca: La mayoría (95.9%) no presenta cardiopatía, mientras que el 4.1% sí. Pese a su baja proporción, esta variable resulta importante en el análisis por su asociación con complicaciones metabólicas y cardiovasculares.
  - Historial de tabaquismo: La distribución es más heterogénea: un 35.8% nunca fumó, un 9.7% son exfumadores, un 9.6% fumadores actuales, un 6.6% “no actuales” y un 4.2% “ever”. Además, un 34.2% está clasificado como No Info, lo que indica que no se dispone de información sobre el historial de tabaquismo para ese grupo. Aunque no se trata de valores faltantes, esta categoría limita el análisis detallado de la relación entre tabaquismo y diabetes.
  - Diabetes (variable objetivo): El dataset está fuertemente desbalanceado: el 91.2% no presenta diabetes frente a un 8.8% con diagnóstico positivo. Este desbalance debe considerarse en los análisis y modelos predictivos, ya que puede sesgar los resultados hacia la clase mayoritaria.
- 

## Conclusión

El dataset representa una población mayoritariamente sin hipertensión, cardiopatía ni diabetes, con un ligero predominio de mujeres. El tabaquismo aparece como un factor con categorías variadas y un grupo amplio sin información disponible (No Info). El aspecto más relevante es el desbalance en la variable objetivo (diabetes), que constituye una consideración clave tanto para los análisis estadísticos como para la construcción de modelos de machine learning.

---

## Explorando Variables Numéricas

---

### Resumen Estadístico

---

```
[ ] 
# Variables cuantitativas
num_vars = ["age", "bmi", "HbA1c_level", "blood_glucose_level"]

# Tabla descriptiva extendida
desc = df[num_vars].describe(percentiles=[0.01, 0.1, 0.15, 0.25, 0.5, 0.75,
0.9, 0.99]).T

# Añadir skewness y kurtosis
desc["skew"] = df[num_vars].skew()
desc["kurtosis"] = df[num_vars].kurtosis()

# Redondear para mejor visualización
desc = desc.round(2)

desc
```

---

	count	mean	std	min	1%	10%	15%	25%	50%	75%	90%	99%	max	skew	kurtosis
age	96128.0	41.80	22.46	0.08	1.00	10.0	15.0	24.0	43.00	59.00	73.0	80.00	80.00	-0.06	-1.00
bmi	96128.0	27.32	6.77	10.01	14.55	19.0	20.8	23.4	27.32	29.86	35.7	48.97	95.69	1.02	3.27
HbA1c_level	96128.0	5.53	1.07	3.50	3.50	4.0	4.0	4.8	5.80	6.20	6.6	8.80	9.00	-0.05	0.24
blood_glucose_level	96128.0	138.22	40.91	80.00	80.00	85.0	90.0	100.0	140.00	159.00	200.0	280.00	300.00	0.84	1.76

### Interpretación por variable:

#### ◊ Edad (age)

Media: 41.8 años.

Mediana (50%): 43 años, muy cerca de la media, lo que indica simetría.

Mínimo–Máximo: 0.08 a 80 años

Skew = -0.06, prácticamente simétrica.

Curtosis = -1.0, distribución más “aplanada” que la normal (platicúrtica).

La edad en la muestra está bien distribuida, sin sesgo.

#### ◊ Índice de Masa Corporal (bmi)

Media: 27.3 (categoría sobrepeso, OMS).

Mediana: 27.3, igual a la media, lo que sugiere cierta simetría central.

Percentiles: el 75% está debajo de 29.8, pero el 99% llega a 48.9 y el máximo a 95.7, hay outliers extremos.

Skew = 1.02, distribución sesgada a la derecha.

Curtosis = 3.27, leptocúrtica, con colas más pesadas que la normal.

la mayoría tiene BMI entre 20 y 35, pero existen valores muy altos que empujan la distribución hacia la derecha.

◊ **HbA1c (HbA1c\_level)**

Media: 5.53% , dentro de lo normal (diabetes  $\geq 6.5\%$ ).

Mediana: 5.8%, muy próxima a la media.

Rango: 3.5 a 9, cubre desde niveles bajos hasta valores compatibles con diabetes.

Skew = -0.05, simétrica.

Curtosis = 0.24, cercana a la normal.

Variable bien comportada, sin sesgo fuerte; la mayoría de los paciente s no llega al umbral de diabetes por HbA1c.

◊ **Glucosa en sangre (blood\_glucose\_level)**

Media: 138.2 mg/dL (un poco por encima del valor normal en ayunas < 126 mg/dL).

Mediana: 140 mg/dL, muy próxima a la media.

Rango: 80 a 300, incluye tanto valores normales como casos severos.

Skew = 0.84, distribución sesgada a la derecha (colas largas).

Curtosis = 1.76, leptocúrtica, con colas más pesadas que la normal.

Distribución con valores extremos hacia arriba, consistente con alguno s pacientes con hiperglucemia importante.

