**Diabetes dataset**

**sama Mohamed 247669**  **Dina Ahmed 244445**

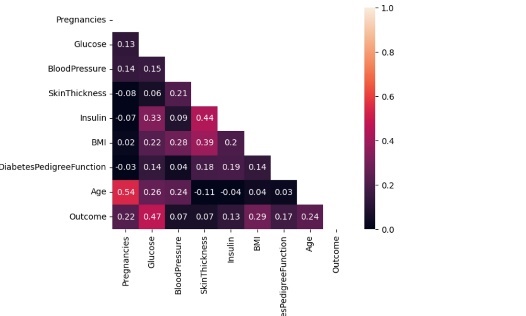
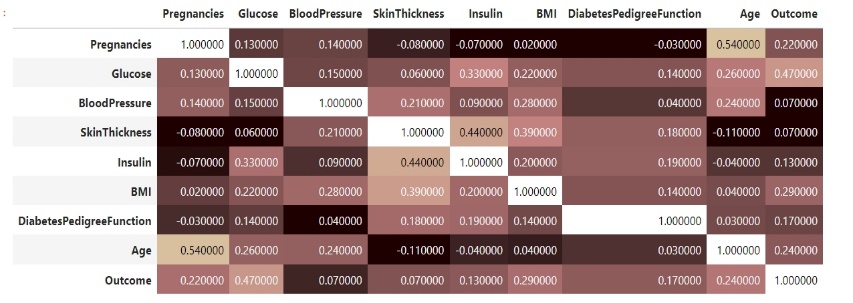
**Visualizations**

1. **Correlation Matrix:**

Type: Heatmap

Definition : A correlation matrix is a table showing correlation coefficients between variables and coff in the dataset .

The heatmap helps identify which features are strongly correlated with each other



Insights:

**Glucose and outcome:** there is a high positive correlation between glucose and outcome=0.47, so the higher the glucose level the higher the risk of diabetes

**BMI and outcome**: there is a high positive correlation between BMI and Outcome =0.29

Therefore the higher BMI the higher the risk of diabetes.

**age and outcome**: there is a high positive correlation between age and outcome =0.24 .Therefore, the older the person is the higher the risk of diabetes .

**skinThickness and outcome :** A week correlation between SkinThickness and outcome =0.07indicates that there is no relation between the SkinThickness and outcome

**pregnancies and age**: A correlation between pregnancies and age=0.54 indicate that younger women have more pregnancies.

1. **Scotter plot :**

Definition: A scatter plot is a type of data visualization that displays the relationship between two continuous variables. Each point on the plot represents an individual data point from the dataset

**Insights:**

The higher your bmi is the higher the risk of having a high bloodpressure **A diagram showing a number of blood pressure

Description automatically generated**

1. **Pie-Chart:**

a comparison between individuals from the dataset that have diabetes and don’t have diabetes.

**Insights:**

It indicates that the majority have diabetes as 65.1% from the dataset have it comparing to the minority of them that don’t have it as only 34%.

A blue and pink pie chart with text

Description automatically generated

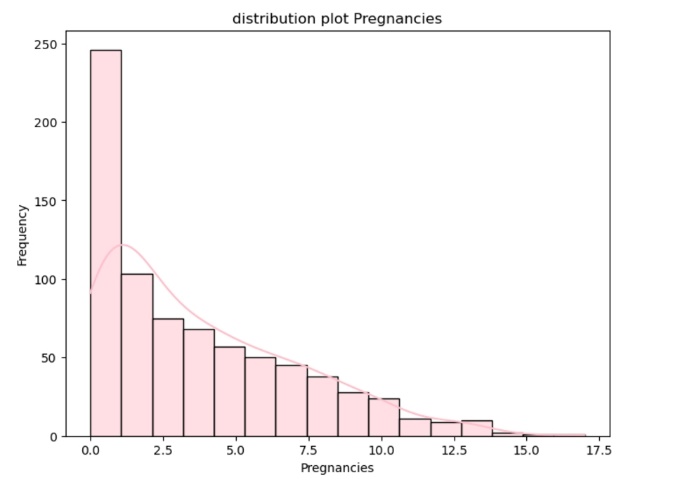
4)**Distribution Plot**

* **Pregnancy**

this plot shows the frequency distribution of the number of pregnancies among the individuals in the dataset.

**Insights:**

It is right skew and it indicates that most of the individuals in the dataset



* Age

this plot shows the frequency distribution of the age among the individuals in the dataset.

It is right skew , It indicates that most individuals in the dataset are around 20-30 years old.

A diagram of a distribution plot

Description automatically generated

this plot shows the frequency distribution of the skinThickness among the individuals in the dataset.

most individuals have a skin thickness of around 20 mm

A graph showing skin thickness

Description automatically generated

**5)Boxplot:** graphical representation of the distribution of a dataset based on min, max,

Q1,Q2, Q3.

**SkinThickness**: In this box plot, the lower quartile (Q1) = 0, the median is =23, and the upper quartile (Q3) is= 32, and the minimum is 0, the maximum is around 62.

A graph showing a box plot

Description automatically generated

**Diabetespedigreefunction**: .In this box plot, the lower quartile (Q1) = 0.24, (Q2) = 0.37, and the upper quartile (Q3) is 0.63, and the minimum is around 0.1, the maximum is around 1.2.

A diagram of a box plot

Description automatically generated

**BloodPressure:** In this box plot, the lower quartile (Q1) is =62.0

, the median(Q2) is 72.0 and the upper quartile (Q3) is=80and the minimum is around 0.1, the maximum is around 1.2.

A diagram of a box plot

Description automatically generated

**Data Statistics:**

**1)Pregnancies**

The mean:3.84

Median:3.0

Mode:1

standard deviation: 3.367

Variance:11

min:0

max:17

**2)Glucose:**

The mean:120.8

Median:117

Mode:100

standard deviation: 31.9

Variance: 1020.9

min:0

max:199

**3) Blood Pressure**

The mean:69.1

Median:72

Mode:70

standard deviation: 19.3

Variance: 374

min:0

max:122

**4)Skin Thickness**

The mean:20.5

Median:23

Mode:0

standard deviation: 15.9

Variance: 254,14

min:0

max:99

**5)Insulin**

The mean:79.7

Median:30.5

Mode:0

standard deviation: 115.1

Variance: 13263.88

min:0

max:846

**6)BMI**

The mean:31.99

Median:32.0

Mode:32.0

standard deviation: 7.8

Variance: 62.0

min:0

max:67.1

**7) DiabetesPedigreeFunction**

The mean:0.47

Median:0.37

Mode:0.254

standard deviation: 0.109

Variance: 0.109

min:0

max:2.42

**8)** **Age**

The mean:33.24

Median:29.0

Mode:22

standard deviation: 11.75

Variance: 138.122

min:21

max:81

**9)outcome**

The mean:0.34

Median:0

Mode:0

standard deviation: 0.47

Variance: 0.227

min:0

max:1