In [ ]:
```
A00840366 Nicolás Alvarez Gonzalez - Licenciatura en Derecho
```

In [38]:
```
import pandas as pd
#importa librerías
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import sklearn
```

In [4]:
```
df = pd.read_csv("diabetes.csv")
```

In [10]:
```
df.head
```

Out[10]:
```
<bound method NDFrame.head of      Pregnancies  Glucose  BloodPressure  SkinThickn
ess  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]>
```

In [11]:
```
df.tail
```

Out[11]:
```
<bound method NDFrame.tail of      Pregnancies  Glucose  BloodPressure  SkinThickn
ess  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]>
```

In [12]:
```
df.info
```

Out[12]:
```
<bound method DataFrame.info of      Pregnancies  Glucose  BloodPressure  SkinThic
kness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]>
```

In [13]:
```
df.nunique
```

Out[13]:
```
<bound method DataFrame.nunique of      Pregnancies  Glucose  BloodPressure  SkinT
hickness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]>
```

In [15]:
```
df[['BMI', 'Age']].describe()
```

Out[15]:

| | BMI | Age |
|---|---|---|
| **count** | 768.000000 | 768.000000 |
| **mean** | 31.992578 | 33.240885 |
| **std** | 7.884160 | 11.760232 |
| **min** | 0.000000 | 21.000000 |
| **25%** | 27.300000 | 24.000000 |
| **50%** | 32.000000 | 29.000000 |
| **75%** | 36.600000 | 41.000000 |
| **max** | 67.100000 | 81.000000 |

In [16]:
```
media_bp = df["BMI"].mean()
mediana_bp = df["BMI"].median()
desv_bp = df["BMI"].std()
media_skin = df["Age"].mean()
mediana_skin = df["Age"].median()
desv_skin = df["Age"].std()
print("BMI:")
```

```
print(f" Media: {media_bp:.2f}")
print(f" Mediana: {mediana_bp:.2f}")
print(f" Desviación estándar: {desv_bp:.2f}\n")
print("Age:")
print(f" Media: {media_skin:.2f}")
print(f" Mediana: {mediana_skin:.2f}")
print(f" Desviación estándar: {desv_skin:.2f}")
```

```
BMI:
 Media: 31.99
 Mediana: 32.00
 Desviación estándar: 7.88

Age:
 Media: 33.24
 Mediana: 29.00
 Desviación estándar: 11.76
```

