Sylvie Aylin Zabetoglu González
A01234692
19/09/2025

Comprensión de los Datos

In [2]:
```
#importa librerías
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import sklearn
```

Descripción de Variables

Pregnancies - Número de embarazos: Cuantitativa discreta

Glucose - Nivel de glucosa en plasma a 2 horas: Cuantitativa continua

BloodPressure - Presión arterial (mm Hg): Cuantitativa continua

SkinThickness - Grosor del pliegue cutáneo (mm): Cuantitativa continua

Insulin - Insulina sérica a 2 horas (mu U/ml): Cuantitativa continua

BMI - Índice de masa corporal (kg/m^2): Cuantitativa continua

DiabetesPedigreeFunction - Función de predisposición genética a diabetes: Cuantitativa continua

Age - Edad (años): Cuantitativa discreta

Outcome - Diagnóstico de diabetes (0 = No, 1 = Sí): Categórica Nominal

In [4]:
```
#Lee archivo csv
diabetes = pd.read_csv("diabetes.csv")
```

In [76]:
```
diabetes.head()
```

Out[76]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

In [75]:
```
#Revisa los últimos 5 renglones del dataset usando la función tail()
diabetes.tail()
```

Out[75]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

In [6]:
```
# Número de filas y columnas
print("Dimensiones:", diabetes.shape)

# Información de columnas y tipos de datos
diabetes.info()

# Número de valores únicos
diabetes.nunique()
```

```
Dimensiones: (768, 9)
<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
```

Out[6]:
```
Pregnancies                 17
Glucose                    136
BloodPressure               47
SkinThickness               51
Insulin                    186
BMI                        248
DiabetesPedigreeFunction   517
Age                         52
Outcome                      2
dtype: int64
```

Exploración de Datos

In [8]:
```
#utiliza la función describe() para obtener estadística básica. se puede incluir -0
diabetes.describe()
```

Out[8]:

	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

In [9]:
```
for col in diabetes.columns:
    print(f"{col}:")
    print(f"  Mínimo: {diabetes[col].min()}")
    print(f"  Máximo: {diabetes[col].max()}")
    print(f"  Media: {diabetes[col].mean():.5f}")
    print(f"  Mediana: {diabetes[col].median():.2f}")
    print(f"  Desviación estándar: {diabetes[col].std():.2f}")
    print("-"*30)
```

```
Pregnancies:
  Mínimo: 0
  Máximo: 17
  Media: 3.84505
  Mediana: 3.00
  Desviación estándar: 3.37
------------------------------
Glucose:
  Mínimo: 0
  Máximo: 199
  Media: 120.89453
  Mediana: 117.00
  Desviación estándar: 31.97
------------------------------
BloodPressure:
  Mínimo: 0
  Máximo: 122
  Media: 69.10547
  Mediana: 72.00
  Desviación estándar: 19.36
------------------------------
SkinThickness:
  Mínimo: 0
  Máximo: 99
  Media: 20.53646
  Mediana: 23.00
  Desviación estándar: 15.95
------------------------------
Insulin:
  Mínimo: 0
  Máximo: 846
  Media: 79.79948
  Mediana: 30.50
  Desviación estándar: 115.24
------------------------------
BMI:
  Mínimo: 0.0
  Máximo: 67.1
  Media: 31.99258
  Mediana: 32.00
  Desviación estándar: 7.88
------------------------------
DiabetesPedigreeFunction:
  Mínimo: 0.078
  Máximo: 2.42
  Media: 0.47188
  Mediana: 0.37
  Desviación estándar: 0.33
------------------------------
Age:
  Mínimo: 21
  Máximo: 81
  Media: 33.24089
  Mediana: 29.00
  Desviación estándar: 11.76
------------------------------
Outcome:
  Mínimo: 0
  Máximo: 1
  Media: 0.34896
  Mediana: 0.00
  Desviación estándar: 0.48
------------------------------
```

In [10]:
```python
# Conteo y porcentaje de cada clase en Outcome
conteo = diabetes['Outcome'].value_counts()
porcentaje = diabetes['Outcome'].value_counts(normalize=True) * 100

# Mostrar resultados
print("Distribución de Outcome (diagnóstico de diabetes tipo 2):\n")
for valor in conteo.index:
    print(f"{valor} -> {conteo[valor]} casos ({porcentaje[valor]:.2f}%)")
```

```
Distribución de Outcome (diagnóstico de diabetes tipo 2):

0 -> 500 casos (65.10%)
1 -> 268 casos (34.90%)
```

Conclusiones generales

El conjunto de datos contiene información de 768 pacientes con diversas variables clínicas y demográficas que pueden influir en el desarrollo de Diabetes. Se observan varios puntos importantes:

1. **Composición general de los datos**: Se identifican nueve variables: número de embarazos (Pregnancies), nivel de glucosa (Glucose), presión arterial diastólica (BloodPressure), grosor del pliegue cutáneo (SkinThickness), nivel de insulina (Insulin), índice de masa corporal

(BMI), historial genético de diabetes (DiabetesPedigreeFunction), edad (Age) y resultado del diagnóstico (Outcome). La base contiene únicamente valores numéricos (enteros o flotantes), y un total de 768 registros sin valores nulos explícitos, aunque varias variables contienen ceros que representan valores faltantes.

2. **Presencia de valores atípicos y posibles datos faltantes**: Variables como Glucose, BloodPressure, SkinThickness, Insulin y BMI presentan valores de 0, lo cual no es fisiológicamente posible y probablemente corresponde a datos faltantes que deben ser tratados antes de cualquier análisis avanzado. Insulin destaca por su alta dispersión y un valor máximo extremadamente alto (846), lo que sugiere la existencia de outliers que podrían influir de forma importante en los resultados.

3. **Distribución de las variables clínicas**: Los niveles medios de Glucose (120.9 mg/dl) y BMI (31.99) indican que la mayoría de las personas de la muestra presentan sobrepeso y niveles de glucosa moderadamente altos, ambos factores de riesgo para el desarrollo de Diabetes. La presión arterial media (69.1 mmHg) y la edad promedio (33 años) reflejan que se trata de una población principalmente joven y con presión en rangos normales, aunque con algunos casos de valores elevados. El DiabetesPedigreeFunction tiene valores bajos en la mayoría de los casos, lo que indica que la predisposición genética es leve para la mayoría de los pacientes, aunque existen algunos casos con predisposición alta.

4. **Distribución del diagnóstico de diabetes**: Aproximadamente un 34% de las personas en el conjunto de datos tienen un diagnóstico positivo (Outcome = 1), lo que representa una proporción significativa de casos de diabetes en la muestra

5. **Relación entre media, mediana y dispersión**: En la mayoría de las variables la media y la mediana son cercanas, lo que indica distribuciones relativamente simétricas. Sin embargo, en Insulin y SkinThickness la gran diferencia entre media y mediana, así como su alta desviación estándar, sugiere distribuciones sesgadas y la existencia de valores muy alejados del promedio.

