DIABETES PREDICTION REPORT

**INTRODUCTION**

For building the prediction model I have used the dataset available from [kraggle](https://www.kaggle.com), which is a widely used site for obtaining various datasets. For this one we particularly used this time is “Pima Indians Diabetes Database” which I have made available below.

* **Data Set:** [Pima Indians Diabetes Database](https://www.kaggle.com/uciml/pima-indians-diabetes-database)
* **Description**: The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The following is the description of the 6 variables:

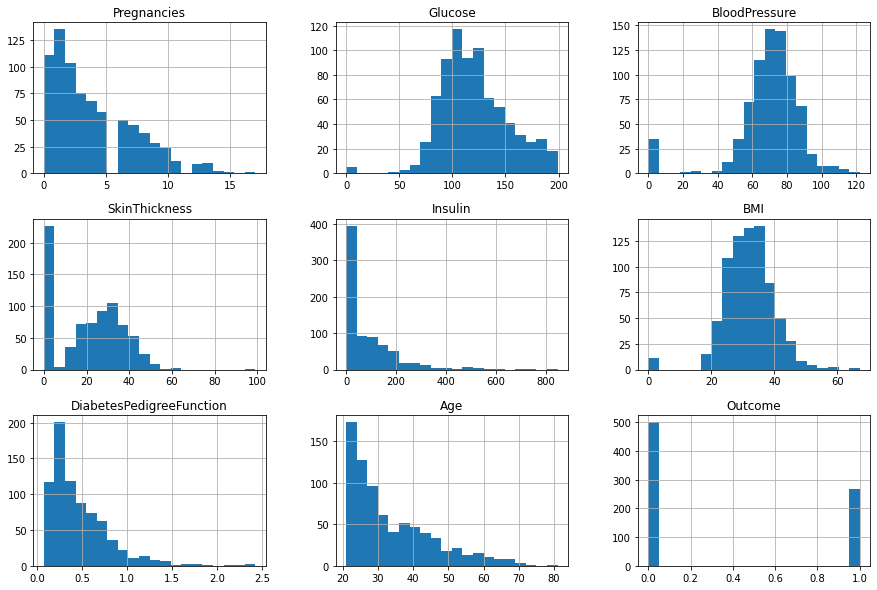
1. **Pregnancies**: It is no of pregnancies the patient has had which is in terms of numerical values.
2. **Glucose**: It is blood glucose levels.
3. **Blood** **Pressure**: Its Diastolic Blood pressure.
4. **Skin** **Thickness**: Triceps skin fold thickness in mm
5. **Insulin:** It’s the insulin levels in the patient blood.
6. **BMI**: Body Mass Index i.e., (weight in kg/(height in m)^2).
7. **Diabetes Pedigree Function.**
8. **Age:** In years.
9. **Outcome:** In 0 or 1 where 1 represents presents of diabetes and 0 shows there is no diabetes.

**DATA CLEANING AND FEATURE ENGINNERING**

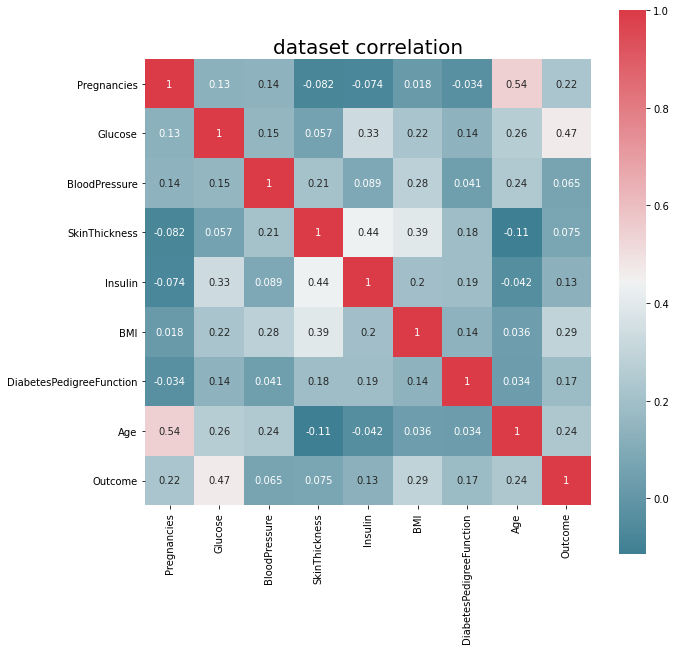
* In the dataset there are no null values so there is no need to worrier about cleaning the data.
* But the there are few places where the values are zero which need to be some values so we imputer the values with the help of simple imputer from sklearn.
* After checking the correlation matrix we can decide the columes which are less contributing towards the outcome. They are:
  + Pregnancies,
  + Blood pressure,
  + Skin Thickness,
* After separating predictor variables and outcome we use feature scaling on the predicator variables to ensure they remain in fixed range of values(-1 to 1).

**INSIGHTS**

**Plot 1**

****

**Plot 2**



**Plot 3**

****