In [17]:
```
df.iloc[0]
```

Out[17]:
```
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 [18]:
```
df.iloc[0:2]
```

Out[18]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunc |
|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0 |

In [19]:
```
df[['BMI', 'Age']]
```

Out[19]:

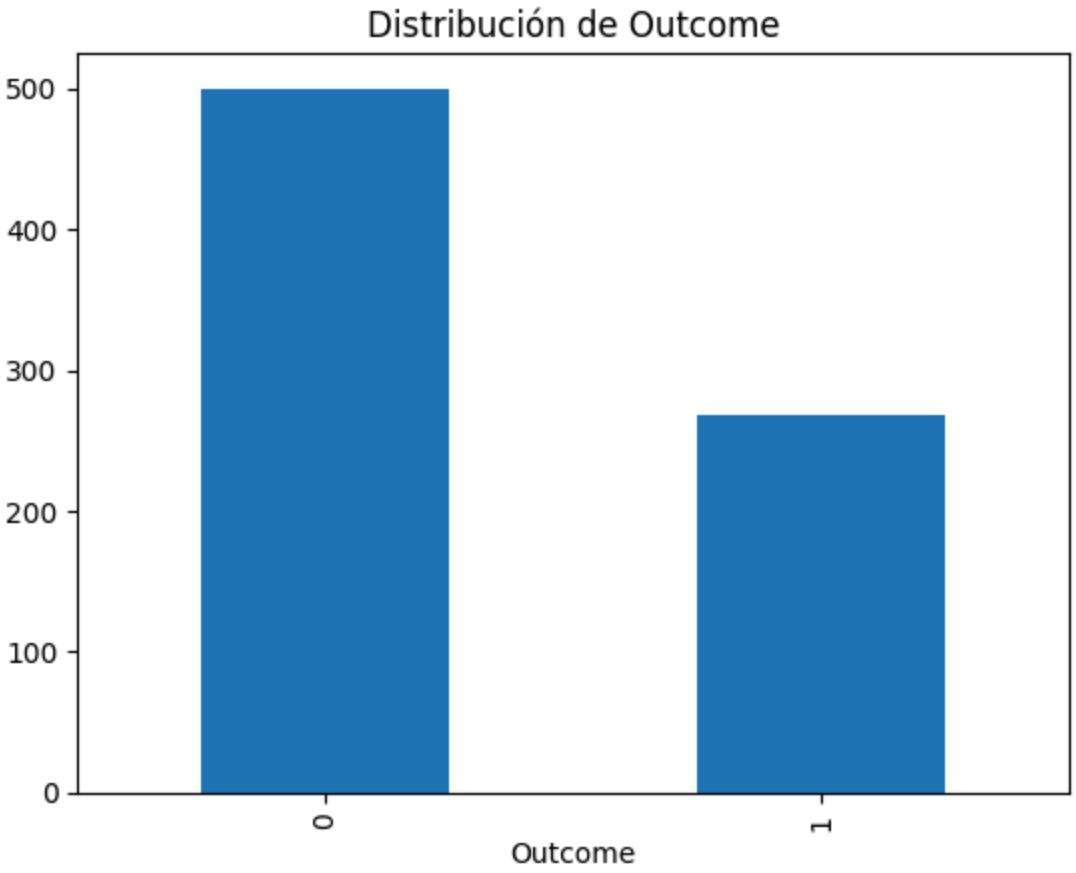
| | BMI | Age |
|---|---|---|
| **0** | 33.6 | 50 |
| **1** | 26.6 | 31 |
| **2** | 23.3 | 32 |
| **3** | 28.1 | 21 |
| **4** | 43.1 | 33 |
| **...** | ... | ... |
| **763** | 32.9 | 63 |
| **764** | 36.8 | 27 |
| **765** | 26.2 | 30 |
| **766** | 30.1 | 47 |
| **767** | 30.4 | 23 |

768 rows × 2 columns

# Visualización de Datos, variable 1

## outcome

In [72]:
```
indice_de_masa_corporal = df['Outcome'].value_counts()
indice_de_masa_corporal.plot(kind='bar')
plt.title('Distribución de Outcome')
plt.show()
```

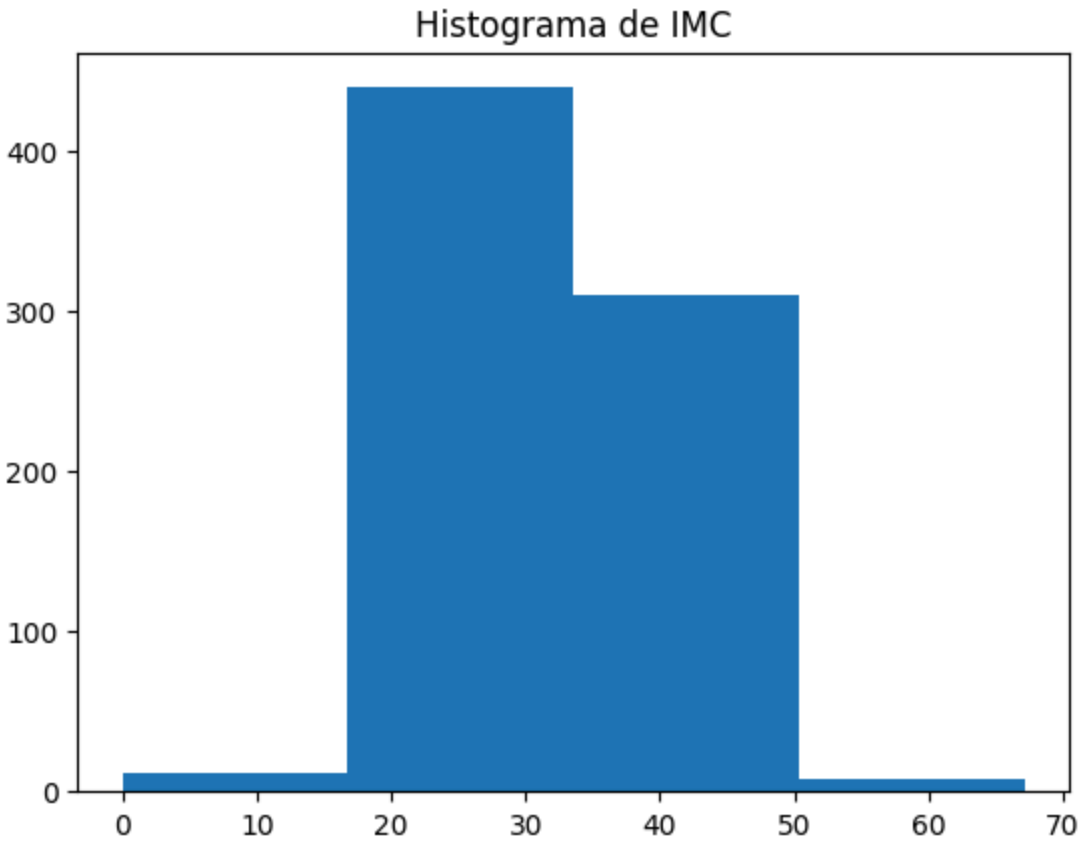



se muestra la distribución de la variable objetivo, que indica si el paciente tiene diabetes o no. Se observa que la mayoría de los pacientes en el conjunto de datos no tienen diabetes, con un recuento cercano a 500, mientras que el número de pacientes que sí tienen diabetes es significativamente menor, aproximadamente 270.

# IMC

## histograma

In [73]:
```
plt.hist(df['BMI'], bins=4)
plt.title('Histograma de IMC')