Variables Cuantitativas

Medidas de tendencia central

In [11]:
```python
# Lista de variables cuantitativas
variables_cuantitativas = ['Pregnancies', 'Glucose', 'BloodPressure',
                           'SkinThickness', 'Insulin', 'BMI',
                           'DiabetesPedigreeFunction', 'Age']

# Recorrer cada variable y calcular media, mediana y moda
for var in variables_cuantitativas:
    mean_val = diabetes[var].mean()
    median_val = diabetes[var].median()
    mode_val = diabetes[var].mode()[0]  # Tomamos la primera moda si hay varias
    print(f"{var}:")
    print(f"  Media: {mean_val:.2f}")
    print(f"  Mediana: {median_val:.2f}")
    print(f"  Moda: {mode_val}")
    print("-"*30)
```

```
Pregnancies:
  Media: 3.85
  Mediana: 3.00
  Moda: 1
------------------------------
Glucose:
  Media: 120.89
  Mediana: 117.00
  Moda: 99
------------------------------
BloodPressure:
  Media: 69.11
  Mediana: 72.00
  Moda: 70
------------------------------
SkinThickness:
  Media: 20.54
  Mediana: 23.00
  Moda: 0
------------------------------
Insulin:
  Media: 79.80
  Mediana: 30.50
  Moda: 0
------------------------------
BMI:
  Media: 31.99
  Mediana: 32.00
  Moda: 32.0
------------------------------
DiabetesPedigreeFunction:
  Media: 0.47
  Mediana: 0.37
  Moda: 0.254
------------------------------
Age:
  Media: 33.24
  Mediana: 29.00
  Moda: 22
------------------------------
```

In [72]:
```
# Contar ceros en todas las columnas numéricas
ceros_por_columna = (diabetes == 0).sum()
print(ceros_por_columna)
```

```
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

Variables Categóricas

In [12]:
```
#Para conteo  de cada valor en una columna, en orden descendente usar función value_counts():
# nombreDataframe.columna.value_counts()
# nombreDataframe['columna'].value_counts()
print("Conteo de pacientes según diagnóstico de diabetes:")
print(diabetes['Outcome'].value_counts())
```

```
Conteo de pacientes según diagnóstico de diabetes:
Outcome
0    500
1    268
Name: count, dtype: int64
```

Consulta

In [13]:
```
# df.iloc[i]: Accede a la fila en la posición i.
# Acceder a la primera fila
print("Primera fila del dataset:")
diabetes.iloc[0]
```

```
Primera fila del dataset:
```

Out[13]:
```
Pregnancies                 6.000
Glucose                   148.000
BloodPressure              72.000
SkinThickness              35.000
Insulin                     0.000
BMI                        33.600
DiabetesPedigreeFunction    0.627
Age                        50.000
Outcome                     1.000
Name: 0, dtype: float64
```

In [14]:
```
# Acceder a las dos primeras filas
diabetes.iloc[:2]
```

Out[14]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0

In [84]:
```
#Selección de filas [indicar dataframe[columna] operador valor]
print("\nPacientes con diagnóstico de diabetes:")
pacientes=diabetes[diabetes['Outcome']==1]
pacientes.head()
```

```
Pacientes con diagnóstico de diabetes:
```

Out[84]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
2	8	183	64	0	0	23.3	0.672	32	1
4	0	137	40	35	168	43.1	2.288	33	1
6	3	78	50	32	88	31.0	0.248	26	1
8	2	197	70	45	543	30.5	0.158	53	1

In [77]:
```
# Filtrar pacientes con glucosa alta (Glucose > 140)
glucosa_alta = diabetes[diabetes['Glucose'] > 140]
print("\nPacientes con glucosa > 140 mg/dl:")
glucosa_alta.head()
```

```
Pacientes con glucosa > 140 mg/dl:
```

Out[77]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
2	8	183	64	0	0	23.3	0.672	32	1
8	2	197	70	45	543	30.5	0.158	53	1
11	10	168	74	0	0	38.0	0.537	34	1
13	1	189	60	23	846	30.1	0.398	59	1

In [80]:
```
# Ordenar los pacientes con diabetes por nivel de insulina de mayor a menor
pacientes=diabetes[diabetes['Outcome']==1]
pacientes = pacientes.sort_values(by='Insulin', ascending=False)
print("\nPacientes con diabetes ordenados por Insulina (mayor a menor):")
pacientes.head()
```

```
Pacientes con diabetes ordenados por Insulina (mayor a menor):
```

Out[80]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
13	1	189	60	23	846	30.1	0.398	59	1
584	8	124	76	24	600	28.7	0.687	52	1
409	1	172	68	49	579	42.4	0.702	28	1
8	2	197	70	45	543	30.5	0.158	53	1
655	2	155	52	27	540	38.7	0.240	25	1

In [82]:
```
# Filtrar pacientes con BMI en rango de obesidad  y Glucose alta
obesidad_glucosa_alta = diabetes[(diabetes['BMI'] >= 30) & (diabetes['Glucose'] > 140)]
print("\nPacientes obesos con glucosa alta:")
obesidad_glucosa_alta.head()
```

```
Pacientes obesos con glucosa alta:
```

Out[82]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
8	2	197	70	45	543	30.5	0.158	53	1
11	10	168	74	0	0	38.0	0.537	34	1
13	1	189	60	23	846	30.1	0.398	59	1
22	7	196	90	0	0	39.8	0.451	41	1

In [83]:
```
# Filtrar pacientes con SkinThickness y DiabetesPedigreeFunction altas
skin_dpf_alto = diabetes[(diabetes['SkinThickness'] > 23) & (diabetes['DiabetesPedigreeFunction'] > 0.3725)]