---

## Conclusiones

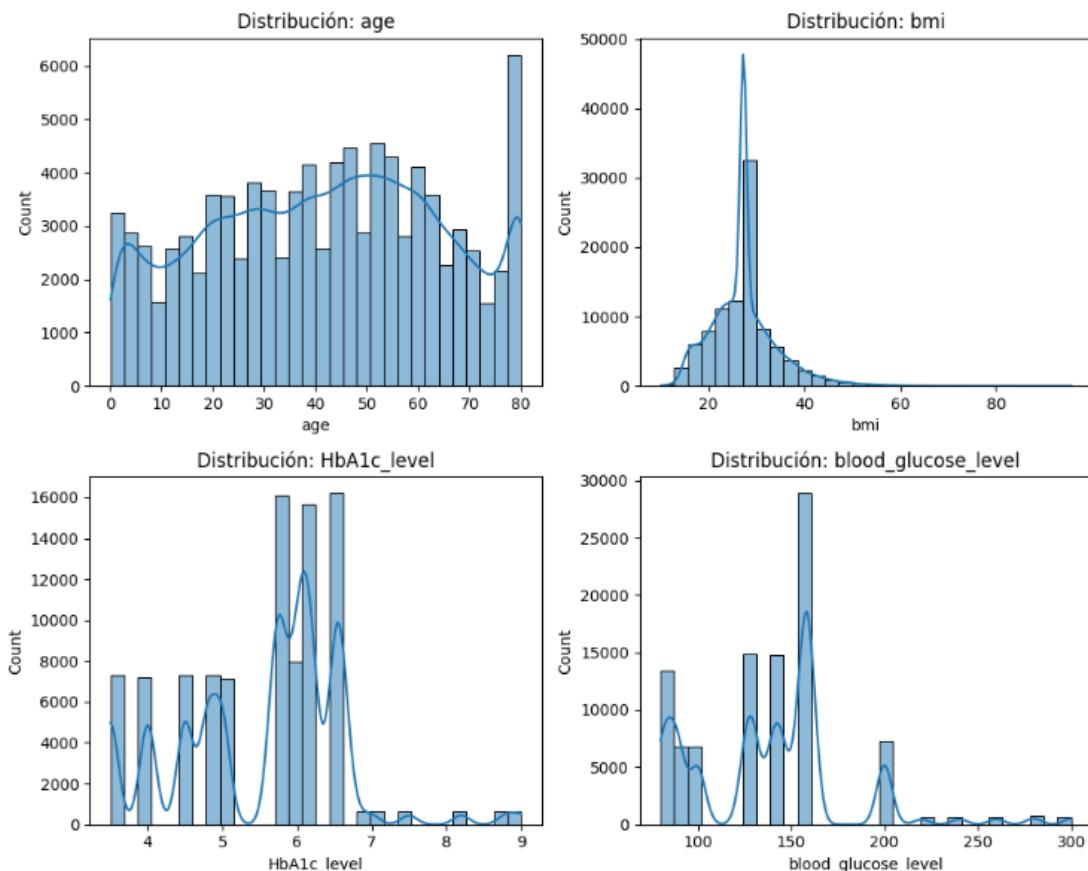
- Edad y HbA1c: variables parecen bastante simétricas y normales.
  - BMI y Glucosa: variables sesgadas a la derecha, con outliers extremos que pueden influir en correlaciones y regresiones.
  - Curtosis positiva en BMI y Glucosa: colas más pesadas, riesgo de outliers que distorsionen el análisis.
- 

## Recomendaciones en Análisis Inferencial:

- Probar normalidad formal (Shapiro, Kolmogorov-Smirnov, Anderson-Darling).
- Usar transformaciones (log) o pruebas no paramétricas si los outliers influyen mucho.

## Distribución de Variables

```
[ ] 
# Creando las visualizaciones de las distribuciones
fig, axs = plt.subplots(2, 2, figsize=(10, 8))
axs = axs.ravel()
for i, v in enumerate(num_vars):
    sns.histplot(df[v].dropna(), bins=30, kde=True, ax=axs[i])
    axs[i].set_title(f'Distribución: {v}')
plt.tight_layout()
plt.show()
```



## Análisis:

En la tabla de estadísticos inicial observamos que edad y HbA1c presentaban valores de asimetría cercanos a cero y curtosis bajas, lo que sugería distribuciones relativamente normales. Esto se corrobora en los histogramas, donde ambas variables muestran una forma simétrica y sin colas extremas relevantes.

Por otro lado, tanto IMC como glucosa en sangre aparecían con asimetría positiva y curtosis elevada en la tabla, lo que anticipaba distribuciones sesgadas con presencia de valores extremos. Los histogramas confirman esta situación: el IMC se concentra en torno a 20–35 pero con colas largas hacia valores muy altos, y la glucosa se centra alrededor de 140 mg/dL con casos severos que alcanzan 300 mg/dL.

En conclusión, la inspección visual de los histogramas respalda lo cuantificado en la tabla: edad y HbA1c son más estables y cercanas a la normalidad, mientras que IMC y glucosa muestran sesgo a la derecha y outliers que deberán considerarse en análisis posteriores.

---

## Boxplot de las Variables

---

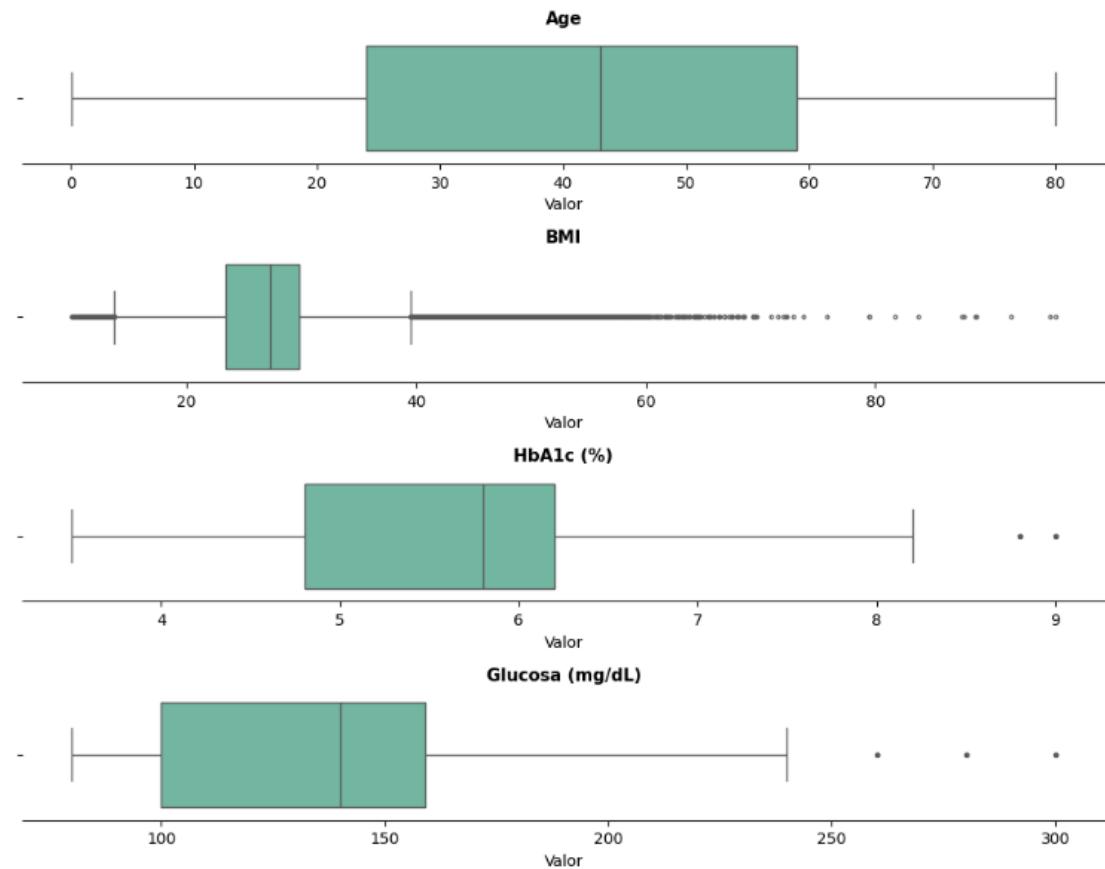
```
[ ] vars_plot = [
    ("age", "Age"),
    ("bmi", "BMI"),
    ("HbA1c_level", "HbA1c (%)"),
    ("blood_glucose_level", "Glucosa (mg/dL)"),
]

fig, axes = plt.subplots(nrows=4, ncols=1, figsize=(10, 8), sharex=False)

for ax, (col, label) in zip(axes, vars_plot):
    sns.boxplot(x=df[col], orient="h", ax=ax, color=sns.color_palette("Set2", 4)[0], fliersize=2)
    ax.set_title(label, fontsize=11, fontweight="bold")
    ax.set_ylabel("")
    ax.set_xlabel("Valor")

sns.despine(left=True)
plt.tight_layout()
plt.show()
```

---



### Interpretación por Variable:

#### ◊ Age

La edad es estable y bastante normal

El rango intercuartílico (IQR) es amplio, lo que indica variabilidad moderada.

