**CapstoneProject -HealthCare Case Study**

**Project Report**

**Aparna Nair**

**aparna.nair.vijay@gmail.com**

**Business Scenario**

* NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.
* The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
* Build a model to accurately predict whether the patients in the dataset have diabetes or not.

**DATASET DESCRIPTION**

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Pregnancies**: Number of times pregnant

**Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test

**BloodPressure**: Diastolic blood pressure (mm Hg)

**SkinThickness**: Triceps skin fold thickness (mm)

**Insulin**: 2-Hour serum insulin (mu U/ml)

**BMI**: Body mass index (weight in kg/(height in m)^2)

**DiabetesPedigreeFunction**: Diabetes pedigree function

**Age**: Age (years)

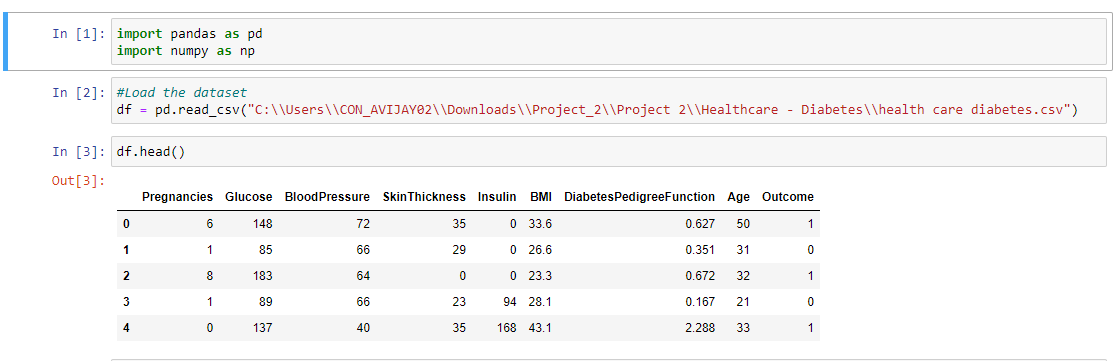
**Outcome**: Class variable (0 or 1) 268 of 768 are 1, the others are 0

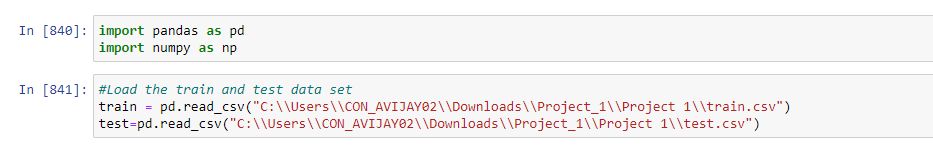
**Statistical algorithm execution – Python code and outputs**

1. Objective1) Perform descriptive analysis. It is very important to understand the variables and corresponding values. We need to think through - Can minimum value of below listed columns be zero (0)? On these columns, a value of zero does not make sense and thus indicates missing value.

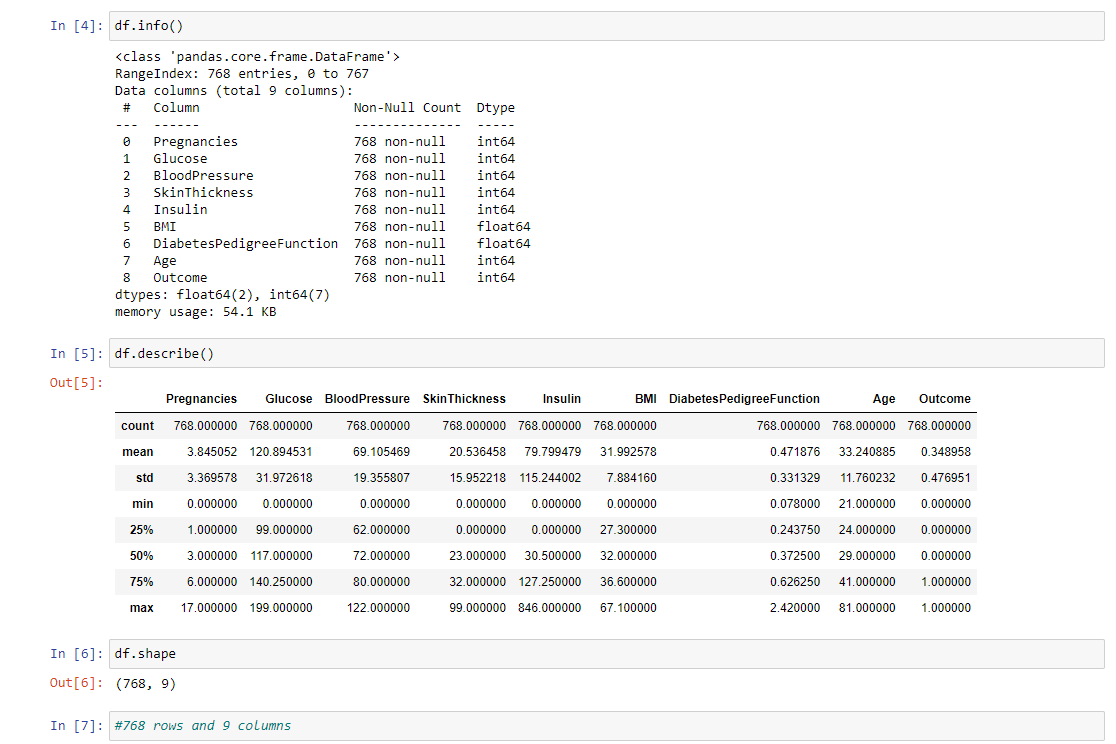
* Glucose
* BloodPressure
* SkinThickness
* Insulin
* BMI