plt.show()
```




se muestra la distribución de los valores del Índice de Masa Corporal. La mayor concentración de datos se encuentra en el segundo bin, que parece cubrir un rango de IMC aproximadamente entre 18 y 35. Los bins inicial y final tienen muy pocos datos, y hay una concentración notable de datos en el rango de sobrepeso/obesidad, lo cual es relevante en el contexto de la diabetes.

## boxplot

In [75]:
```
print("Mediana:",df['BMI'].median())
print("Media:",df['BMI'].mean())
print(df['BMI'].dtype)
df["BMI"].isnull().sum()
```

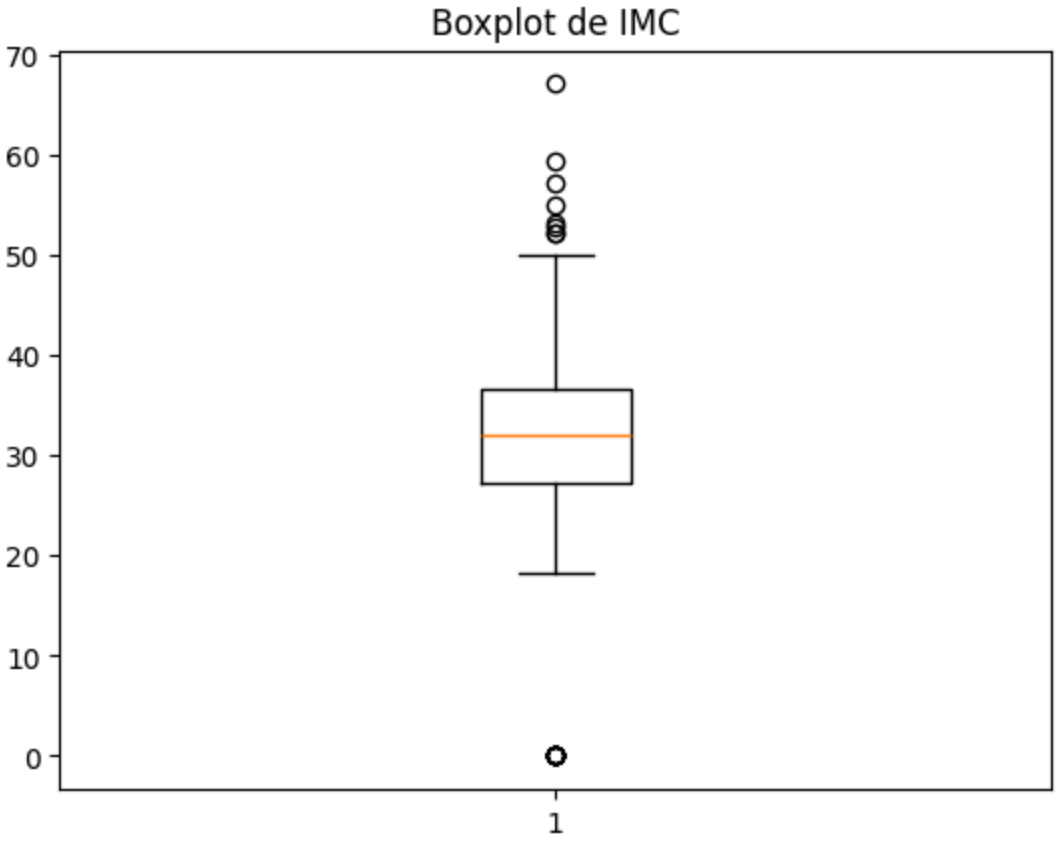
```
Mediana: 32.0
Media: 31.992578124999998
float64
```

Out[75]: np.int64(0)

In [76]:
```
df['BMI']=df['BMI'].fillna(df['BMI'].median())
```

In [77]:
```
plt.boxplot(df['BMI'])
plt.title('Boxplot de IMC')
```

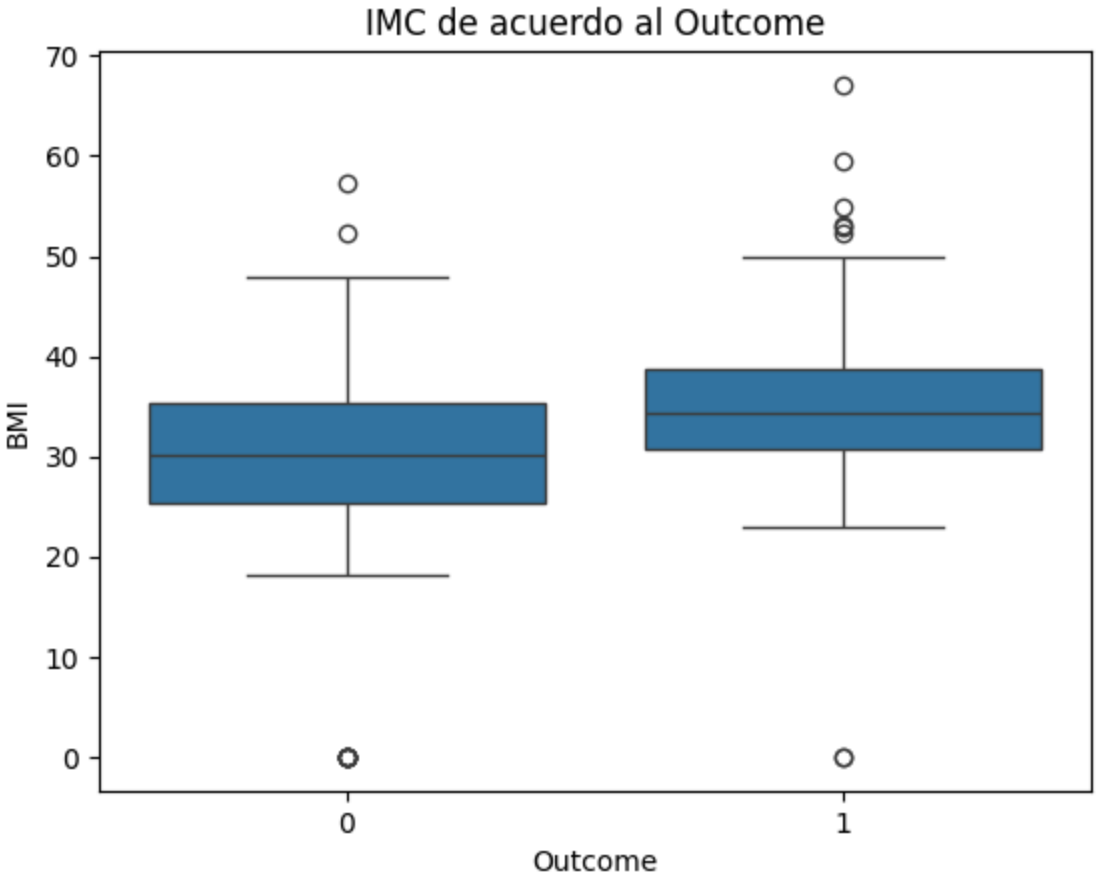
Out[77]: Text(0.5, 1.0, 'Boxplot de IMC')




La distribución de la variable se presenta con una mediana de 32.0. El 50/% central de los datos se encuentra en el rango entre el primer cuartil y el tercer cuartil. se observan valores atípicos, representados por puntos individuales, tanto en el extremo superior como en el extremo inferior. Destaca un valor atípico de 0.0 para el IMC que está cerca del extremo inferior y podría ser indicativo de datos faltantes o un error de registro.