No se observan outliers relevantes.

La edad es estable y bastante normal

#### ◊ BMI

La caja está comprimida entre 20 y 35 (la mayoría de los valores).

Aparecen muchísimos outliers a la derecha, que llegan hasta 96. Que confirma sesgo positivo y colas largas.

la mayoría está en sobrepeso/obesidad moderada, pero los outliers extremos dominan visualmente y pueden distorsionar análisis paramétricos.

### ◊ HbA1c\_level

Caja más equilibrada, sin sesgo marcado.

Se ven algunos outliers, pero en menor cantidad y no tan extremos.

la variable se ve bastante controlada, pero con ciertos valores elevados que conviene vigilar.

### ◊ Blood\_glucose\_level

La caja se concentra entre 100 y 160 mg/dL.

Se destacan varios outliers altos. Esto sugiere una distribución con sesgo positivo y valores atípicos que reflejan casos de glucosa elevada en la muestra.

la mayoría está cerca del rango limítrofe de diabetes (126-140 mg/dL), pero hay casos con glucosa muy elevada que generan colas pesadas.

## Conclusiones

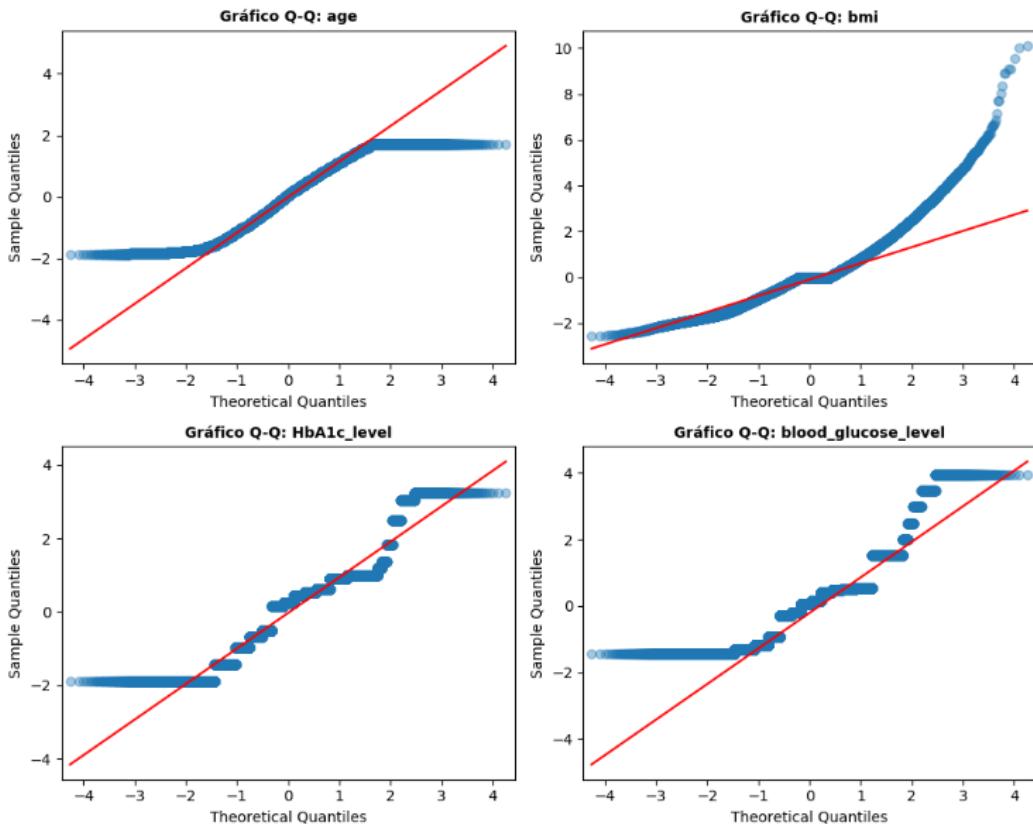
- Age y HbA1c: distribuciones más regulares, sin gran presencia de outliers.
- BMI y Glucosa: variables con clara presencia de outliers y sesgo a la derecha, lo que confirma lo visto en histogramas y en los estadísticos (asimetría positiva, curtosis alta).
- En análisis posteriores (correlaciones, regresión), habrá que evaluar si se tratan los outliers (winsorización, transformaciones logarítmicas) o si se emplean métodos robustos/no paramétricos.

## Prueba de Normalidad (Q-Q plots)

```
[ ] 
fig, axes = plt.subplots(2, 2, figsize=(10, 8))

for i, var in enumerate(num_vars):
    sm.qqplot(df[var], line='q', fit=True, ax=axes[i//2, i%2], alpha=0.4, lw=2)
    axes[i//2, i%2].set_title(f"Gráfico Q-Q: {var}", fontsize=10, fontweight="bold")

plt.tight_layout()
plt.show()
```



### Interpretación de los Q-Q plots:

#### 1. Age

Los puntos siguen bastante bien la línea roja, salvo leves desviaciones en colas. Distribución aproximadamente normal.

#### 2.BMI

Los puntos se separan mucho de la línea en la cola superior (valores altos). No hay normalidad, sesgo positivo y muchos outliers.

#### 3.HbA1c\_level

Se ven escalones (efecto de valores discretos) y cierta desviación en colas. Aunque no es perfectamente normal, no tiene el sesgo extremo de BMI.

#### 4.Blood\_glucose\_level

Igual que HbA1c, escalones y desviaciones marcadas en colas. Muy alejado de la normalidad.

- **Conclusión:** Los gráficos Q-Q corroboran que age y HbA1c tienen distribuciones más estables, mientras que BMI y blood\_glucose\_level presentan asimetría positiva marcada. Esto condiciona el uso de

pruebas paramétricas y justifica considerar transformaciones logarítmicas o correlaciones no paramétricas en los siguientes análisis.

