EstadisticaDescriptiva-Brian2

September 19, 2025

0.1 Brian Roberto Gómez Martínez A00841404

1 Análisis Descriptivo del Dataset de Diabetes

1.1 1. Carga de Datos

[2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

# Cargar los datos desde el archivo CSV
diabetes = pd.read_csv('diabetes.csv')
```

1.2 2. Verificación de Datos

En esta sección revisamos las dimensiones del dataset, las variables que contiene y sus tipos de datos.

[3]:
```
# Verificar la cantidad de datos (filas, columnas)
print("Dimensiones del dataset (filas, columnas):")
diabetes.shape
```

```
Dimensiones del dataset (filas, columnas):
```

[3]:
```
(768, 9)
```

[4]:
```
# Ver las variables que contiene cada vector de datos
print("Nombres de las variables:")
diabetes.columns
```

```
Nombres de las variables:
```

[4]:
```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

[5]:
```
# Identificar el tipo de variables y si hay valores nulos
diabetes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

1.3 3. Análisis de Variables

Utilizamos describe() para obtener un resumen estadístico que nos muestra qué representa cada variable y en qué rangos se encuentran sus valores.

[6]:
```
# Analizar rangos (mínimo y máximo) y otras estadísticas clave
diabetes.describe()
```

[6]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Descripción y Rango de las Variables:

Variable	Descripción	Rango (Mín - Máx)
Pregnancies	Número de embarazos	0 - 17
Glucose	Concentración de glucosa en plasma a 2 horas	0 - 199
BloodPressure	Presión arterial diastólica (mm Hg)	0 - 122
SkinThickness	Grosor del pliegue cutáneo del tríceps (mm)	0 - 99
Insulin	Insulina sérica de 2 horas (mu U/ml)	0 - 846
BMI	Índice de Masa Corporal (kg/m²)	0 - 67.1
DiabetesPedigreeFunction	Función de pedigrí de diabetes (influencia genética)	0.078 - 2.42
Age	Edad (años)	21 - 81
Outcome	Diagnóstico (0 = No, 1 = Sí)	0 - 1

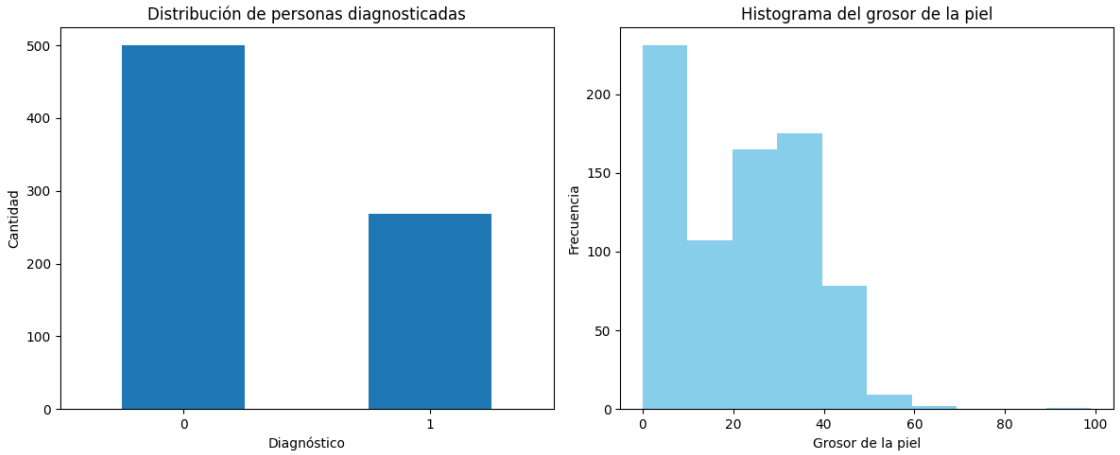
1.4 4. Visualización y análisis de Datos

En esta sección se analizan las variables **Outcome** y **SkinThickness**.

[7]:
```
fig, axes = plt.subplots(1, 2, figsize=(12, 5))
diagnosticados = diabetes['Outcome'].value_counts()
diagnosticados.plot(kind='bar', ax=axes[0])
axes[0].set_title('Distribución de personas diagnosticadas')
axes[0].set_xlabel('Diagnóstico')
axes[0].set_ylabel('Cantidad')
axes[0].tick_params(axis='x', rotation=0)