In [78]:
```
sns.boxplot(df, x="Outcome", y="BMI")
plt.title("IMC de acuerdo al Outcome")
```

Out[78]:
```
Text(0.5, 1.0, 'IMC de acuerdo al Outcome')
```




Al comparar las distribuciones de Índice de Masa Corporal para los dos grupos de Outcome, es evidente que las personas con diabetes presentan un IMC medio y una dispersión más altos que las personas sin diabetes. La mediana del IMC para el grupo 1 es notablemente superior a la mediana del grupo 0, lo que sugiere una asociación clara entre la presencia de diabetes y valores más elevados de IMC.

## Matriz de Correlación

In [70]:
```
variables_numericas = df.select_dtypes(include='number')
matriz_correlacion = variables_numericas.corr().round(2)
matriz_correlacion
```

Out[70]:

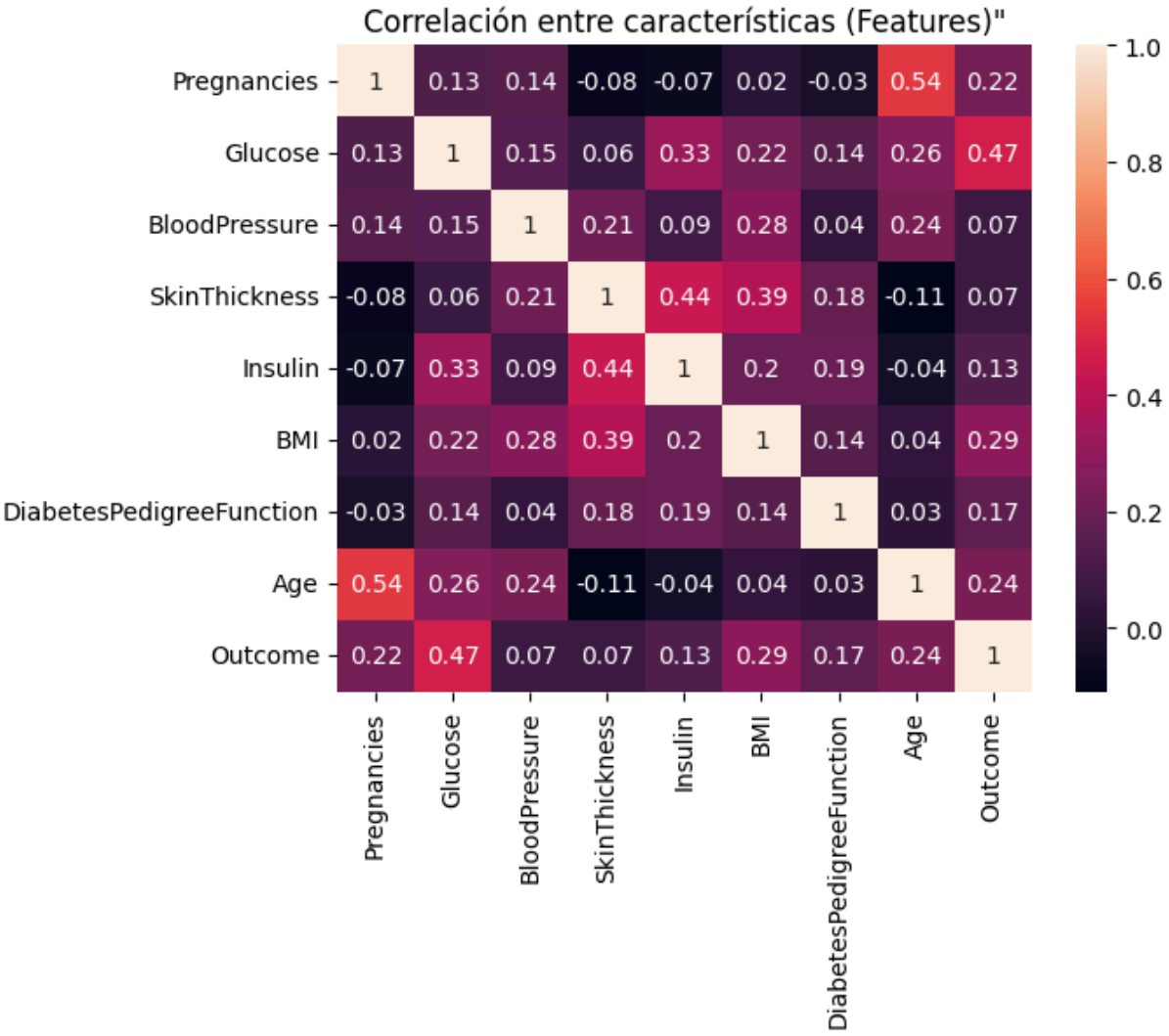
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BM |
|---|---|---|---|---|---|---|
| **Pregnancies** | 1.00 | 0.13 | 0.14 | -0.08 | -0.07 | 0.0 |
| **Glucose** | 0.13 | 1.00 | 0.15 | 0.06 | 0.33 | 0.2 |
| **BloodPressure** | 0.14 | 0.15 | 1.00 | 0.21 | 0.09 | 0.2 |
| **SkinThickness** | -0.08 | 0.06 | 0.21 | 1.00 | 0.44 | 0.3 |
| **Insulin** | -0.07 | 0.33 | 0.09 | 0.44 | 1.00 | 0.2 |
| **BMI** | 0.02 | 0.22 | 0.28 | 0.39 | 0.20 | 1.0 |
| **DiabetesPedigreeFunction** | -0.03 | 0.14 | 0.04 | 0.18 | 0.19 | 0.1 |