---

## 4. Estudio de Correlación

---

### Enfoque para el Análisis de Correlación

---

#### 1. Volumen de datos:

- Con tamaños muestrales muy grandes, las pruebas formales de normalidad (Shapiro-Wilk, Kolmogorov-Smirnov) suelen rechazar el supuesto incluso ante desviaciones mínimas.
- Por ello, la decisión se apoya principalmente en la inspección visual (histogramas, Q-Q plots, boxplots) y en los estadísticos de asimetría y curtosis.

#### 2. Transformaciones:

- Cuando una variable presenta fuerte sesgo (ejemplo: BMI y glucosa en este dataset), se consideran transformaciones logarítmicas para mejorar simetría y linealidad.
- Se trabaja tanto con variables crudas como transformadas, comparando resultados.

#### 3. Selección del método de correlación:

- Pearson → adecuado cuando se busca medir relaciones lineales, aplicable a variables originales o transformadas si mejoran la normalidad.
- Spearman → recomendado en presencia de sesgo, outliers o distribuciones no normales, ya que mide relaciones monótonas.
- Kendall → menos sensible a outliers pero computacionalmente más costoso; se suele reservar para muestras pequeñas o confirmación adicional.

#### 4. Visualización:

- Se generan mapas de calor (heatmaps) para comparar matrices de correlación (Pearson vs Spearman).

- Se incluyen scatterplots con línea de tendencia para validar visualmente si la relación observada es efectivamente lineal o sólo monótona.

### 5. Selección de Variables:

- Se priorizan las variables que muestran correlaciones más altas y estables en ambos métodos.
- Si Spearman es claramente mayor que Pearson, se interpreta como una relación monotónica no lineal, y por lo tanto, la regresión lineal simple puede no ser el modelo más apropiado.

## Aplicación al Dataset de Diabetes

**1. Volumen:** al contar con más de 90.000 registros, los tests formales de normalidad pierden utilidad práctica.

**2. Distribuciones:** según histogramas, Q-Q plots y boxplots, las variables BMI y glucosa están sesgadas a la derecha con outliers, mientras que edad y HbA1c son más estables.

**3. Método recomendado:** Spearman es la opción más robusta y segura para evaluar asociaciones. Pearson se mantiene como referencia, sobre todo después de aplicar transformaciones logarítmicas a BMI y glucosa.

**Decisión final:** Spearman servirá como punto de partida para determinar pares relevantes, pero se contrastará con Pearson (crudo y transformado) para verificar si existe también evidencia de linealidad suficiente que justifique un modelo de regresión lineal.

## Matrices de Correlación

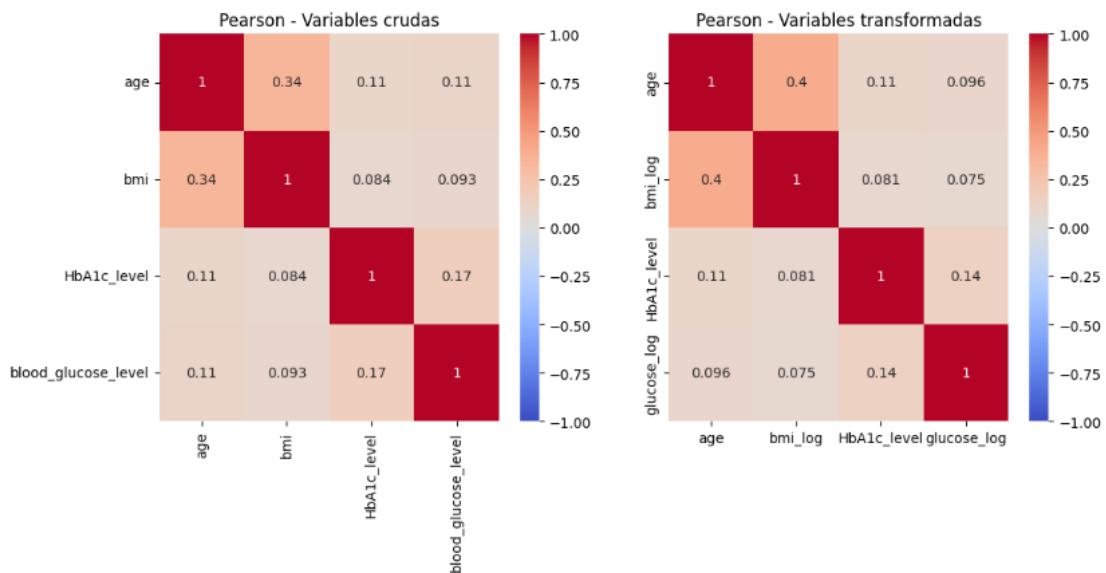
```
[ ]  
# Matriz Pearson con datos crudos  
corr_pearson_raw = df[num_vars].corr(method="pearson")  
  
# Matriz Pearson con variables transformadas (log)  
df_trans = df.copy()  
df_trans["bmi_log"] = np.log1p(df_trans["bmi"])  
df_trans["glucose_log"] = np.log1p(df_trans["blood_glucose_level"])  
corr_pearson_log = df_trans[["age", "bmi_log", "HbA1c_level", "glucose_log"]].corr(method="pearson")  
  
# Graficar  
fig, axes = plt.subplots(1, 2, figsize=(12,5))
```

```

sns.heatmap(corr_pearson_raw, annot=True, cmap="coolwarm", vmin=-1, vmax=1, ax=axes[0])
axes[0].set_title("Pearson - Variables crudas")

sns.heatmap(corr_pearson_log, annot=True, cmap="coolwarm", vmin=-1, vmax=1, ax=axes[1])
axes[1].set_title("Pearson - Variables transformadas")
plt.show()

```



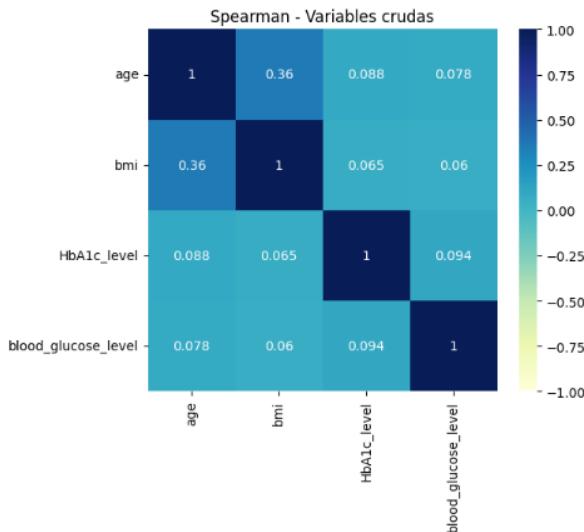

---

```

[ ]
corr_spearman = df[num_vars].corr(method="spearman")

plt.figure(figsize=(6,5))
sns.heatmap(corr_spearman, annot=True, cmap="YlGnBu", vmin=-1, vmax=1)
plt.title("Spearman - Variables crudas")
plt.show()

```



## Interpretación de Correlaciones

Al comparar los métodos de correlación (Pearson en crudo, Pearson con variables transformadas y Spearman), se observan las siguientes conclusiones clave:

- **Age–BMI:**

Pearson crudo: 0.34

Pearson log-transformado: 0.40

Spearman: 0.36

◊ Este par de variables muestra la asociación más consistente. La transformación logarítmica de BMI mejora la linealidad, y Spearman confirma que existe una relación monotónica positiva.