print("\nPacientes con SkinThickness y DPF altos:")
skin_dpf_alto.head()
```

```
Pacientes con SkinThickness y DPF altos:
```

Out[83]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
4	0	137	40	35	168	43.1	2.288	33	1
16	0	118	84	47	230	45.8	0.551	31	1
19	1	115	70	30	96	34.6	0.529	32	1
20	3	126	88	41	235	39.3	0.704	27	0

Conclusiones

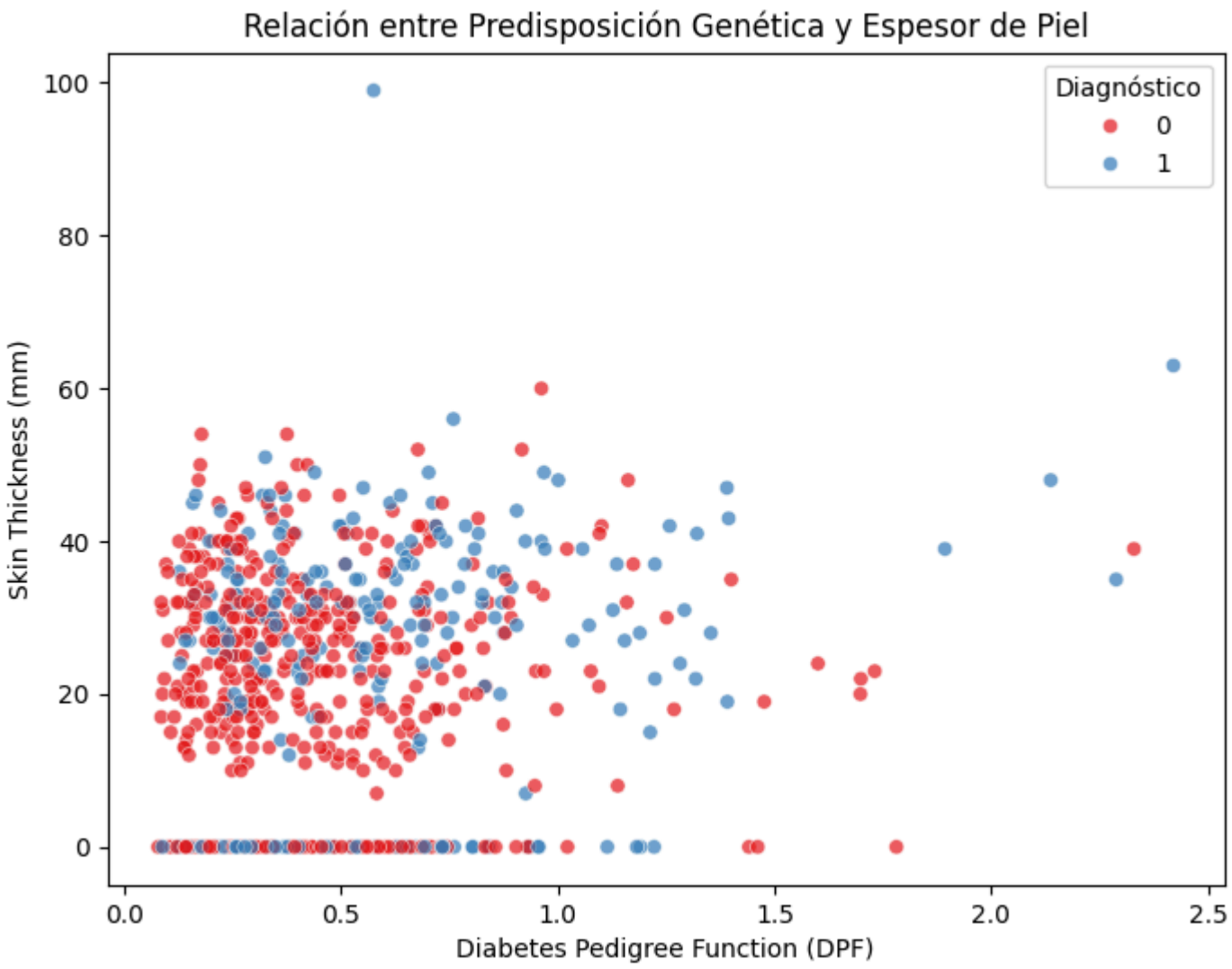
Al explorar las filas y realizar filtrados específicos en el conjunto de datos de diabetes, se obtienen varias observaciones relevantes. La primera fila del dataset corresponde a un paciente de 50 años con antecedentes de 6 embarazos, glucosa de 148 mg/dl, BMI de 33.6 y diagnóstico positivo de diabetes, lo que nos permite contextualizar los registros. Al filtrar todos los pacientes con diagnóstico positivo (Outcome = 1), se observa que estos presentan una amplia variabilidad en sus mediciones clínicas, incluyendo niveles de glucosa que van desde moderados hasta extremadamente altos. Cuando se consideran únicamente los pacientes con glucosa superior a 140 mg/dl, se identifica un grupo significativo de individuos con riesgo elevado, incluyendo algunos con valores de insulina muy altos o BMI en rango de obesidad. Y al filtrar simultáneamente pacientes con obesidad (BMI ≥ 30) y glucosa alta (> 140 mg/dl), se identifica un subgrupo de pacientes con riesgo elevado combinado de diabetes, sobrepeso y glucosa alta, lo que resalta la importancia de realizar análisis bivariados y multivariados para comprender mejor los factores de riesgo y la distribución de las variables clínicas en la población estudiada. Adicionalmente, al analizar las variables SkinThickness y DiabetesPedigreeFunction, se identifican pacientes con valores altos en ambas medidas, lo que podría indicar un riesgo genético combinado con características físicas asociadas a la diabetes.

Análisis bivariado

In [40]:
```
corr = diabetes[['DiabetesPedigreeFunction', 'SkinThickness']].corr().iloc[0,1]
print(f"Correlación entre DiabetesPedigreeFunction y SkinThickness: {corr:.2f}")
```