| **Age** | 0.54 | 0.26 | 0.24 | -0.11 | -0.04 | 0.0 |
| **Outcome** | 0.22 | 0.47 | 0.07 | 0.07 | 0.13 | 0.2 |

In [ ]:

## Mapa de calor

In [71]:
```
sns.heatmap(matriz_correlacion, annot=True)
plt.title('Correlación entre características (Features)"')
```

Out[71]:
```
Text(0.5, 1.0, 'Correlación entre características (Features)"')
```

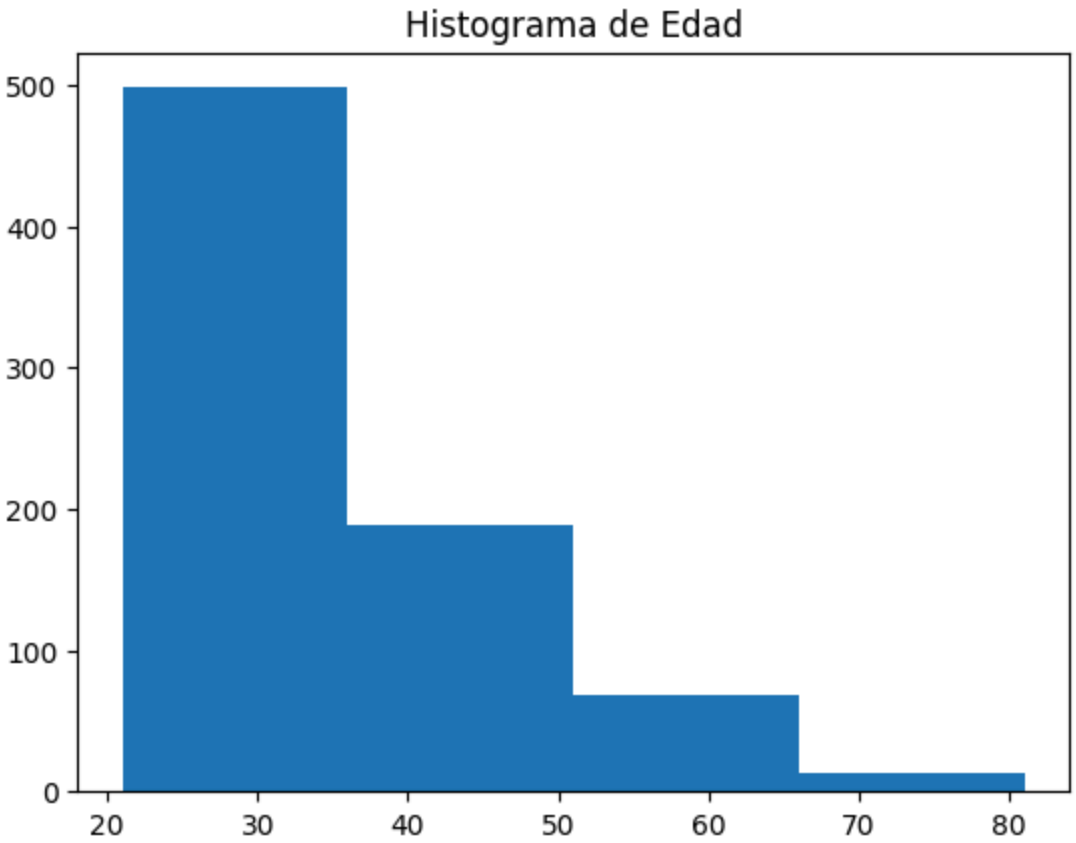



# Visualización de Datos, variable 2

## Edad

### histograma

In [59]:
```
plt.hist(df['Age'], bins=4)
plt.title('Histograma de Edad')
plt.show()
```




se muestra que la mayoría de los pacientes en el conjunto de datos se encuentran en el grupo de edad más joven, específicamente en el primer bin, que va de los 20 a aproximadamente los 35 años. Este grupo inicial presenta una frecuencia de alrededor de 500 individuos. La frecuencia de pacientes disminuye significativamente a medida que aumenta la edad, lo que resulta en una distribución sesgada a la derecha.

## boxplot

In [60]:
```
print("Mediana:",df['Age'].median())
print("Media:",df['Age'].mean())
print(df['Age'].dtype)
df["Age"].isnull().sum()
```

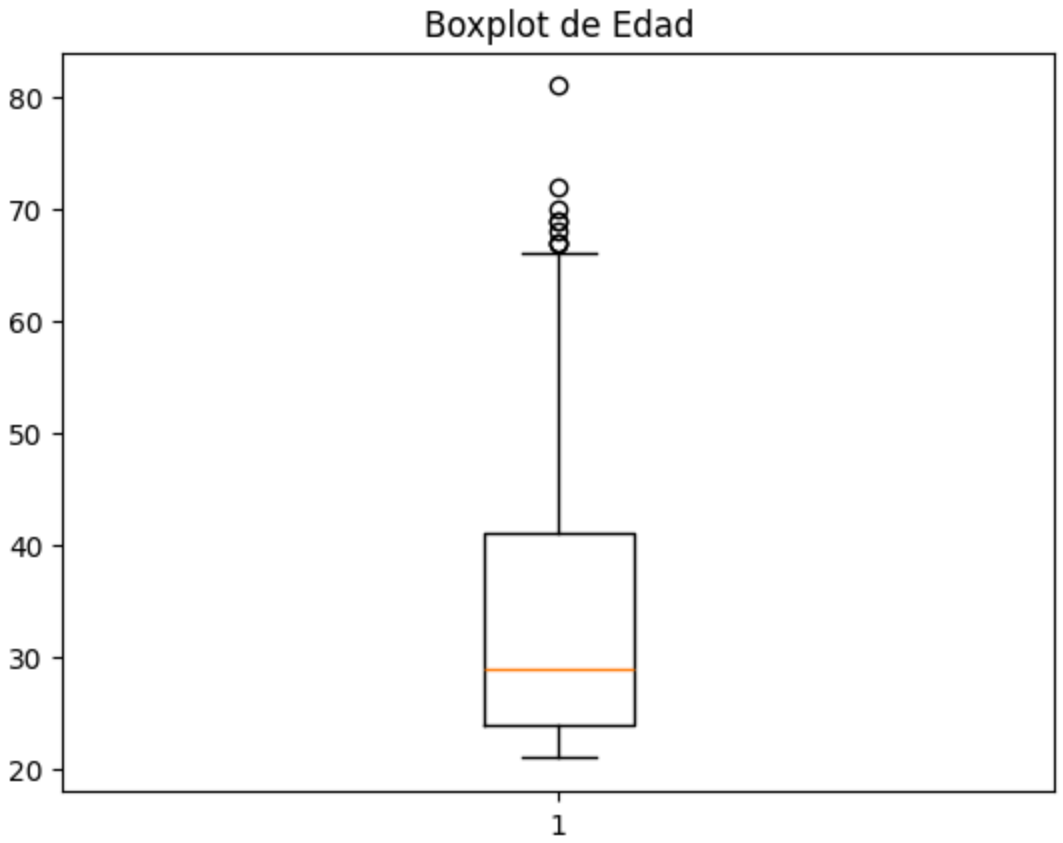
```
Mediana: 29.0
Media: 33.240885416666664
int64
```

Out[60]: np.int64(0)

In [61]:
```
df['Age']=df['Age'].fillna(df['Age'].median())
```

In [62]:
```