◊ Es el par más adecuado para exemplificar un modelo de regresión lineal.

- **HbA1c–Glucosa:**

Pearson crudo: 0.17

Pearson log-transformado: 0.14

Spearman: 0.094

◊ A pesar de su relación clínica conocida, en este dataset la correlación es baja. No se observa una relación fuerte ni estrictamente lineal.

### Otros pares (Age–HbA1c, BMI–Glucosa, etc.):

- ◊ Todas presentan correlaciones débiles ( $<0.15$ ), lo que indica ausencia de asociación relevante en este contexto.

### Conclusión

De acuerdo con los resultados:

Age–BMI es el único par que justifica un análisis de regresión lineal, al ser el que muestra mayor fuerza de asociación y mejora con transformaciones.

El resto de las combinaciones presentan correlaciones débiles, por lo que no resultan útiles para construir modelos lineales en este análisis exploratorio.

---

### Diagramas de dispersión con ajuste lineal para las variables seleccionadas

---

Para los scatterplots se seleccionaron las combinaciones con mayor correlación.

En Age–BMI, se utilizó BMI transformado en logaritmo, ya que la transformación mejoró la linealidad ( $r=0.40$ ).

En Glucosa–HbA1c, se usaron las variables en crudo, dado que la transformación no aportó mejoras y resulta más interpretable en sus unidades originales.

---

```
[ ]
# Hacer copia del df original
df_scatter = df.copy()

# Crear variable transformada solo en la copia
df_scatter["bmi_log"] = np.log1p(df_scatter["bmi"])      # log(1+x) para
evitar problemas con ceros

# Pares a graficar: mezclando crudo y transformado
pares = [
    ("age", "bmi_log", "Age vs BMI (log-transformado)"),
    ("blood_glucose_level", "HbA1c_level", "Glucosa vs HbA1c (crudo)")
]

# Generar scatterplots con regresión
fig, axes = plt.subplots(1, 2, figsize=(12,5))

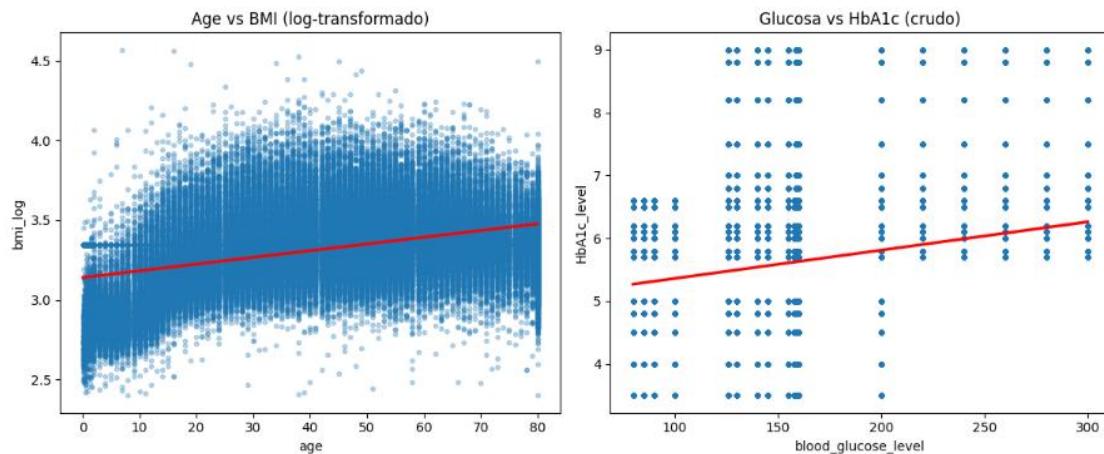
for i, (x, y, titulo) in enumerate(pares):
    sns.regplot(
```

```

data=df_scatter, x=x, y=y,
scatter_kws={'alpha':0.3, 's':10},    # puntos más pequeños y s
uaves
line_kws={'color':'red'},           # línea de tendencia
ax=axes[i]
)
axes[i].set_title(titulo)

plt.tight_layout()
plt.show()

```



## Análisis:

Se presentan los scatterplots para los pares de variables con mayor interés según los coeficientes de correlación (Pearson y Spearman):

- **Age vs BMI (log-transformado):** se observa una **tendencia lineal positiva moderada**, donde a mayor edad tiende a aumentar el índice de masa corporal. La transformación logarítmica ayudó a mejorar la simetría y a estabilizar la relación, lo que respalda su uso en un modelo de regresión lineal.
- **Glucosa (mg/dL) vs HbA1c (%):** existe una **tendencia creciente débil**, lo que coincide con la baja correlación hallada. Aunque clínicamente ambas variables deberían estar asociadas, en este dataset la dispersión de puntos es elevada y limita la capacidad explicativa de un modelo lineal directo.