Importing the data





Exploring the dataset

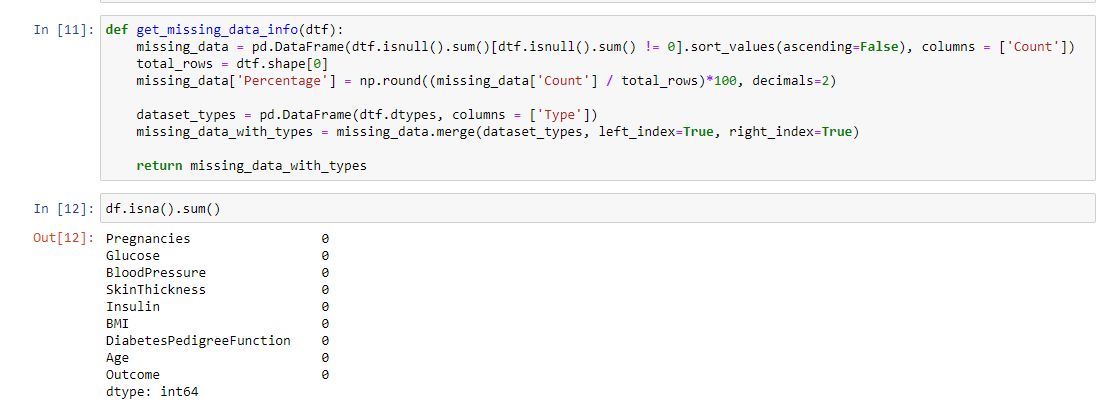


Data Cleaning

Checking if duplicate rows are present

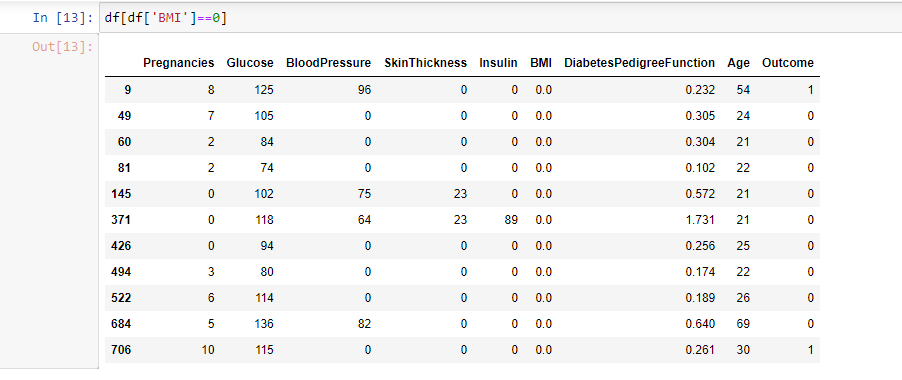


Checking for missing values

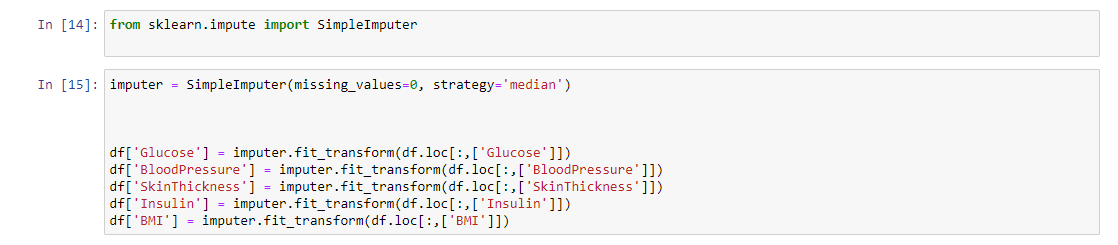


1. Objective 2)Visually explore these variables using histograms. Treat the missing values accordingly.

Checking the rows where BMI is zero which is not correct

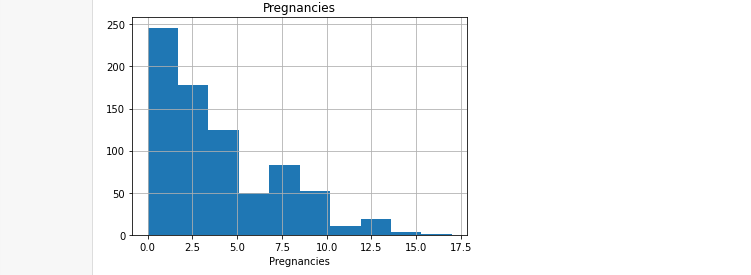


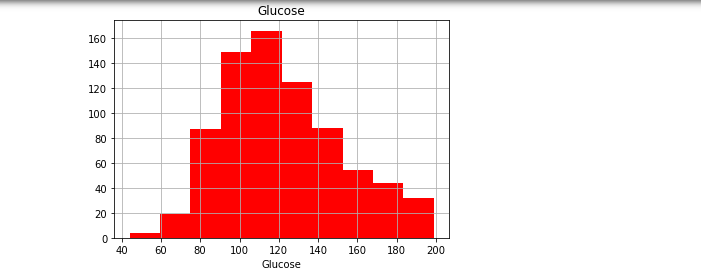
Using Simple imputer replacing the zero values for all variables with the median as it is incorrect to have zero for fields like BMI, Glucose, Blood pressure etc

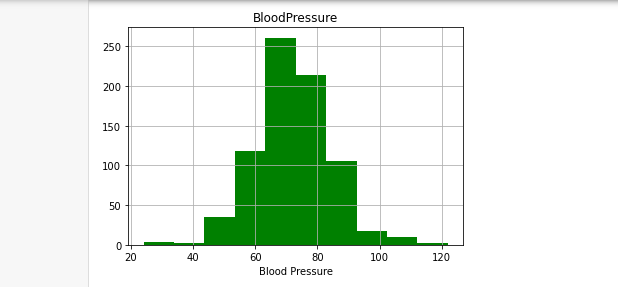


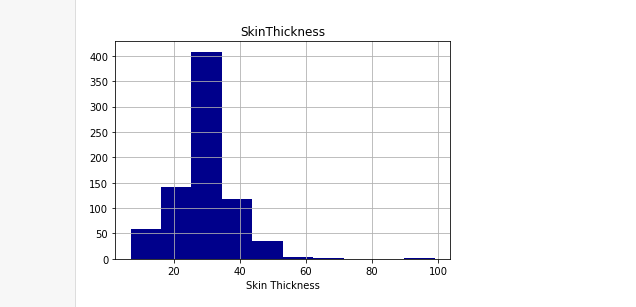
Plotting the histogram for all variables