```
Correlación entre DiabetesPedigreeFunction y SkinThickness: 0.18
```

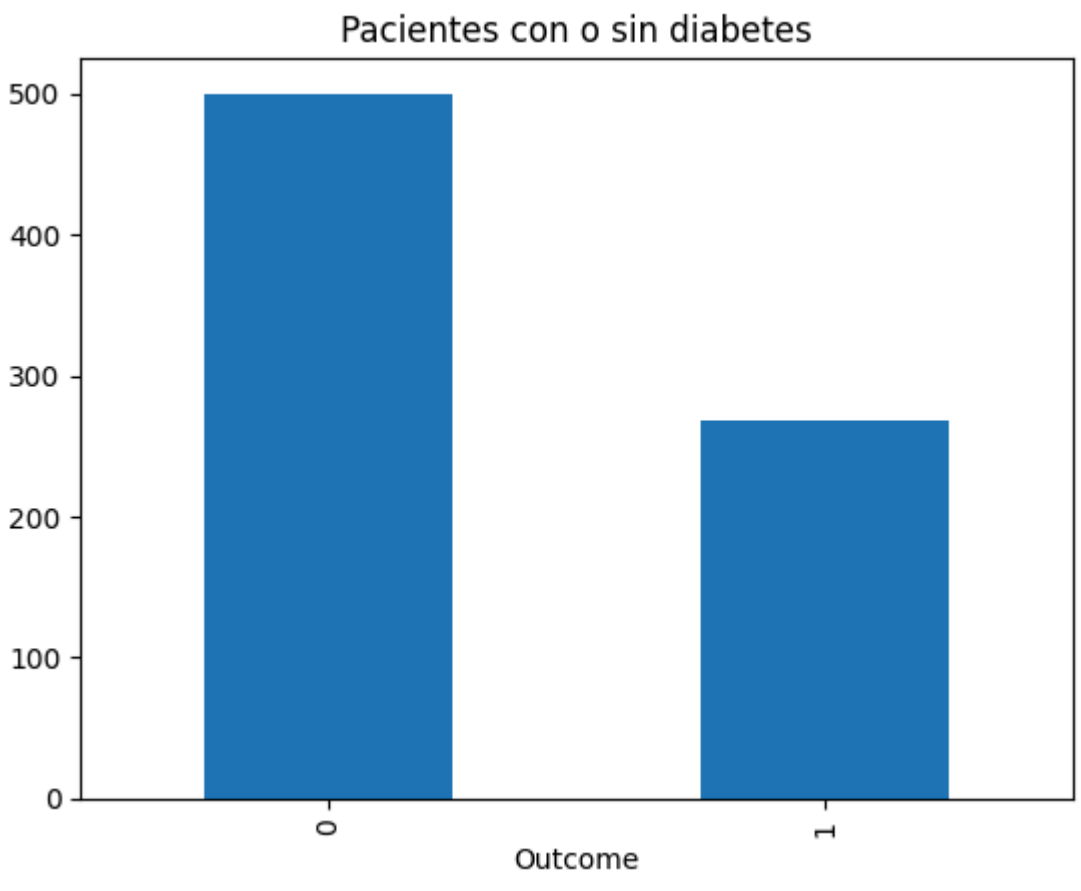
In [42]:
```
# Scatterplot
plt.figure(figsize=(8,6))
sns.scatterplot(
    data=diabetes,
    x='DiabetesPedigreeFunction',
    y='SkinThickness',
    hue='Outcome',
    palette='Set1',
    alpha=0.7
)
plt.title("Relación entre Predisposición Genética y Espesor de Piel")
plt.xlabel("Diabetes Pedigree Function (DPF)")
plt.ylabel("Skin Thickness (mm)")
plt.legend(title="Diagnóstico")
plt.show()
```



En la gráfica de dispersión se observa la relación entre la función de predisposición genética a la diabetes (DPF) y el espesor de la piel. Los puntos están diferenciados por diagnóstico (pacientes con y sin diabetes). Los datos muestran que, en promedio, los pacientes con diabetes presentan mayores valores de SkinThickness y DPF en comparación con quienes no tienen diabetes. La correlación entre ambas variables es positiva pero baja (0.18), lo que indica una ligera tendencia de que a mayor predisposición genética corresponda un mayor grosor de piel, aunque la relación no es muy fuerte. Además, se observan varios valores atípicos en ambos grupos, especialmente entre los pacientes con diabetes, reflejando mediciones excepcionales que se alejan de la mayoría de los datos. .

Visualización y Análisis de Datos

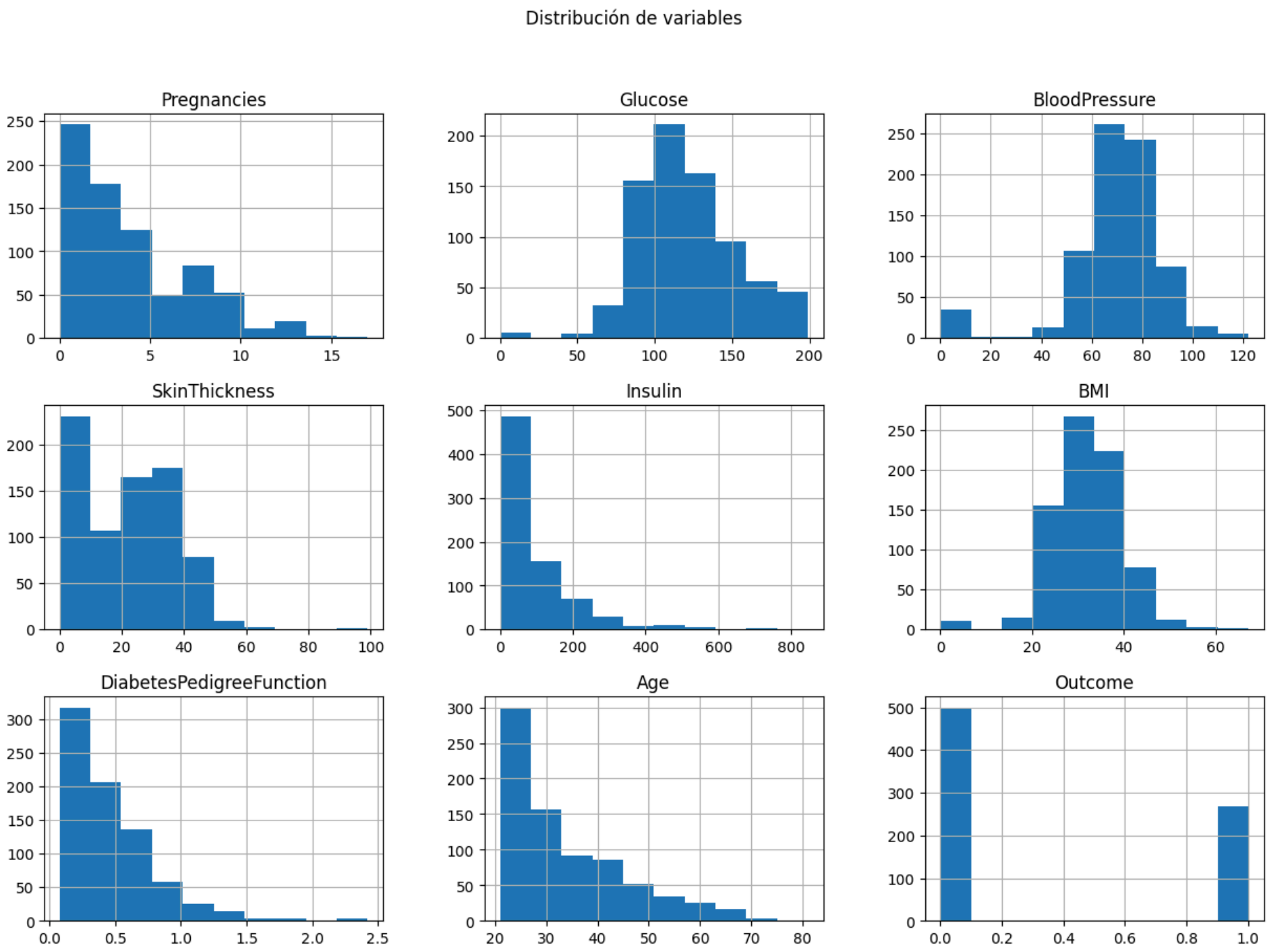
In [31]:
```
# Gráfico de barras
pacientes= diabetes['Outcome'].value_counts()
pacientes.plot(kind='bar')
plt.title('Pacientes con o sin diabetes')
plt.show()
```



Con este gráfico podemos visualizar que pacientes fueron diagnosticados con diabetes, teniendo 500 pacientes con Outcome=0 siendo que no presentan esta condición. En este caso la variable de Outcome es más fácil de visualizar en este tipo de gráfico debido a solo utilizar dos valores.

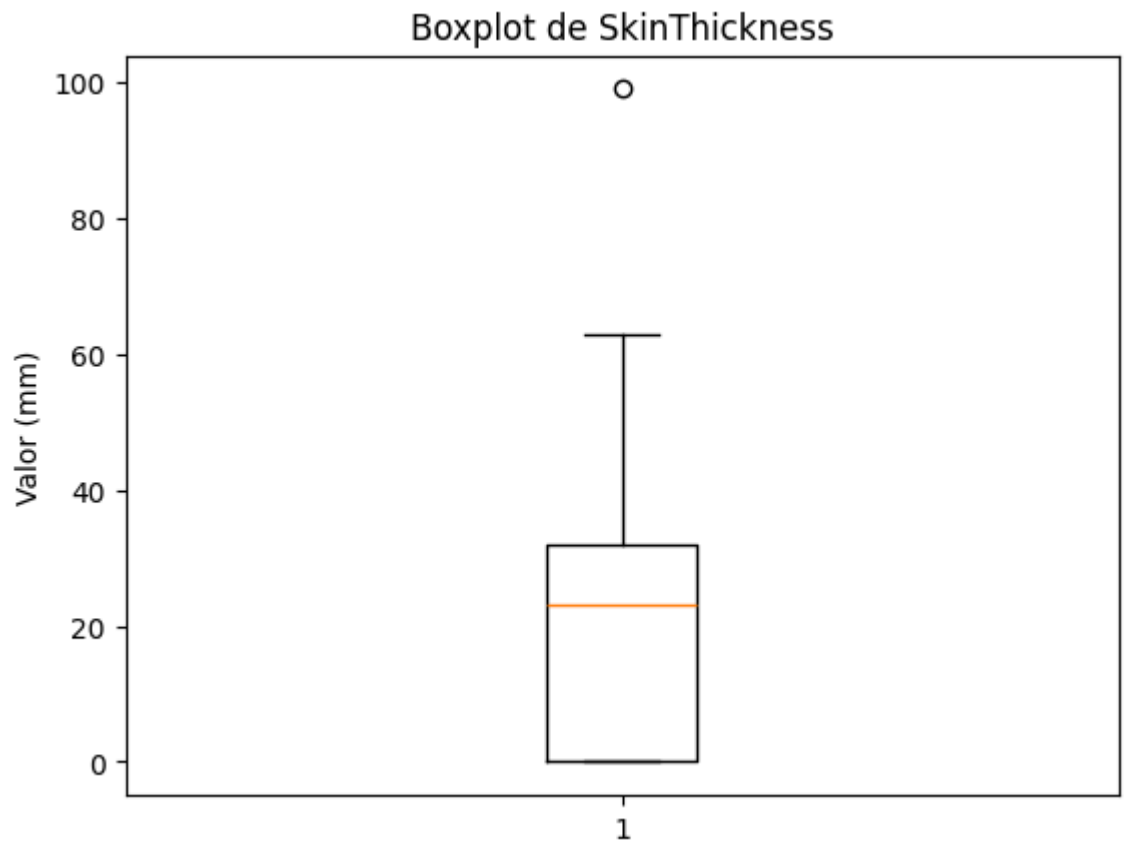
In [27]:
```
# distribución de variables en gráficos de barras
diabetes.hist(figsize=(15,10))
```