plt.boxplot(df['Age'])
plt.title('Boxplot de Edad')
```

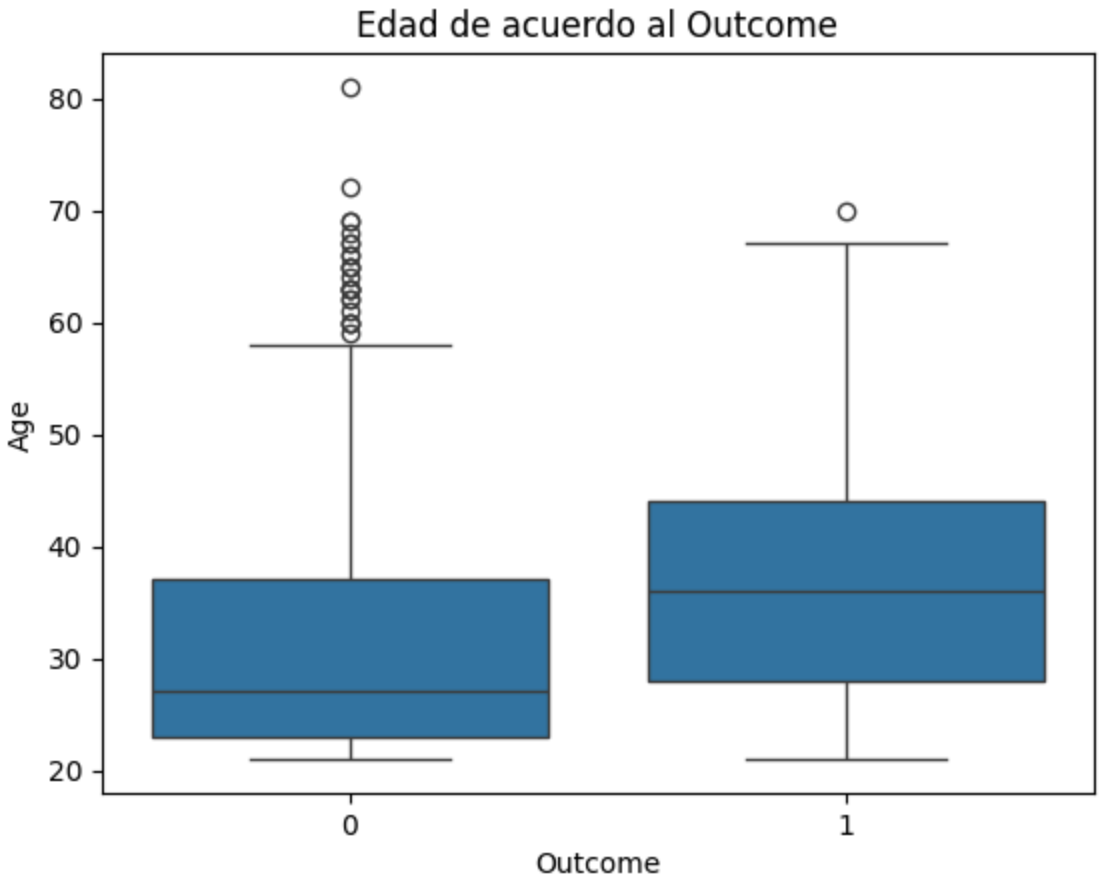
Out[62]: Text(0.5, 1.0, 'Boxplot de Edad')




La distribución de la edad presenta una mediana que se sitúa en 29.0 años. El 50/% central de los datos se encuentra en el rango comprendido entre el primer cuartil y el tercer cuartil. Se aprecian numerosos valores atípicos en el extremo superior, lo que indica que hay una cantidad considerable de pacientes de edad avanzada entre 65 y 81 años en la muestra.

In [80]:
```
sns.boxplot(df, x="Outcome", y="Age")
plt.title("Edad de acuerdo al Outcome")
```

Out[80]:
```
Text(0.5, 1.0, 'Edad de acuerdo al Outcome')
```




Al comparar las distribuciones de edad para los grupos con y sin diabetes, se observa que las personas con diabetes tienden a ser más mayores que las personas sin diabetes. Esta diferencia se refleja en el hecho de que el rango intercuartílico y la mediana son visiblemente más altos en el grupo con diabetes. La mediana de edad para el grupo con diabetes parece estar en el rango de los 35 a 40 años, mientras que para el grupo sin diabetes se sitúa alrededor de los 25 a 30 años.

## Matriz de Correlación

In [64]:
```
variables_numericas = df[['BMI', 'Age', 'Outcome']]
matriz_correlacion = variables_numericas.corr().round(2)
matriz_correlacion