axes[1].hist(diabetes['SkinThickness'], bins=10, color='skyblue')
axes[1].set_title('Histograma del grosor de la piel')
axes[1].set_xlabel('Grosor de la piel')
axes[1].set_ylabel('Frecuencia')

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



El gráfico de barras de la izquierda muestra la distribución de la variable **Outcome**. El número de pacientes que no tienen diabetes (0) es considerablemente mayor, aproximadamente el doble, que el número de pacientes que sí la tienen (1).

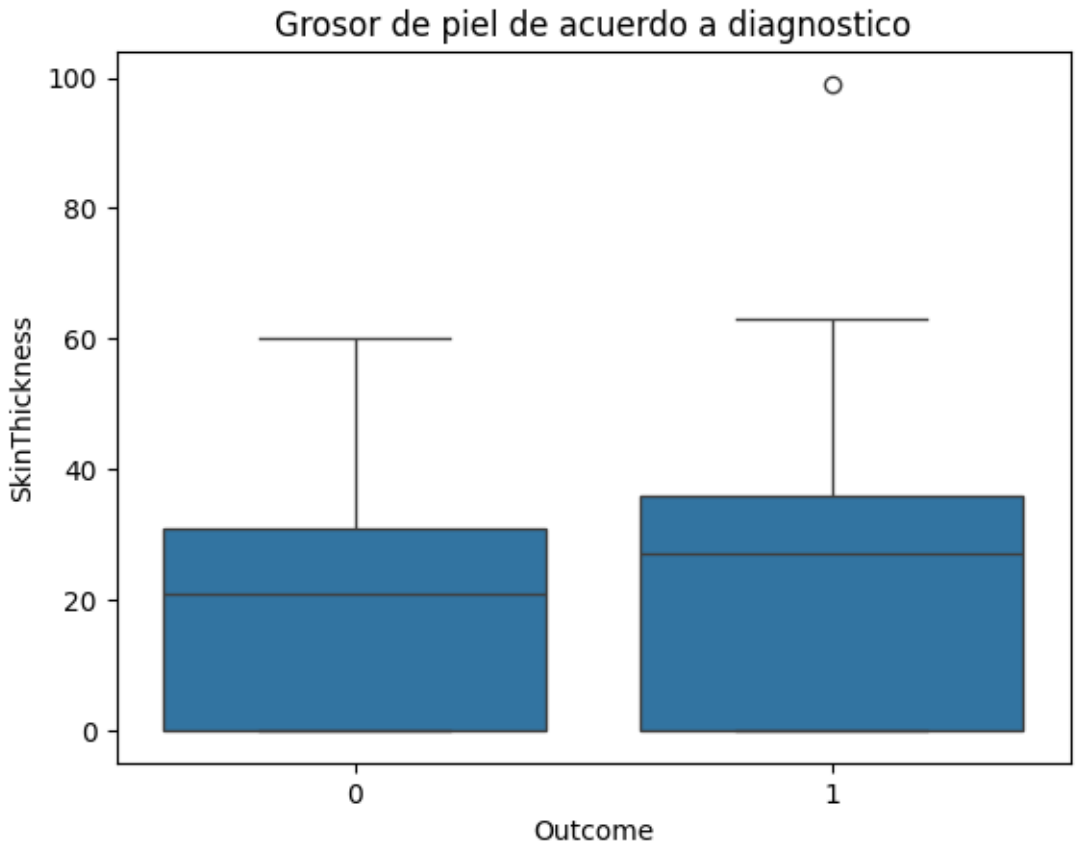
El histograma de la derecha, que representa el **SkinThickness**, revela que hay una altísima frecuencia de valores en cero y podríamos inferir que estos son datos faltantes o anómalos. Para el resto de los puntos, la mayoría de los valores de grosor de la piel se concentran en el rango de 20 a 40 mm, con una distribución que disminuye rápidamente para valores más altos.