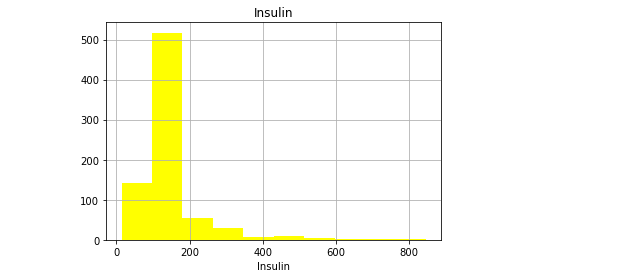


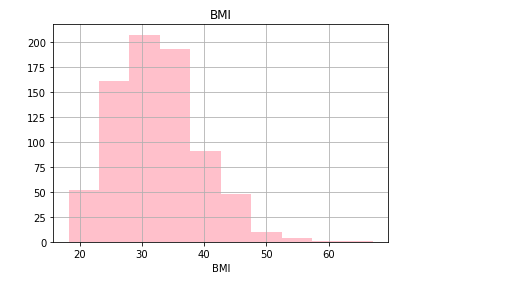


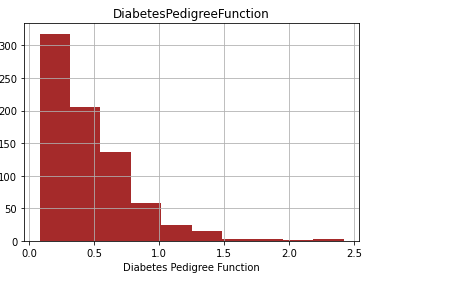


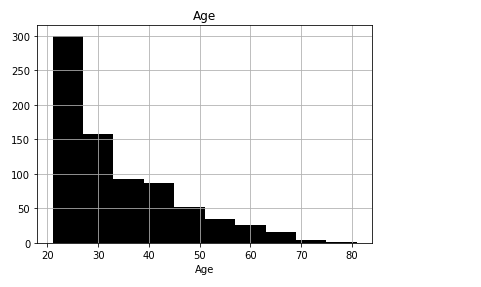


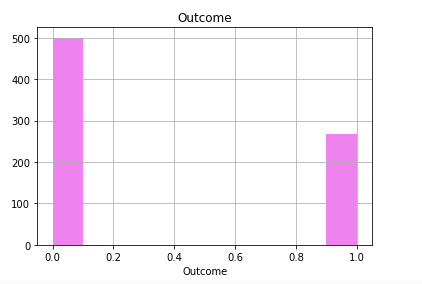










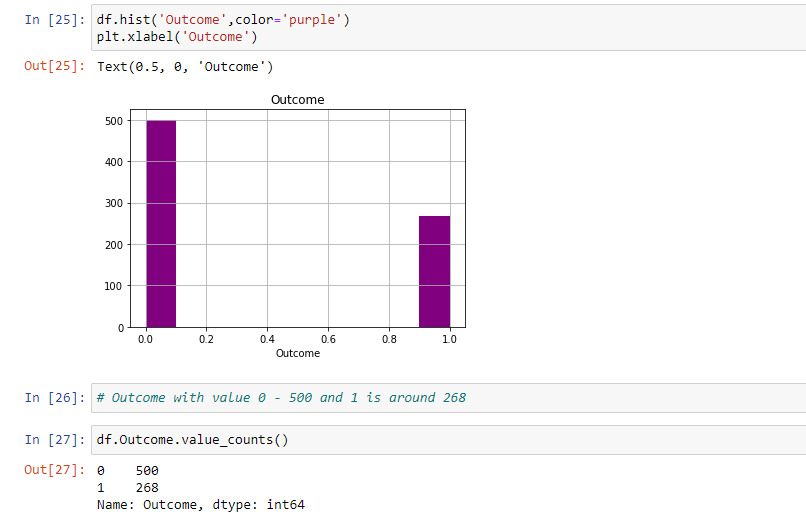


1. Objective 3 ) We observe integer as well as float data-type of variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.



We have 6 fields with float datatype and 3 fields with integer datatype

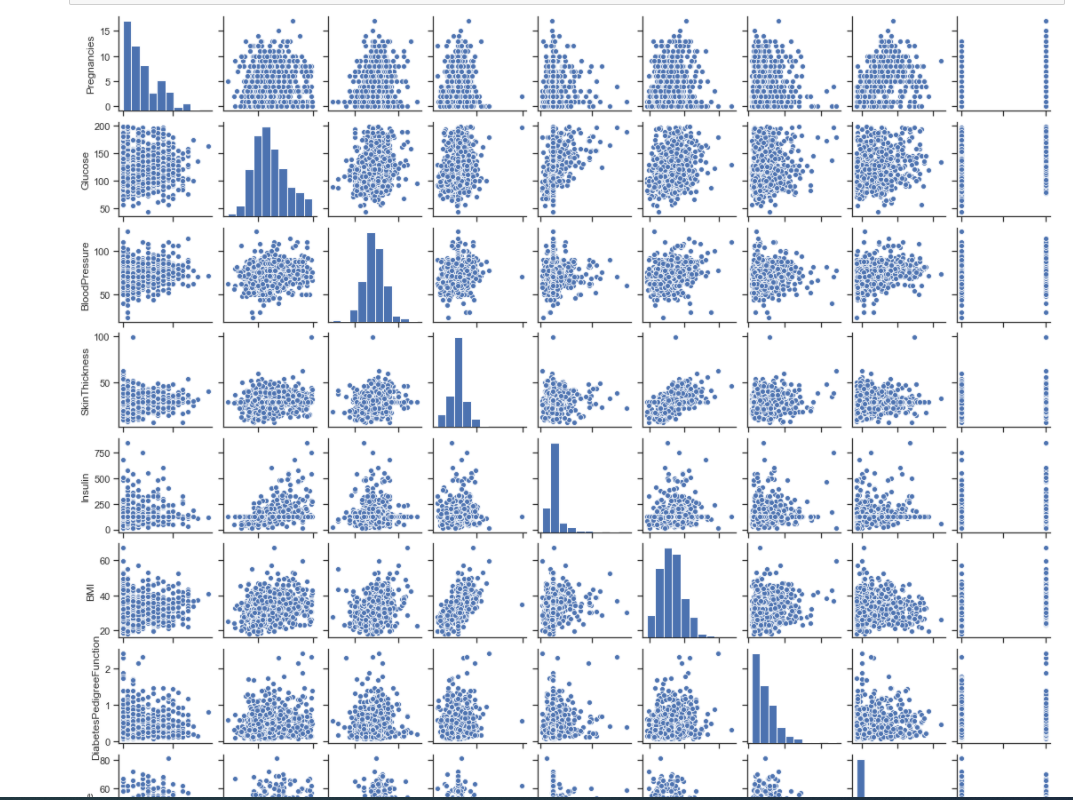
Objective 4) Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions.

