Actividad Evaluable: Obtención de estadisticas descriptivas

Regina Echavarría Torres A00841096

Carga de Datos

In [1]:
```
import pandas as pd
```

In [13]:
```
#1. Cargar los datos
df = pd.read_csv('diabetes.csv')
```

In [10]:
```
#2.
df.shape
```

Out[10]: (768, 9)

Hay 768 datos y 9 variables diferentes.

In [11]:
```
#3. Informacion general
df.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
```

Analisis de variables seleccionadas: "DiabetesPedigreeFunction" y "Age"

Descripción

DiabetesPedigreeFunction

cuantitativa y continua

Age

cuantitativa discreta

Outcome

categórica binaria

In [37]:
```
#Analisis de variables
df[["DiabetesPedigreeFunction", "Age", "Outcome"]].describe()
```

Out[37]:

	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000
mean	0.471876	33.240885	0.348958
std	0.331329	11.760232	0.476951
min	0.078000	21.000000	0.000000
25%	0.243750	24.000000	0.000000
50%	0.372500	29.000000	0.000000
75%	0.626250	41.000000	1.000000
max	2.420000	81.000000	1.000000

que representan y sus rangos

DiabetesPedegreeFunction representa la probabilidad de tener diabetes basandose en la historia familiar su rango es de 0.078 a 1, no se porque dice que el valor máximo es de 2.42 si lo que describe es una probabilidad.

Age representa la edad de las personas con diabetes, su rango es de 21 años a 81 años.

Outcome es una variable dependiente de las demás variables, solo puede tener dos valores: "1" idíca que sí tiene diabetes, "0" indica que no tiene diabetes.

Conclusiones de los datos

A juzgar por la media, mediana y desviación estandar de los datos, pude llegar a las siguientes conclusiones.

De la variable DiabetesPedegreeFunction puedo concluir que hay en promedio un 37% de probabilidad que una persona tenga diabetes basandose en el historial familiar. Como

la desviación estántar es de 33% se puede concluir que los datos estan muy dispersos, por lo que puede que para algunas personas sea muy poca la infuencia familiar y para otras sea un factor determinante.

De la variable Age sabemos que en promedio las personas que fueron encuestadas tienen 33 años, con una desviación estandar de 11.76 años. Lo que indica que se encuestaron en su mayoria adultos de mediana edad.

De la variable outcome la media es de 0.35, como es una variable que solo puede tomar valor 1 o 0, puedo concluir que son menos las personas que tienen diabetes. Si fueran la misma cantidad de valores 1 y 0, la media seria 0.5. Como es menor a 0.5, se puede decir que hay más personas que no tienen diabetes.

Consulta

1. Cuantos jovenes diabeticos hay

In [38]:
```
jovendiabetes = (df['Outcome'] == 1) & (df['Age'] <= 30)
df_jovendiabetes = df.loc[jovendiabetes, ['DiabetesPedigreeFunction', 'Outcc
print(df_jovendiabetes)
df_jovendiabetes.shape[0]
```

```
     DiabetesPedigreeFunction  Outcome  Age
6                       0.248        1   26
23                      0.263        1   29
31                      0.851        1   28
38                      0.503        1   27
45                      1.893        1   25
..                        ...      ...  ...
731                     0.259        1   22
732                     0.646        1   24
746                     0.358        1   27
750                     1.182        1   22
753                     0.222        1   26

[90 rows x 3 columns]
```

Out[38]: 90

De las 768 personas encuestadas 90 son menores de 30 años y tienen diabetes.

2. Personas con alto riesgo

In [39]:
```
altoriesgo = df['DiabetesPedigreeFunction'] > 1.0
altoriesgop = df.loc[altoriesgo, ['DiabetesPedigreeFunction', 'Outcome', 'Ag
altoriesgop.shape[0]
```

Out[39]: 51

In [43]:
```
altoriesgop.head()
```

Out[43]:

	DiabetesPedigreeFunction	Outcome	Age
4	2.288	1	33
12	1.441	0	57
39	1.390	1	56
45	1.893	1	25
58	1.781	0	44

De las 768 personas encuestadas hay 51 que están en alto riezgo de tener diabetes debido a su historia familiar.

3. Cuantos jovenes tienen alto riezgo

In [41]:
```
jovenesriezgo = (df['DiabetesPedigreeFunction'] > 1.0)& (df['Age'] <= 30)
df_jovenesriezgo = df.loc[jovenesriezgo, ['DiabetesPedigreeFunction', 'Outcc
df_jovenesriezgo.shape[0]
```

Out[41]: 18

In [44]:
```
df_jovenesriezgo.head()
```

Out[44]:

	DiabetesPedigreeFunction	Outcome	Age
45	1.893	1	25
220	1.072	1	21
267	1.101	0	24
308	1.391	1	25
370	2.137	1	25

De las 51 personas que tienen alto riezgo de tener diabetes por historial familiar, 18 son menores de 30 años. Lo que indica que la mayoria de las personas que estan en riezgo de contraer diabetes son mayores a 30 años.