```
plt.suptitle("Distribución de variables")
plt.show()
```



Con este histograma podemos visualizar la distribución de cada variable del conjunto de datos. Se observa que varias variables presentan distribuciones asimétricas y concentraciones en ciertos rangos específicos. Al enfocarnos en las variables de este respectivo análisis, DiabetesPedigreeFunction muestra una distribución sesgada a la derecha mientras que SkinThickness tiene una distribución asimética donde los datos varían en su rango de datos.

In [52]:
```
# Gráfico boxplot
plt.boxplot(diabetes['SkinThickness'].dropna())
plt.title('Boxplot de SkinThickness')
plt.ylabel('Valor (mm)')
plt.show()
```

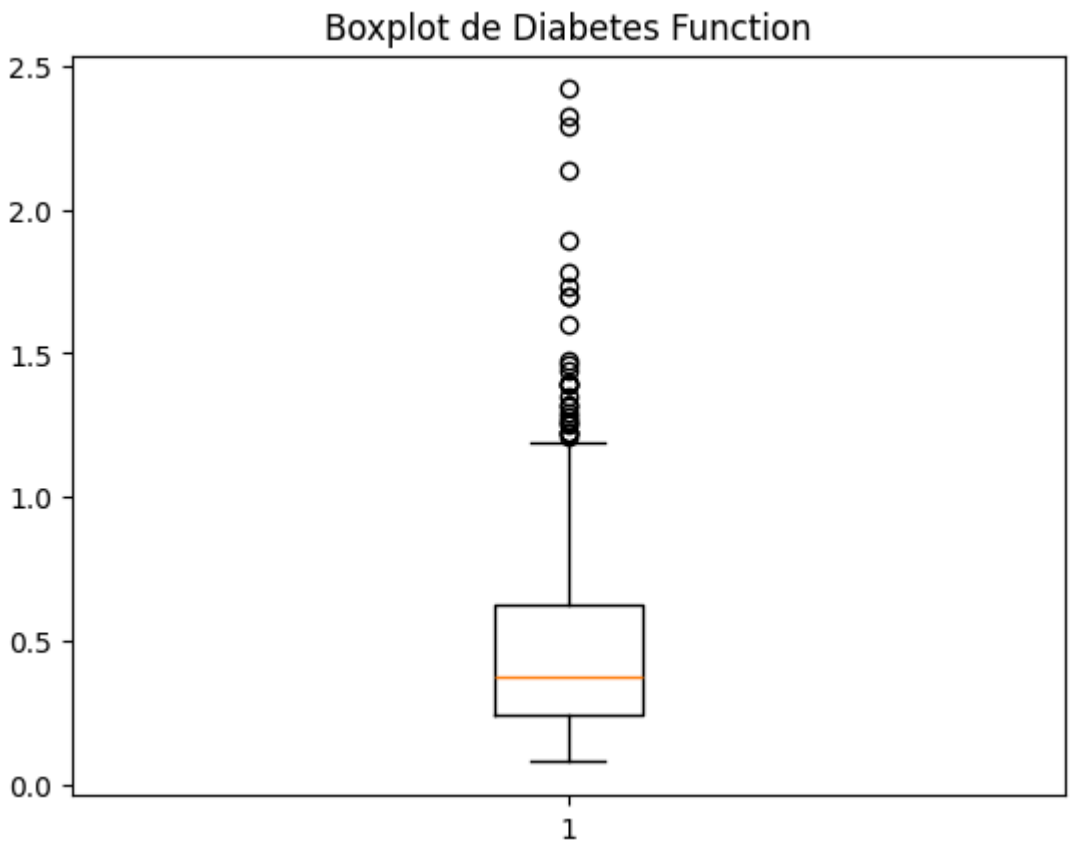


El análisis del boxplot de SkinThickness muestra que el 25% de los pacientes tiene un grosor de piel por debajo de 0.0, lo que indica la presencia de registros faltantes o atípicos en los datos. La mediana se encuentra en 23.0, reflejando el valor central de la distribución. Además, el 25% de los pacientes presenta un grosor de piel mayor a 32.0. . Además, se observa un valor atípico extremo cercano a 100, mucho más alto

que el resto de los datos. Esto indica que, aunque la mayoría de los pacientes se concentra en un rango bajo o medio, existen registros muy alejados de la tendencia central que podrían influir en el análisis si no se consideran o se tratan adecuadamente.

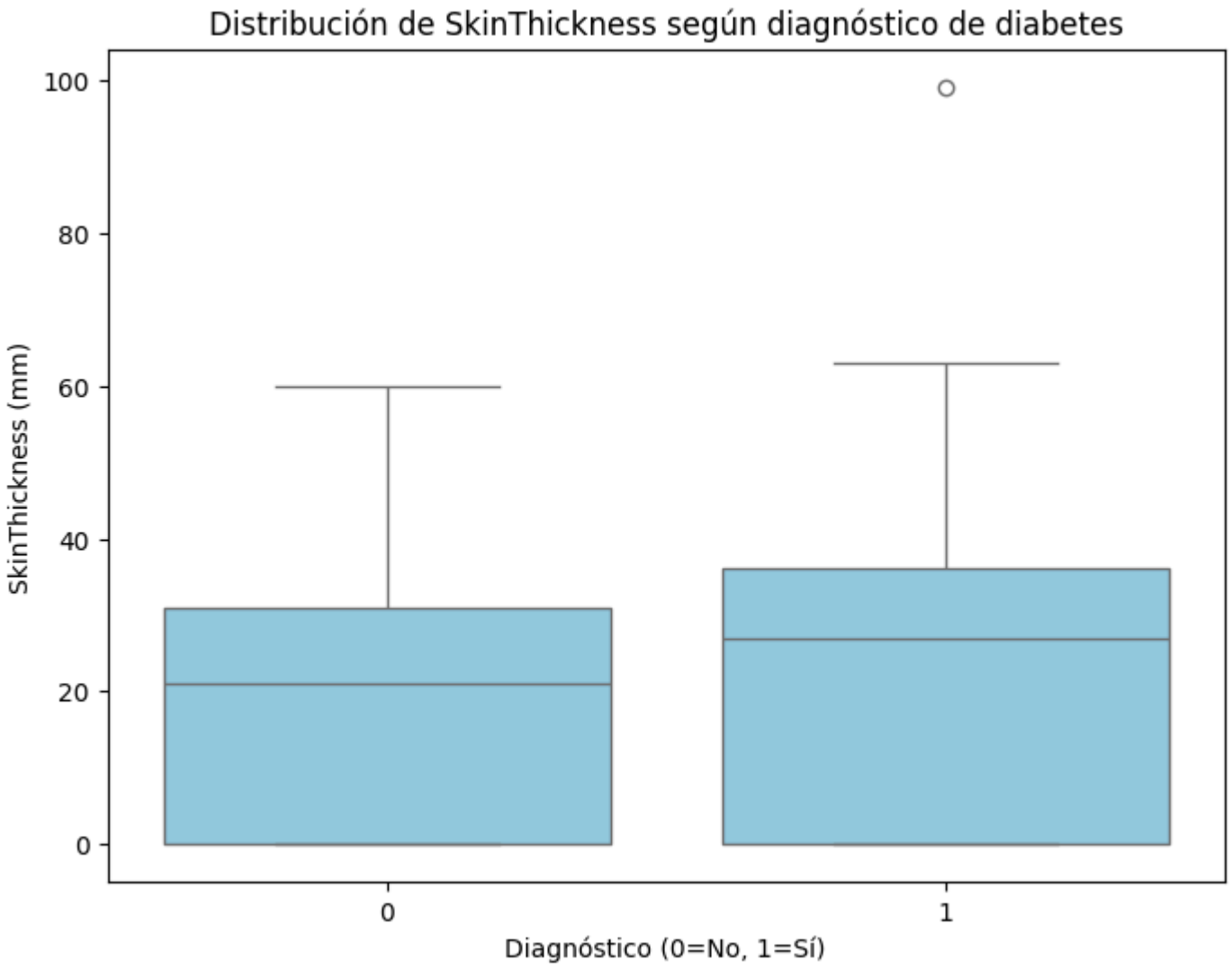
In [56]:
```
plt.boxplot(diabetes['DiabetesPedigreeFunction'])
plt.title('Boxplot de Diabetes Function')
```

Out[56]: Text(0.5, 1.0, 'Boxplot de Diabetes Function')