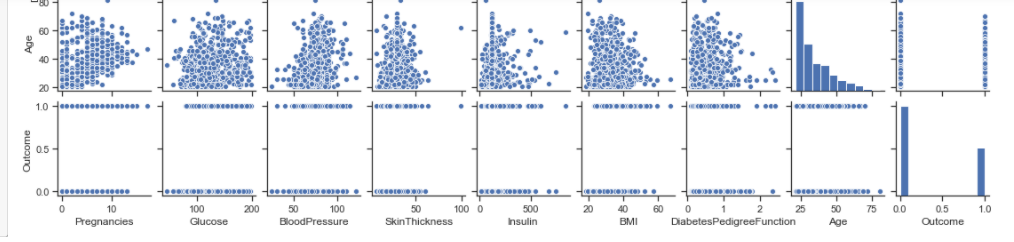


Outcome with value 0 - 500 and 1 is around 268

* Objective 5) Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

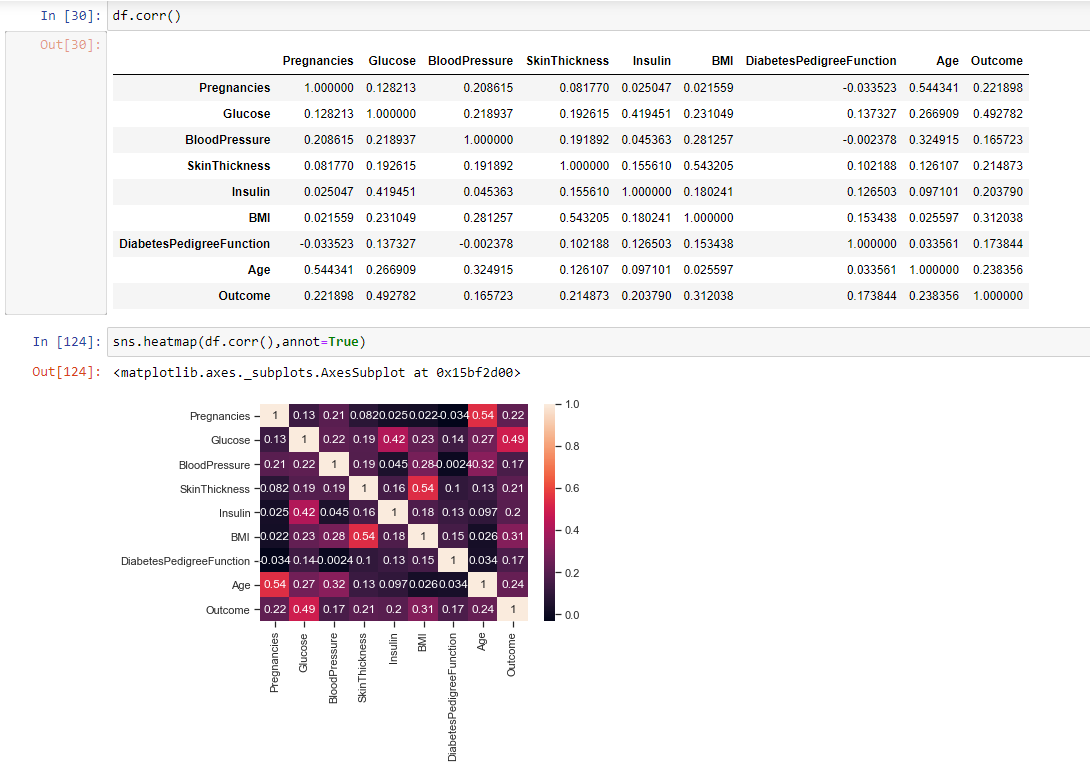






Skin Thickness and BMI, Age and Pregnancies, Glucose & Insulin is having a linear relationship

Objective 6 ) Perform correlation analysis. Visually explore it using a heat map.



Glucose is having a positive correlation with Insulin and outcome

Skin thickness is having a positive correlation with BMI

Age is having a positive correlation with Pregnancies

Objective 7 )Devise strategies for model building. It is important to decide the right validation framework. Express your thought process. Would Cross validation be useful in this scenario?

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Cross-Validation is a very powerful tool. It helps us better use our data, and it gives us much more information about our algorithm performance. In complex machine learning models, it's sometimes easy not pay enough attention and use the same data in different steps of the pipeline.

Trying different Model Validation techniques

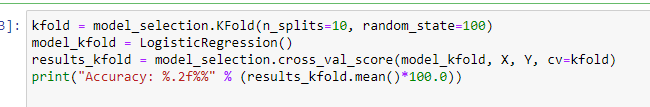
Holdout Validation



We can see that the accuracy for the model on the test data is approximately 73percent. The above technique is useful but it has pitfalls. The split is very important and, if it goes wrong, it can lead to model overfitting or underfitting the new data. This problem can be rectified using resampling methods, which repeat a calculation multiple times using randomly selected subsets of the complete data.

K-fold Cross-Validation

In k-fold cross-validation, the data is divided into k folds. The model is trained on k-1 folds with one fold held back for testing. This process gets repeated to ensure each fold of the dataset gets the chance to be the held back set. Once the process is completed, we can summarize the evaluation metric using the mean or/and the standard deviation.

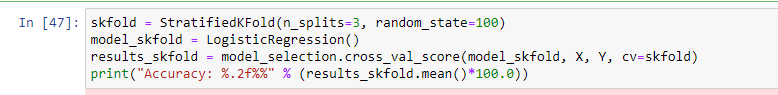




The mean accuracy for the model using k-fold cross-validation is 76.95 percent, which is better than the 74 percent we achieved in the holdout validation approach.

Stratified K-fold Cross-Validation

Stratified K-Fold approach is a variation of k-fold cross-validation that returns stratified folds, i.e., each set containing approximately the same ratio of target labels as the complete data

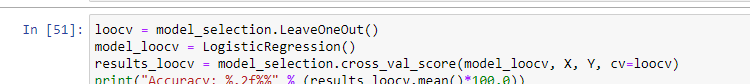




The mean accuracy for the model using stratified k-fold cross-validation is 76.96 percent.

Leave One Out Cross-Validation (LOOCV)

LOOCV is the cross-validation technique in which the size of the fold is “1” with “k” being set to the number of observations in the data. This variation is useful when the training data is of limited size and the number of parameters to be tested is not high.