```

Out[64]:

| | BMI | Age | Outcome |
|---|---|---|---|
| **BMI** | 1.00 | 0.04 | 0.29 |
| **Age** | 0.04 | 1.00 | 0.24 |
| **Outcome** | 0.29 | 0.24 | 1.00 |

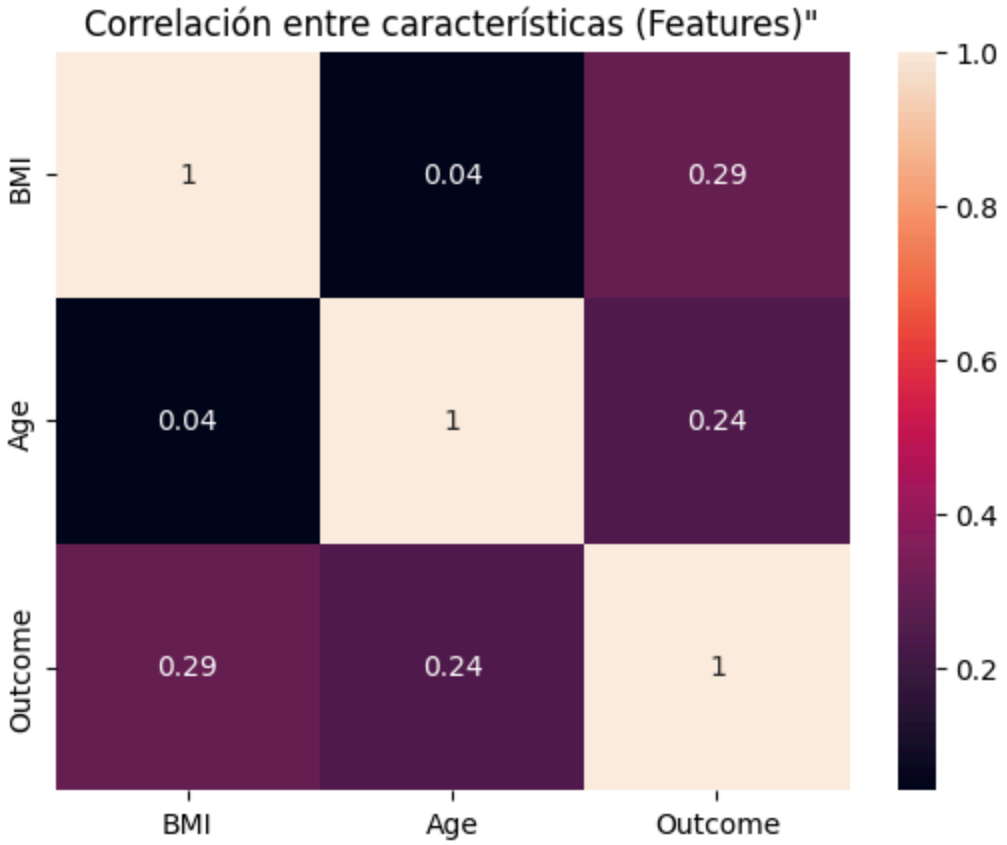
El análisis de correlación revela que la variable Outcome muestra una correlación positiva moderada tanto con el Índice de Masa Corporal, con un valor de 0.29, como con la Edad, con un valor de 0.24. Este hallazgo sugiere que un IMC y una edad mayores están

moderadamente asociados con una mayor probabilidad de tener diabetes. Es importante notar que la correlación entre el IMC y la Edad es muy baja, lo que indica que estas dos variables predictoras son casi independientes entre sí.

### Mapa de Calor

In [65]:
```
sns.heatmap(matriz_correlacion, annot=True)
plt.title('Correlación entre características (Features)"')
```

Out[65]:
```
Text(0.5, 1.0, 'Correlación entre características (Features)"')
```




# Preguntas Finales

## ¿Hay alguna variable que no aporta información?

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

La variable DiabetesPedigreeFunction es la que menos aporta al resultado, mostrando la correlación más baja con el Outcome. Además, Insulin y SkinThickness contienen un alto número de valores cero, que son biológicamente imposibles y representan datos faltantes mal codificados. Si se necesitara reducir variables, yo eliminaría estas dos últimas debido a la

alta distorsión que introducen sus 0's, o al menos tratarlas como datos faltantes antes de modelar.

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

No, las variables no están en rangos similares, lo que hace necesaria una estandarización de los datos. Los rangos varían drásticamente, desde el muy estrecho de DiabetesPedigreeFunction, pasando por el rango moderado de BMI y Age, hasta el muy amplio de Insulin.

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

Sí, existen datos atípicos en varias variables. Los boxplots confirman claramente la presencia de datos atípicos en BMI y en Age. Además, se infiere que Insulin, SkinThickness, Glucose y BloodPressure también contienen datos atípicos, principalmente debido a los valores de 0, que son anómalos en un contexto biométrico.

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

Sí, existe una correlación positiva moderada a alta entre algunas variables. La correlación más fuerte entre predictores es positiva entre Pregnancies y Age. Respecto a la variable objetivo, la correlación más alta es positiva con Glucose, seguida por BMI y Age. Esto indica que a mayor glucosa, mayor IMC y mayor edad, mayor es la probabilidad de tener diabetes.