El boxplot de Diabetes Pedigree Function (DPF) indica que el 25% de los pacientes tiene un valor de DPF por debajo de 0.244, la mediana se encuentra en 0.373 y el 25% superior de los pacientes tiene valores por encima de 0.626. Esto muestra que la mayoría de los pacientes tiene una predisposición genética baja o moderada a la diabetes, mientras que un pequeño grupo presenta valores relativamente altos, aunque no tan extremos como para considerarse outliers tan marcados como en otras variables. En la gráfica también se identifican varios puntos por encima del límite superior del boxplot, los cuales corresponden a valores atípicos y representan pacientes cuya predisposición genética a la diabetes es significativamente más alta que la mayoría del grupo.

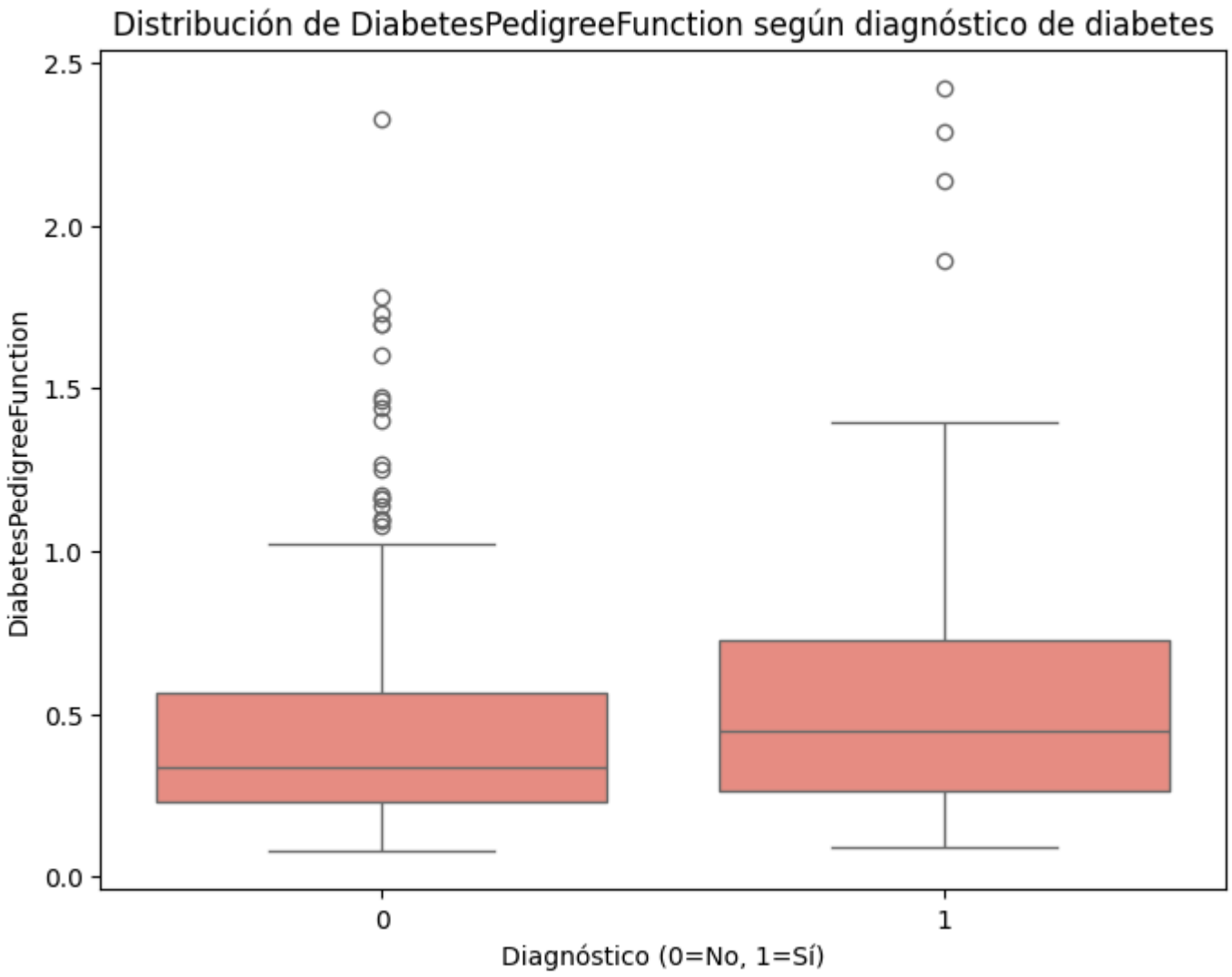
In [66]:
```
# Boxplot de SkinThickness según Outcome
plt.figure(figsize=(8,6))
sns.boxplot(data=diabetes, x='Outcome', y='SkinThickness', color='skyblue')
plt.title('Distribución de SkinThickness según diagnóstico de diabetes')
plt.xlabel('Diagnóstico (0=No, 1=Sí)')
plt.ylabel('SkinThickness (mm)')
plt.show()
```



En este caso analizamos la variable SkinThickness respecto al diagnóstico de diabetes y observamos que pacientes con un resultado positivo (1) presentan valores mayores de grosor de la piel a diferencia de aquellos que no tienen diabetes. En detalle, el 25% de los pacientes sin

diabetes tienen SkinThickness por debajo de 0 mm, la mediana es de 21 mm y el 25% superior tiene valores por encima de 31 mm. Por otro lado, los pacientes con diabetes también tienen un 25% de valores por debajo de 0 mm, pero la mediana aumenta a 27 mm y el 25% superior alcanza valores mayores a 36 mm.

In [69]:
```