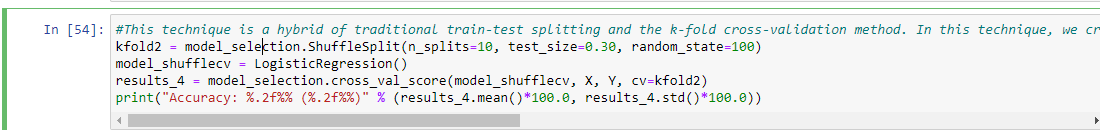




The mean accuracy for the model using stratified k-fold cross-validation is 76.69 percent.

Repeated Random Test-Train Splits

This technique is a hybrid of traditional train-test splitting and the k-fold cross-validation method. In this technique, we create random splits of the data in the training-test set manner and then repeat the process of splitting and evaluating the algorithm multiple times, just like the cross-validation method.





#The mean accuracy result for the various techniques is summarised below:

#Holdout Validation Approach: Accuracy of 73.16%

#Kfold Cross-Validation: Mean Accuracy of 77.21%

#Stratified K-fold Cross-Validation: Mean Accuracy of 76.43%

#Leave One Out Cross-Validation: Mean Accuracy of 76.69%

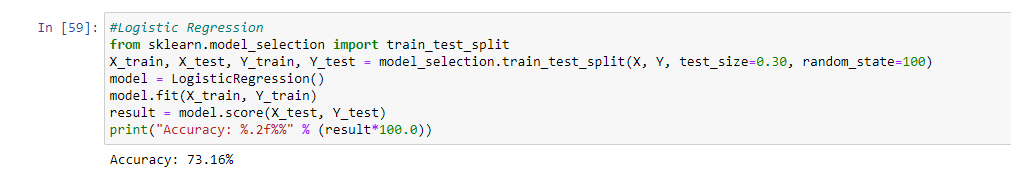
#Repeated Random Test-Train Splits: Mean Accuracy of 73.77%

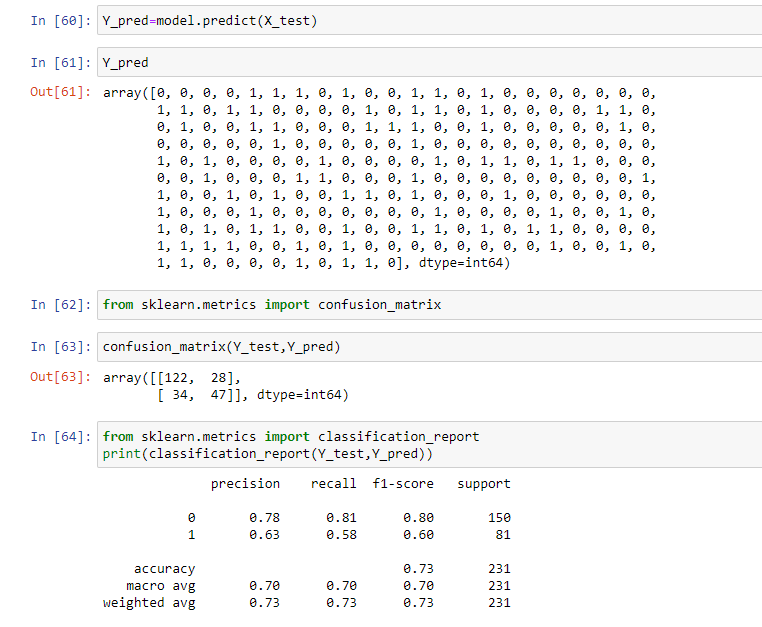
#We can conclude that the cross-validation technique improves the performance of the model and is a better model validation strategy.

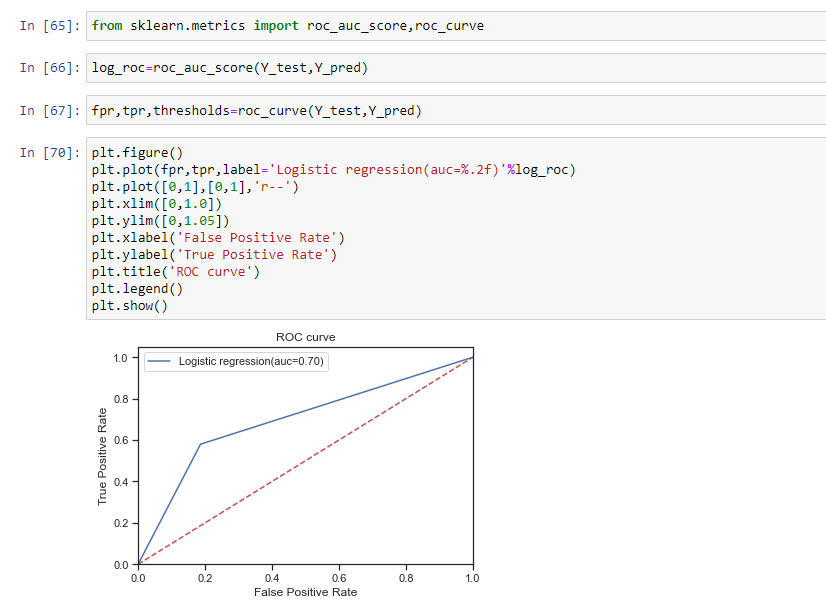
Objective8) Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN.

Objective9)Create a classification report by analysing sensitivity, specificity, AUC(ROC curve) etc. Please try to be as descriptive as possible to explain what values of these parameter you settled for? any why?