En conclusión, los scatterplots confirman lo observado en las matrices de correlación:

- La relación **Edad–BMI (log)** es la más consistente para un análisis de regresión.
  - La relación **Glucosa–HbA1c**, aunque teóricamente esperable, aparece débil en los datos analizados.
- 

## 5. Regresión Lineal Simple (OLS)

---

**Modelo:** `bmi_log ~ age`

---

### 1. Definición y División de Datos

---

```
[ ] 
# Se utilizan las variables 'age' y 'bmi_log' del DataFrame 'df_scatter'

X = df_scatter['age']
y = df_scatter['bmi_log']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

---

### 2. Entrenamiento del Modelo

---

```
[ ] 
X_train_const = sm.add_constant(X_train, prepend=True)
modelo_bmi = sm.OLS(y_train, X_train_const).fit()
print(modelo_bmi.summary())
=====
              OLS Regression Results
=====
Dep. Variable:      bmi_log    R-squared:   0.165
Model:                 OLS        Adj. R-squared:  0.165
Method:                Least Squares   F-statistic:  1.519e+04
Date:           Sun, 21 Sep 2025   Prob (F-statistic):  0.00
Time:            17:51:36       Log-Likelihood:  9302.9
No. Observations:      76902      AIC:             -
Df Residuals:         76900      BIC:             -
Df Model:                  1
```

```
Df Model: 1
Covariance Type: nonrobust
=====
=====
```

	coef	std err	t	P> t	[0.025
0.975]					0.975]
const	3.1386	0.002	1924.405	0.000	3.135
3.142					
age	0.0042	3.44e-05	123.243	0.000	0.004
0.004					
Omnibus:	2752.899		Durbin-Watson:		
1.998					
Prob(Omnibus):	0.000		Jarque-Bera (JB):		
3869.175					
Skew:	0.370		Prob(JB):		
0.00					
Kurtosis:	3.811		Cond. No.		
100.					
=====					

---

**OLS Regression Results**

```
=====
=====
```

Dep. Variable:	bmi_log	R-squared:	0.165			
Model:	OLS	Adj. R-squared:	0.165			
Method:	Least Squares	F-statistic:	1.519e+04			
Date:	Sun, 21 Sep 2025	Prob (F-statistic):	0.00			
Time:	17:51:36	Log-Likelihood:	9302.9			
No. Observations:	76902	AIC:	-1.860e+04			
Df Residuals:	76900	BIC:	-1.858e+04			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
const	3.1386	0.002	1924.405	0.000	3.135	3.142
age	0.0042	3.44e-05	123.243	0.000	0.004	0.004
Omnibus:	2752.899		Durbin-Watson:			1.998
Prob(Omnibus):	0.000		Jarque-Bera (JB):			3869.175
Skew:	0.370		Prob(JB):			0.00
Kurtosis:	3.811		Cond. No.			100.
=====						

---

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

## Análisis:

### 1. Coeficientes del modelo

- **Intercepto (const = 3.1386, p < 0.001):**  
Representa el valor esperado de  $\log(1+\text{BMI})$  cuando la edad es 0. Es estadísticamente significativo.
  - **Coeficiente de age (0.0042, p < 0.001):**  
Cada año adicional de edad incrementa en promedio 0.0042 unidades el logaritmo del BMI. → En términos prácticos: a mayor edad, el BMI (log-transformado) tiende a aumentar ligeramente. Aproximadamente 0.42% por año.  
Ambos coeficientes son **estadísticamente significativos** ( $p < 0.05$ ).
- 

## 2. Bondad de ajuste

- **R-squared = 0.165 (16.5%)**  
La edad explica el 16.5% de la variabilidad en bmi\_log.  
→ Es bajo, lo que indica que hay muchos otros factores (no incluidos en este modelo) que influyen en el BMI.
  - **F-statistic = 1.519e+04, p < 0.001**  
El modelo global es significativo: al menos una variable predictora (aquí, age) aporta información sobre la variable respuesta.
- 

## 3. Pruebas de supuestos

- **Omnibus / Jarque-Bera (JB=3869, p < 0.001):**  
Los residuos **no son normales** (la hipótesis nula de normalidad se rechaza).
  - **Skew = 0.370 y Kurtosis = 3.811:**
    - Asimetría (Skew) positiva ligera → la distribución de residuos está un poco sesgada a la derecha.
    - Curtosis superior a 3 → hay colas más pesadas que la normal, posibles outliers.
  - **Durbin-Watson = 1.998:**  
Muy cercano a 2 → no hay autocorrelación de residuos.
- 

## 4. Intervalos de confianza

- Para age, el 95% CI es [0.004, 0.004], muy estrecho → el efecto estimado es estable y preciso.
- Para const, el 95% CI es [3.135, 3.142].

## 5. Ecuación del modelo

$$\text{bmi\_log}^{\wedge}=3.1386+0.0042 \cdot \text{age}$$


---

## 6. Conclusión general

- El modelo confirma que la edad está asociada positivamente con el BMI log-transformado.
  - Sin embargo, **la capacidad explicativa es baja (16.5%)**, lo que indica que deben incluirse más variables (ej. hábitos, genética, estilo de vida) para mejorar el ajuste.
  - Los residuos no cumplen normalidad estricta, aunque la gran cantidad de observaciones mitiga el problema.
- 

## 3. Visualización del modelo en TRAIN

```
[ ] # IC 95% de coeficientes
intervalos_ci = modelo_bmi.conf_int(alpha=0.05)
intervalos_ci.columns = ['2.5%', '97.5%']
display(intervalos_ci)

      2.5%    97.5%
const  3.135354  3.141747
age   0.004169  0.004303

[ ] # --- Predicciones sobre TRAIN con IC de la MEDIA ---
pred_train = modelo_bmi.get_prediction(X_train_const).summary_frame(alpha=0.05)
pred_train['x'] = X_train.values                      # age (train)
pred_train['y'] = y_train.values                      # bmi_log (train)
pred_train = pred_train.sort_values('x')             # importante para trazar la curva
pred_train.head(4)
```

---

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper	x	y
79692	3.138889	0.001628	3.135698	3.142081	2.718648	3.559131	0.08	3.343568
34663	3.138889	0.001628	3.135698	3.142081	2.718648	3.559131	0.08	3.343568
16028	3.138889	0.001628	3.135698	3.142081	2.718648	3.559131	0.08	2.734368
53946	3.138889	0.001628	3.135698	3.142081	2.718648	3.559131	0.08	3.343568

```
[ ] 
# --- Gráfico ---
fig, ax = plt.subplots(figsize=(7, 4.5))