# Boxplot de DiabetesPedigreeFunction según Outcome
plt.figure(figsize=(8,6))
sns.boxplot(data=diabetes, x='Outcome', y='DiabetesPedigreeFunction', color='salmon')
plt.title('Distribución de DiabetesPedigreeFunction según diagnóstico de diabetes')
plt.xlabel('Diagnóstico (0=No, 1=Sí)')
plt.ylabel('DiabetesPedigreeFunction')
plt.show()
```



El boxplot de DiabetesPedigreeFunction (DPF) según el diagnóstico de diabetes nos permite observar la distribución de la predisposición genética en ambos grupos. Para los pacientes sin diabetes (Outcome=0), el 25% presenta un DPF por debajo de 0.22975, la mediana es de 0.336 y el 25% restante tiene valores por encima de 0.56175. En contraste, los pacientes con diabetes (Outcome=1) muestran valores generalmente más altos: el 25% tiene DPF por debajo de 0.2625, la mediana es de 0.449 y el 25% superior alcanza hasta 0.728. Esto indica que los individuos con diabetes tienden a tener una predisposición genética mayor. Sin embargo existen valores atípicos altos en pacientes que sin diabetes, esto significa que algunos individuos sin diagnóstico positivo presentan una predisposición genética relativamente elevada, incluso por encima del rango típico de su grupo. Estos casos pueden ser observaciones inusuales o simplemente personas con alto riesgo genético que aún no desarrollan la enfermedad.

In [50]:
```
# seleccionar variables numéricas
variables_numericas = diabetes.select_dtypes(include='number')
matriz_correlacion = variables_numericas.corr().round(2)
matriz_correlacion
```

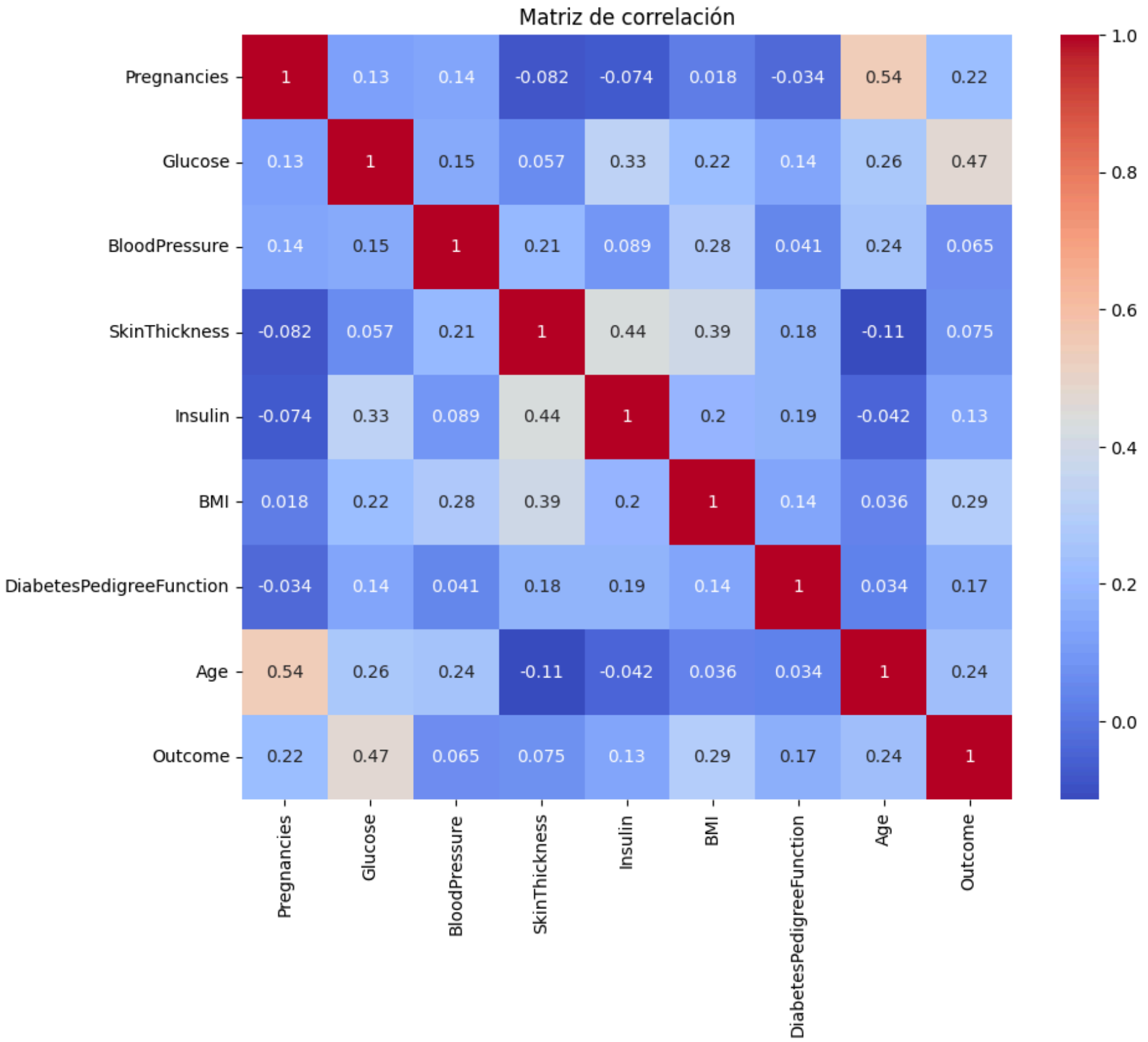
Out[50]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

In [53]:
```
# Matriz de correlación