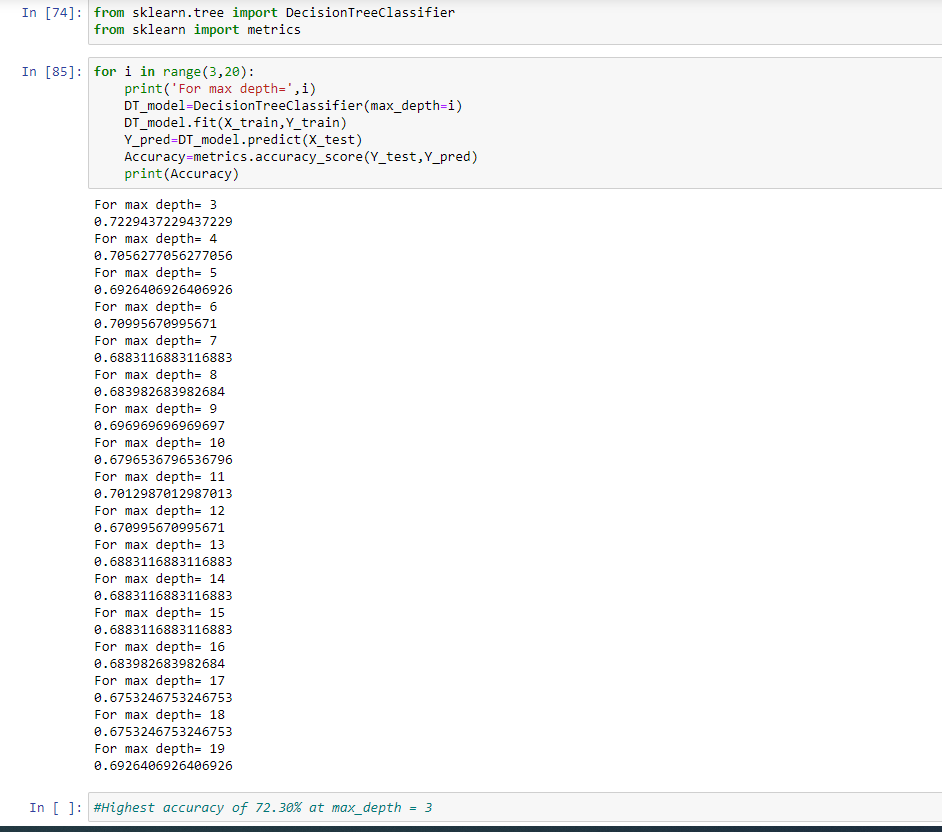
Logistic Regression

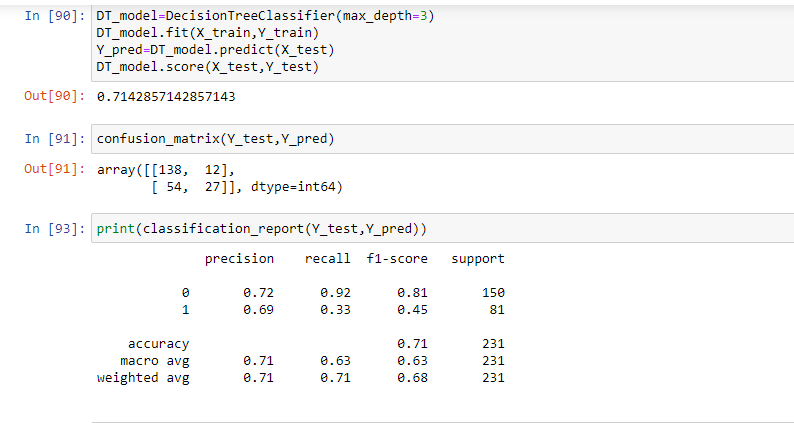


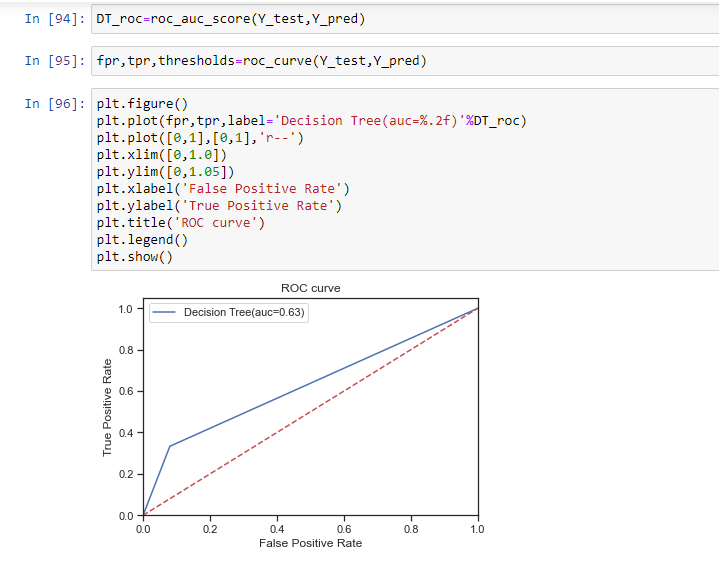




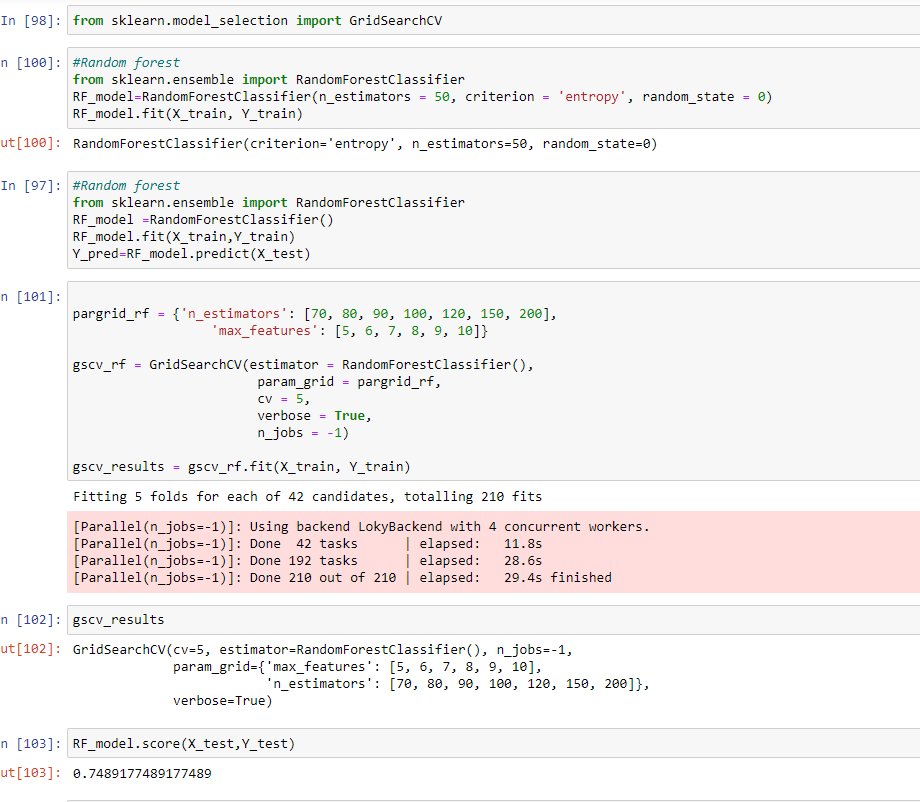
Decision Trees

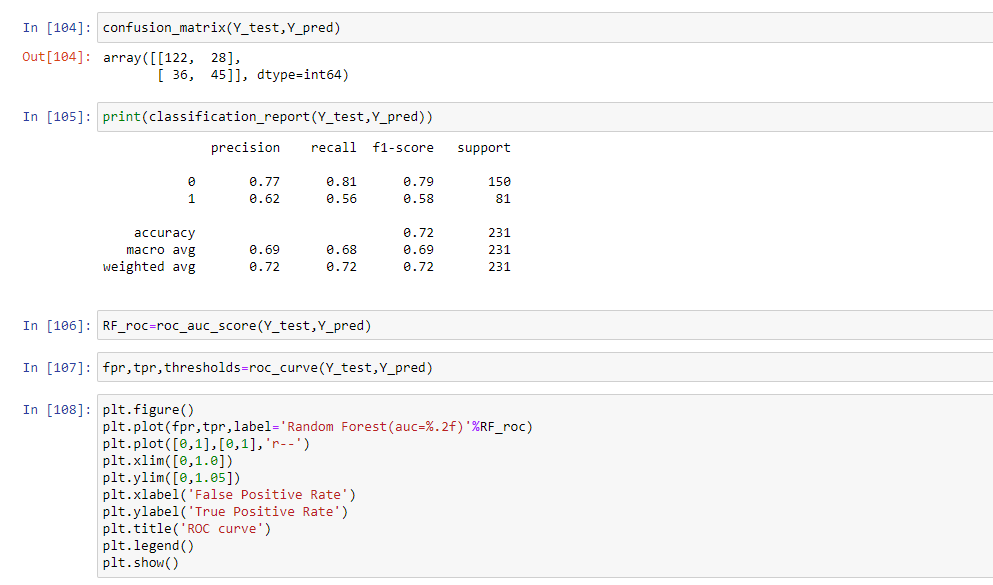


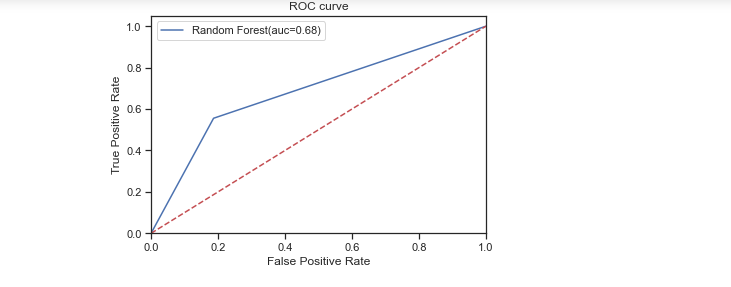




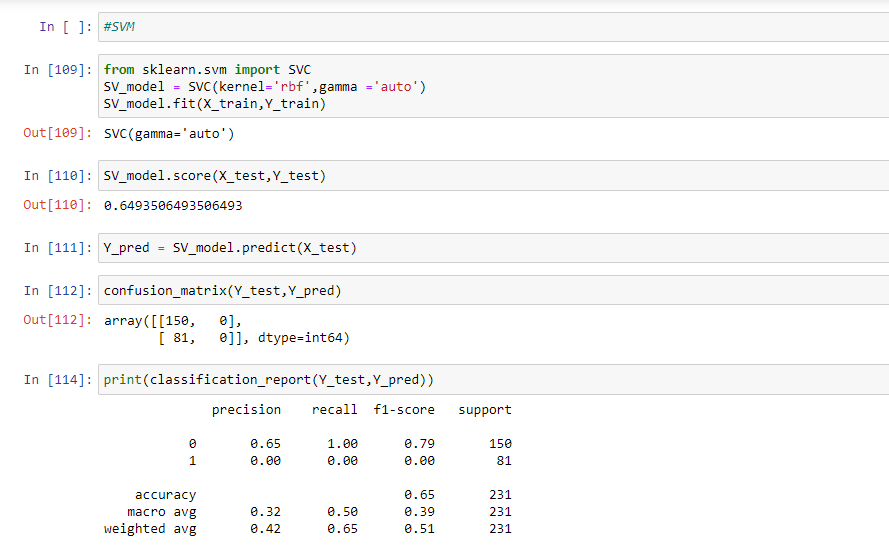
Random Forest



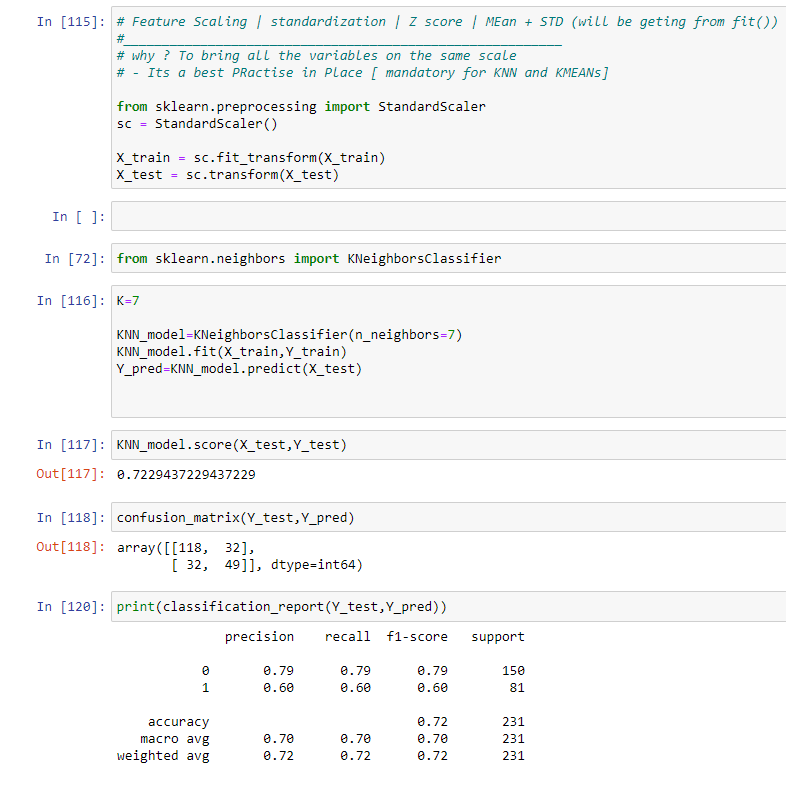