# Datos reales (train)
ax.scatter(pred_train["x"], pred_train["y"], color="gray", alpha=0.3,
label="Train")

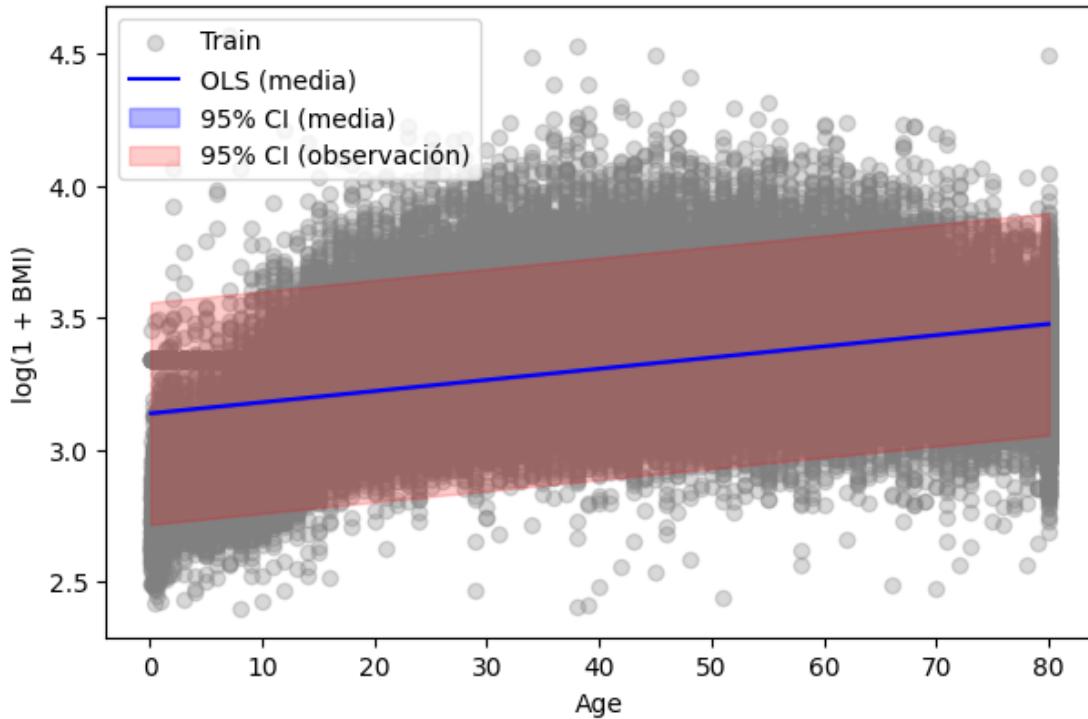
# Recta ajustada (media)
ax.plot(pred_train["x"], pred_train["mean"], color="blue", label="OLS
(media)")

# IC de la media (estrecho)
ax.fill_between(pred_train["x"],
                pred_train["mean_ci_lower"],
                pred_train["mean_ci_upper"],
                color="blue", alpha=0.3, label="95% CI (media)")

# IC de observación (más ancho)
ax.fill_between(pred_train["x"],
                pred_train["obs_ci_lower"],
                pred_train["obs_ci_upper"],
                color="red", alpha=0.2, label="95% CI (observación)")

ax.set_title("Modelo ajustado en TRAIN: recta, IC de media y de observación")
ax.set_xlabel("Age")
ax.set_ylabel("log(1 + BMI)")
ax.legend()
plt.show()
```

### Modelo ajustado en TRAIN: recta, IC de media y de observación



### Análisis

- La recta de regresión confirma que existe una relación positiva entre edad y BMI transformado, pero muy débil.
- La banda azul angosta (IC del 95% para la media), muestra que el modelo estima bien la media poblacional.
- La banda roja amplia (IC del 95% para observaciones individuales), evidencia que las predicciones individuales son muy imprecisas, reflejando que el BMI depende de muchos otros factores además de la edad.

### 4. Evaluación del Modelo con datos de prueba (el 20% restante)

```
[ ] 
# --- Predicciones en TEST con intervalos ---
X_test_const = sm.add_constant(X_test, prepend=True)
pred_test = modelo_bmi.get_prediction(X_test_const).summary_frame(alpha=0.05)

# Añadimos variables al DataFrame
```

```

pred_test["x"] = X_test.values
pred_test["y_real"] = y_test.values

# RMSE en test
y_pred_test = pred_test["mean"].values
rmse = root_mean_squared_error(y_true=y_test, y_pred=y_pred_test)
print(f"RMSE en test: {rmse:.4f}")

# Ordenamos por x para que la curva se vea bien
pred_test = pred_test.sort_values("x")

# --- Gráfico ---
fig, ax = plt.subplots(figsize=(7, 4), dpi=120)

# Puntos reales
ax.scatter(pred_test["x"], pred_test["y_real"],
           color='red', alpha=0.4, s=14, label="Reales (test)")

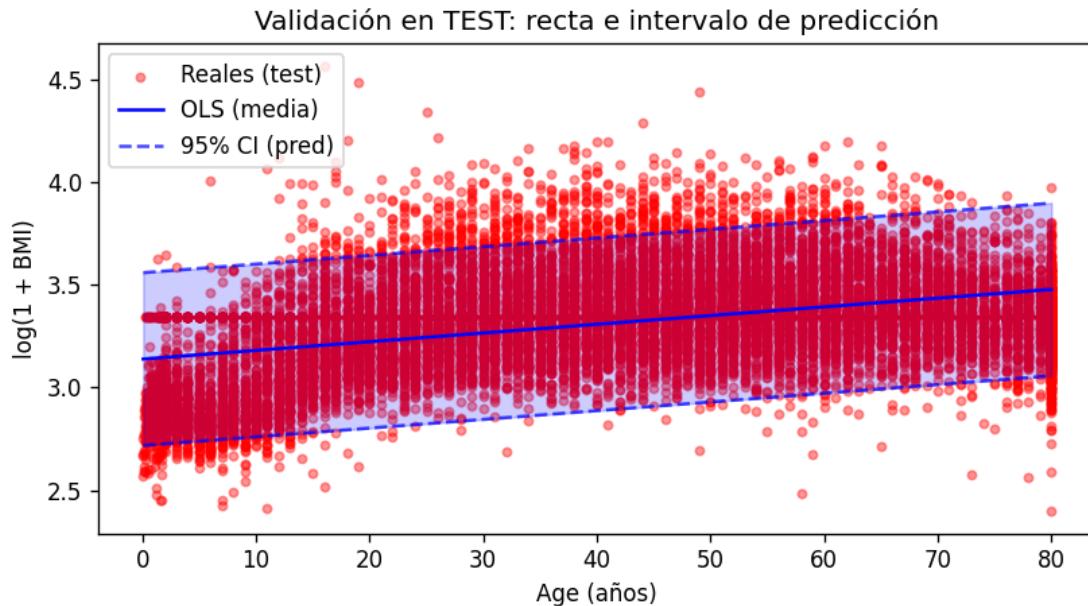
# Línea de predicción
ax.plot(pred_test["x"], pred_test["mean"],
         color="blue", label="OLS (media)")