[8]:
```
q1, mediana, q3 = np.percentile(diabetes['SkinThickness'], [25, 50, 75])
print("el 25% de las personas tienen un grosor de piel de ",q1)
print("el 25% de las personas tienen grosor de piel de ",q3)

sns.boxplot(diabetes, x="Outcome", y="SkinThickness")
plt.title("Grosor de piel de acuerdo a diagnostico")
```

```
el 25% de las personas tienen un grosor de piel de  0.0
el 25% de las personas tienen grosor de piel de  32.0
```

[8]:
```
Text(0.5, 1.0, 'Grosor de piel de acuerdo a diagnostico')
```

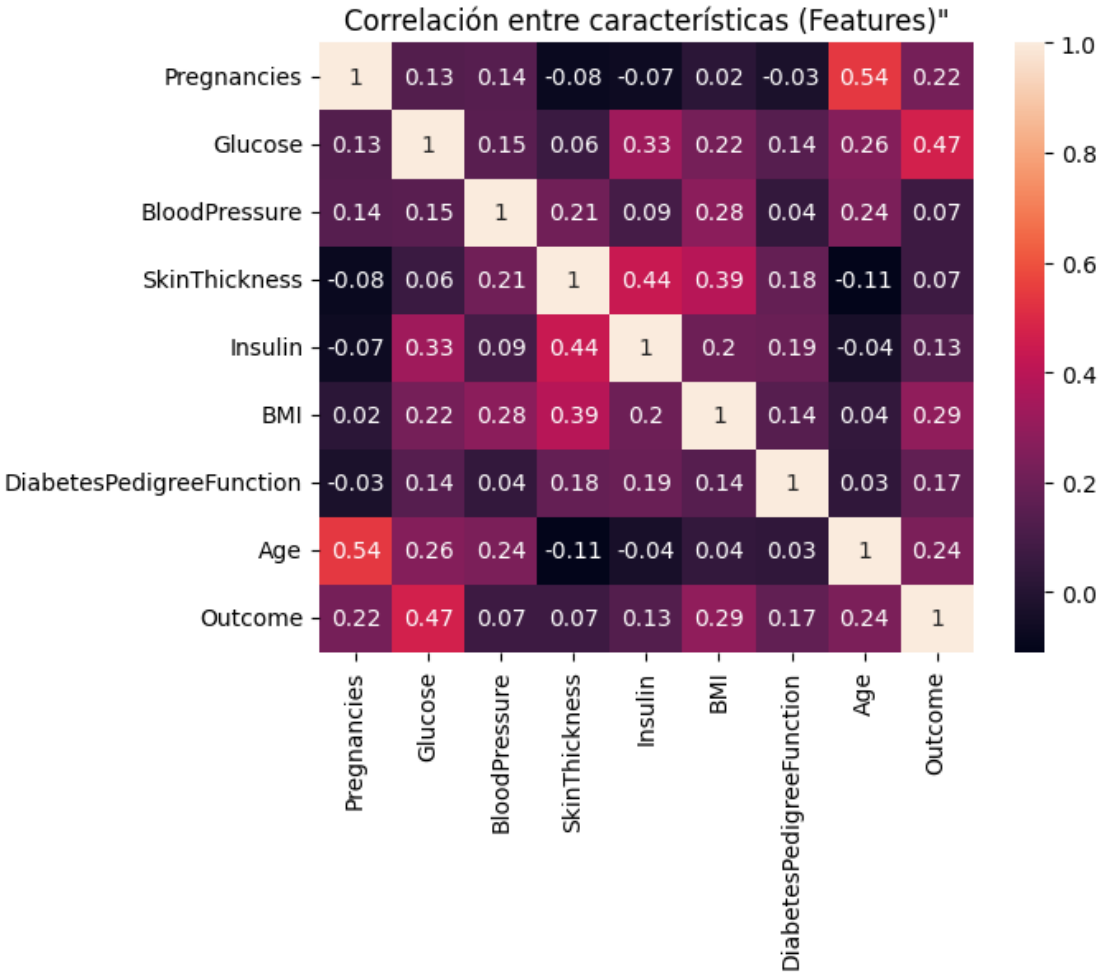


En este boxplot realcionando a las dos variables podemos observar que la media de perosonas que

tienen un diagnostico de diabetes tienen la piel más gruesa con respecto a la media de personas que no tienen un diagnóstico

[46]:
```
#seleccionar varaiables numericas
variables_numericas = diabetes.select_dtypes(include='number')
matriz_correlacion = variables_numericas.corr().round(2)
matriz_correlacion
sns.heatmap(matriz_correlacion, annot=True)
plt.title('Correlación entre características (Features)"')
print("Podemos observar en el heatmap de todas las variables que las dos␣
 ↪variables estudiadas no tienen una correlación significante(0.07 equivalente␣
 ↪al 7%) por tanto podemos concluir que no hay una correlación directa entre␣
 ↪estas dos variables")
```

```
Podemos observar que las dos variables estudiadas no tienen una correlación
significante(0.07 equivalente al 7%) por tanto podemos concluir que no hay una
correlación directa entre estas dos variables
```



-

1.4.1 ¿Hay alguna variable que no aporta información? Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

Creo que la variable de pregnancies no tiene suficiente correlación con las otras variables a excepción de edad, sin embargo esto no es relevante para la invetigación sobre la diabetes

-

1.4.2 Si comparas el rango de las variables (min-max), ¿todas están en rangos similares? Describe sus rangos.