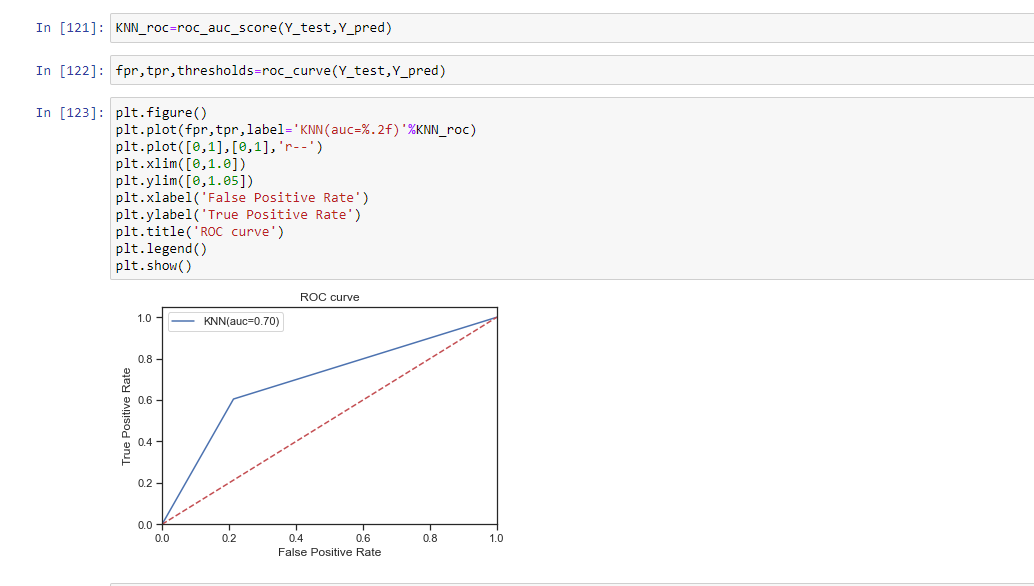


SVM



KNN

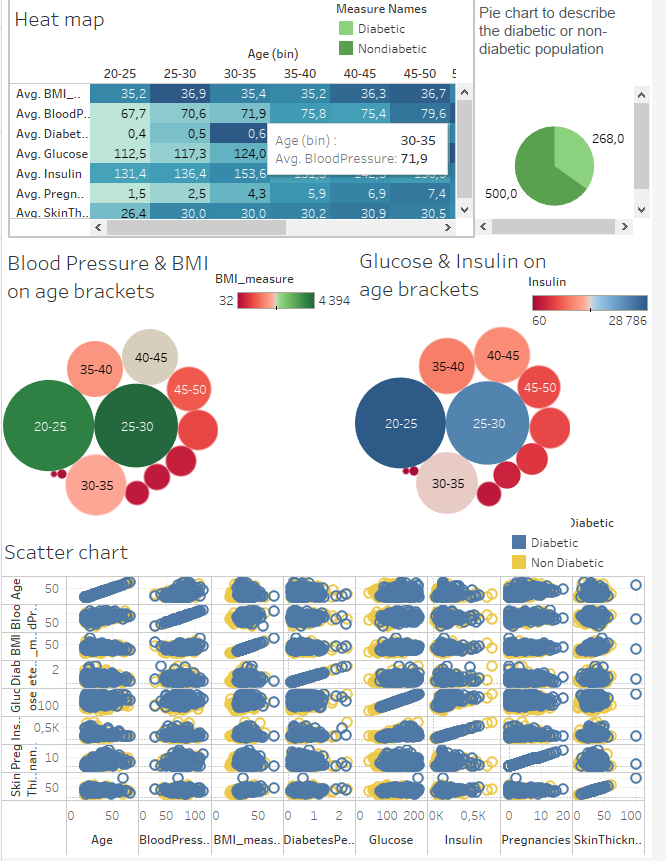




Objective 10) Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

1. Pie chart to describe the diabetic/non-diabetic population
2. Scatter charts between relevant variables to analyse the relationships
3. Histogram/frequency charts to analyse the distribution of the data