# Banda de intervalo de predicción (observaciones)
ax.plot(pred_test["x"], pred_test["obs_ci_lower"],
        linestyle="--",
        color="blue", alpha=0.7, label="95% CI (pred)")
ax.plot(pred_test["x"], pred_test["obs_ci_upper"],
        linestyle="--", color="blue", alpha=0.7)
ax.fill_between(pred_test["x"], pred_test["obs_ci_lower"], pred_test["obs_ci_upper"],
                color="blue", alpha=0.2)

ax.set_xlabel("Age (años)")
ax.set_ylabel("log(1 + BMI)")
ax.set_title("Validación en TEST: recta e intervalo de predicción")
ax.legend()
plt.tight_layout()
plt.show()

```

RMSE en test: 0.2150



### Analisis:

- ◊ 1. Gráfico de validación

En la figura se muestran:

- **Puntos rojos (Reales – Test):**  
Valores reales de  $\log(1 + \text{BMI})$  en el conjunto de prueba.  
Se aprecia una gran dispersión: para cualquier edad, los valores de  $\text{bmi\_log}$  varían de forma amplia.
- **Línea azul (OLS – media):**  
Recta estimada por el modelo lineal.  
Confirma la tendencia positiva ya observada en entrenamiento: a mayor edad, el valor esperado de  $\log(1 + \text{BMI})$  tiende a crecer ligeramente.
- **Banda azul (95% CI – predicción):**  
Intervalo de **predicción** para nuevas observaciones individuales.  
Es bastante amplio, lo que refleja la alta incertidumbre al predecir valores individuales.  
Mientras que la media condicional se estima con precisión, los datos reales están muy dispersos alrededor de la recta.

- ◊ 2. Error de validación (RMSE en test)

El modelo presenta un **RMSE ≈ 0.2150** en el conjunto de prueba.

- Como la variable respuesta está en la escala transformada  $\log(1 + \boxed{\text{BMI}})$ , este valor indica que el error típico de predicción es de  **$\approx 0.21$**  **unidades en esa escala logarítmica.**
  - Convertido a la escala original ( $\boxed{\text{BMI}}$ ), este error equivale aproximadamente a una desviación del **6–7% respecto al BMI real.**
- 

- ◊ 3. Conclusiones sobre el desempeño en TEST
1. El modelo generaliza de forma consistente: el comportamiento en **TEST** es muy similar al observado en **TRAIN**.
  2. La recta ajustada capta correctamente la **tendencia promedio** entre edad y BMI.
  3. La **amplia dispersión de puntos** y el **intervalo de predicción ancho** indican que la edad explica solo una pequeña parte de la variabilidad del BMI.
  4. El RMSE confirma que las predicciones son relativamente precisas en promedio, pero **poco útiles a nivel individual**: el BMI depende fuertemente de otros factores no incluidos en el modelo.
- 

**Resumen:** El modelo captura la tendencia poblacional (edad  $\rightarrow$  BMI), pero tiene **baja capacidad predictiva individual**. Para mejorar el ajuste, se requiere incorporar más variables relevantes en el análisis.

---

## Interpretación General del Modelo

---

- ◊ 1. Ecuación estimada El modelo lineal obtenido es:

$$\text{bmi\_log}^{\wedge}=3.1386+0.0042 \cdot \text{age}$$

Donde:

- $\boxed{\text{bmi\_log}} = \log(1 + \boxed{\text{BMI}})$
- $\boxed{\text{age}}$  = edad de la persona en años.

### Interpretación de los coeficientes:

- **Intercepto (3.1386):** Valor esperado de  $\log(1 + \boxed{\text{BMI}})$  cuando la edad es 0.  
En términos prácticos, equivale a un **BMI  $\approx 22.9$**  (valor típico de inicio en la infancia).

- **Pendiente (0.0042):** Por cada año adicional de edad, el valor de  $\log(1 + \text{BMI})$  aumenta en promedio **0.0042 unidades**.

Al llevarlo a la escala original: implica que el **BMI se incrementa en torno a un 0.42% anual**.

Ejemplo:

- **Persona de 20 años:**

$$\begin{aligned}\text{bmi\_log}^{\wedge} &= 3.1386 + 0.0042 \cdot 20 = 3.2226 \\ \text{BMI}^{\wedge} &= e^{3.2226} - 1 \approx 24.99\end{aligned}$$

- **Persona de 40 años:**

$$\begin{aligned}\text{bmi\_log}^{\wedge} &= 3.1386 + 0.0042 \cdot 40 = 3.3066 \\ \text{BMI}^{\wedge} &= e^{3.3066} - 1 \approx 26.3\end{aligned}$$

**Interpretación:** pasar de 20 a 40 años implica un aumento esperado de  $\approx 1.3$  puntos en BMI, lo que equivale a un incremento acumulado de aproximadamente **5.2%**.

◊ 2. Ajuste del modelo en TRAIN

- El **R<sup>2</sup> ≈ 0.165**, lo que significa que la **edad explica solo un 16.5% de la variabilidad del BMI**.
- Los coeficientes son **altamente significativos (p-value < 0.05)**, confirmando que la relación edad ↔ BMI existe y no es aleatoria.
- El gráfico en TRAIN muestra una **recta clara con intervalos de confianza estrechos para la media**, pero los datos reales están muy dispersos, evidenciando que **otros factores influyen fuertemente en el BMI**.

◊ 3. Validación en TEST

- El modelo predice en el conjunto de prueba con un **RMSE ≈ 0.215** en la escala  $\log(1 + \text{BMI})$ .
- Convertido al BMI real, el error equivale aproximadamente a un **6–7% respecto al valor real de BMI**.
- La **banda de predicción** es amplia, lo que refleja **alta incertidumbre en predicciones individuales**.
- La recta ajustada sigue la misma tendencia observada en TRAIN, indicando que el modelo **generaliza correctamente**, aunque con baja capacidad explicativa.

◊ 4. Evaluación global del modelo

- **Precisión estadística:** El modelo es estadísticamente significativo, con coeficientes robustos y p-values muy bajos.
  - **Poder explicativo limitado:** El  $R^2$  bajo (16.5%) indica que la **edad por sí sola es un predictor débil del BMI**.
  - **Intervalos de predicción amplios:** Aunque la tendencia es clara, las predicciones para individuos concretos presentan mucha incertidumbre.
  - **Uso recomendado:** El modelo es útil para **captar la tendencia promedio poblacional** (cómo varía el BMI con la edad), pero **no es adecuado para predecir el BMI de una persona en particular**.
- 

## Conclusión

El modelo lineal **es válido pero poco preciso a nivel individual**:

- Confirma que la edad se relaciona con el BMI de manera positiva y significativa.
  - Sin embargo, **otros factores (dieta, genética, actividad física, etc.)** no considerados explican la mayor parte de la variabilidad.
  - Es un buen **primer modelo exploratorio**, pero debe complementarse con más variables para lograr predicciones realmente útiles y robustas.
-