No están en rangos similares y esto es por la diferencia de unidades utilizada en cada medición, que esten en rangos similares realmente podría no significar nada y por tanto no es relevante.

-

1.4.3 ¿Existen variables que tengan datos atípicos? Describe cuáles si o no.

Outcome si tiene un valor atipico en el bloxplot, recordando que un valor atipico está a más allá de dos desviaciónes estandar

-

1.4.4 ¿Existe correlación alta entre variables? Describe algunas, indicando si es correlación positiva o negativa.

Se observa una correlación alta (>50%) entre la edad y el numero de embarazos, sin embargo esto no es necesariamente por la diabetes, esta correlación es positiva y por tanto significa que ambas variables crecen y decrecen proporcionalemnte. Tambien existe una correlación considerable (>40%) entre el nivel de insulina y el grosor de la piel, lo que podría considerarse para realizar un analisis más a fondo, esta correlación también es positiva y por tanto significa que ambas variables crecen y decrecen proporcionalmente.

1.5 5. Conclusiones

Basándonos en la media, mediana y desviación estándar, y el analisis gráfico hecho anteriormente podemos decir que:

El análisis de las variables Outcome y SkinThickness revela una conexión directa y significativa entre ambas. La variable Outcome muestra una mayor proporción de individuos sin diabetes que con ella, estableciendo la distribución base de los pacientes. Por su parte, SkinThickness presenta un rango de valores considerable, pero su comportamiento es marcadamente distinto al dividirse por el resultado del diagnóstico. Se observa de manera consistente que los pacientes con un diagnóstico positivo de diabetes tienden a presentar una mediana de grosor de piel notablemente más alta. Esta diferencia sugiere que un mayor grosor en el pliegue cutáneo está directamente asociado con la presencia de diabetes en la población estudiada.