4. Heatmap of correlation analysis among the relevant variables

Create bins of Age values – 20-25, 25-30, 30-35 etc. and analyse different variables for these age brackets using a bubble chart



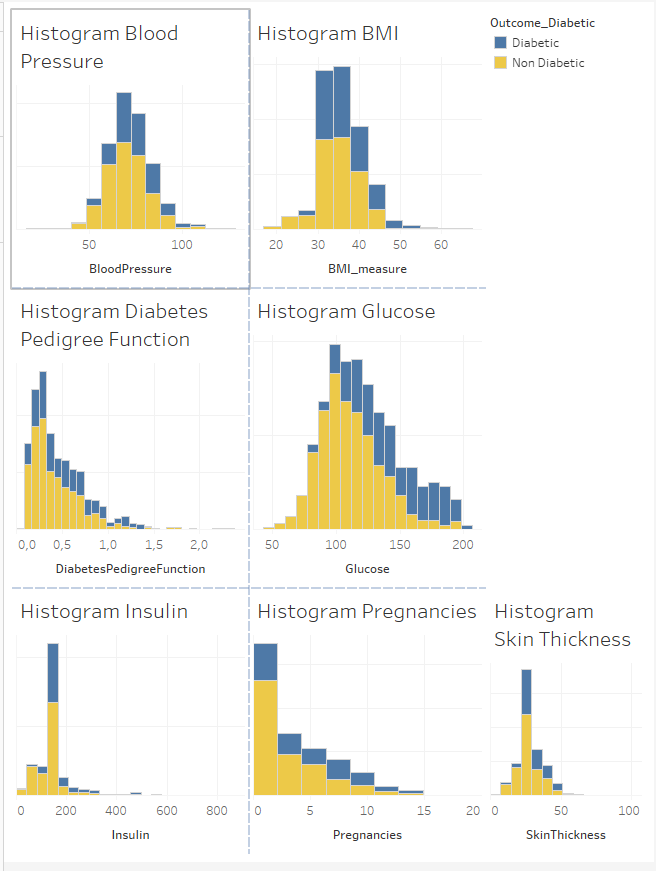


Tableau Public link to dashboard

https://public.tableau.com/app/profile/aparna6616/viz/Capstone\_Project2/Dashboard2