plt.figure(figsize=(10,8))
sns.heatmap(diabetes.corr(), annot=True, cmap="coolwarm")
```

```
plt.title("Matriz de correlación")
plt.show()
```



El heatmap muestra la matriz de correlaciones entre las variables del dataset de diabetes, permitiendo identificar relaciones lineales entre ellas. Se observa que la variable Glucose tiene la correlación positiva más fuerte con el Outcome (0.47), lo que indica que niveles más altos de glucosa están asociados con un diagnóstico de diabetes. Otra relación destacable es la correlación moderada positiva entre SkinThickness e Insulin (0.44), y entre BMI y SkinThickness (0.39), sugiriendo que pacientes con mayor grosor de piel tienden a tener niveles más altos de insulina y un índice de masa corporal mayor. La variable Age se correlaciona positivamente con Pregnancies (0.54), lo que refleja que pacientes de mayor edad suelen haber tenido más embarazos. En general, la mayoría de las correlaciones son bajas o moderadas, indicando que, aunque existen tendencias, no todas las variables están fuertemente relacionadas linealmente. El heatmap es útil para identificar patrones potenciales y relaciones que pueden ser relevantes para análisis posteriores o modelos predictivos.

Preguntas

Responde al final de esta sección, en un apartado, las siguientes preguntas:

¿Hay alguna variable que no aporta información?

- Todas las variables aportan información relevante para el análisis del diagnóstico de diabetes. Sin embargo, algunas columnas tienen muchos ceros que podrían representar datos faltantes más que valores reales, por lo que su utilidad podría verse limitada si no se imputan.

Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

- Insulin y SkinThickness tienen un número alto de ceros (374 y 227 respectivamente), lo que indica muchos datos faltantes. Si no se imputan, podrían distorsionar los análisis estadísticos o modelos predictivos.

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

- Al analizar los rangos de las variables del dataset, se observa que no todas están en rangos similares:
- Pregnancies: de 0 a 17, indicando que la mayoría de pacientes han tenido pocos embarazos, pero hay casos con hasta 17.

- Glucose: de 0 a 199, mostrando una gran variabilidad y que algunos pacientes tienen niveles de glucosa extremadamente altos.
- BloodPressure: de 0 a 122 mmHg, con la mayoría de los valores concentrados alrededor de la mediana (72).
- SkinThickness: de 0 a 99 mm, lo que indica presencia de valores atípicos, dado que el 75% de los pacientes tienen ≤32 mm.
- Insulin: de 0 a 846 µU/mL, también con valores atípicos evidentes, ya que la mediana es 30.5 y el 75% de los pacientes ≤127.25.
- BMI: de 0 a 67.1, con la mayoría de pacientes en rangos normales y solo unos pocos extremadamente altos.
- DiabetesPedigreeFunction (DPF): de 0.078 a 2.42, mostrando que la mayoría de los pacientes tienen valores bajos, pero algunos presentan predisposición genética muy alta.
- Age: de 21 a 81 años, indicando que el dataset cubre un rango amplio de edades adultas.
- Outcome: 0 o 1, ya que es una variable categórica que indica diagnóstico de diabetes.

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

- Sí, existen valores atípicos notables:
- SkinThickness: presenta valores extremos; el máximo es 99 mientras que la mediana es 23 y el tercer cuartil 32, lo que indica pacientes con un grosor de piel mucho mayor al resto.
- Insulin: tiene un valor máximo de 846, muy por encima del tercer cuartil (127.25), mostrando niveles de insulina extremadamente altos en algunos pacientes.
- DiabetesPedigreeFunction (DPF): hay valores superiores a 2.0, mientras que el tercer cuartil es 0.62625, indicando casos atípicos con predisposición genética mucho más alta.
- Glucose: aunque la mayoría se concentra alrededor de valores moderados, hay algunos pacientes con glucosa cercana a 199, alejados del tercer cuartil (140.25).

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

La mayoría de las correlaciones son bajas a moderadas, pero destacan:

- Glucose – Outcome: 0.47 (positiva) → niveles altos de glucosa asociados a diagnóstico de diabetes.
- SkinThickness – Insulin: 0.44 (positiva) → mayor grosor de piel se relaciona con mayores niveles de insulina.
- BMI – SkinThickness: 0.39 (positiva) → pacientes con mayor grosor de piel suelen tener un IMC más alto.
- Age – Pregnancies: 0.54 (positiva) → pacientes mayores suelen tener más embarazos.