# **DATA SCIENCE CAPSTONE PROJECT**

# **Prediction Of Diabetics Using Classification Model**

**Problem Statement**

* 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.

**PLEASE CHECK THE JUPYTER NOTEBOOK PDF FOR DETAILED VIEW OF CODE AND EXPLANATION OF EACH TASK**

**Project Task: Week 1**

**Data Exploration:**

1. **Perform descriptive analysis. Understand the variables and their corresponding values.**

**On the columns below, a value of zero does not make sense and thus indicates missing value:**

• Glucose

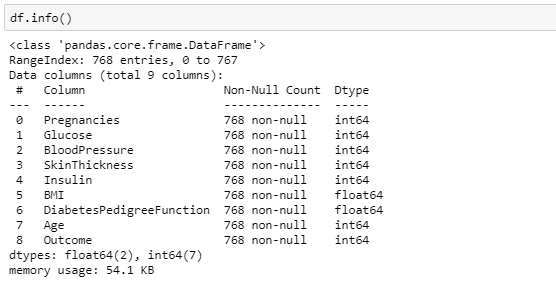
• BloodPressure

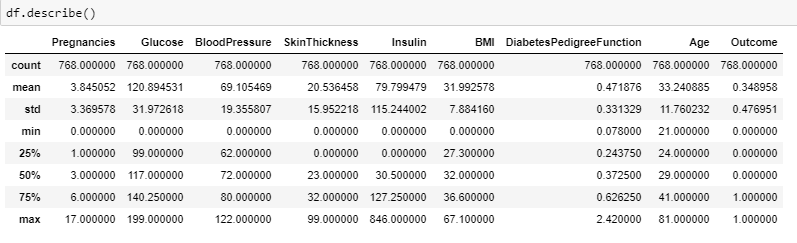
• SkinThickness

• Insulin

• BMI

**DESCRIPTIVE ANALYSIS**



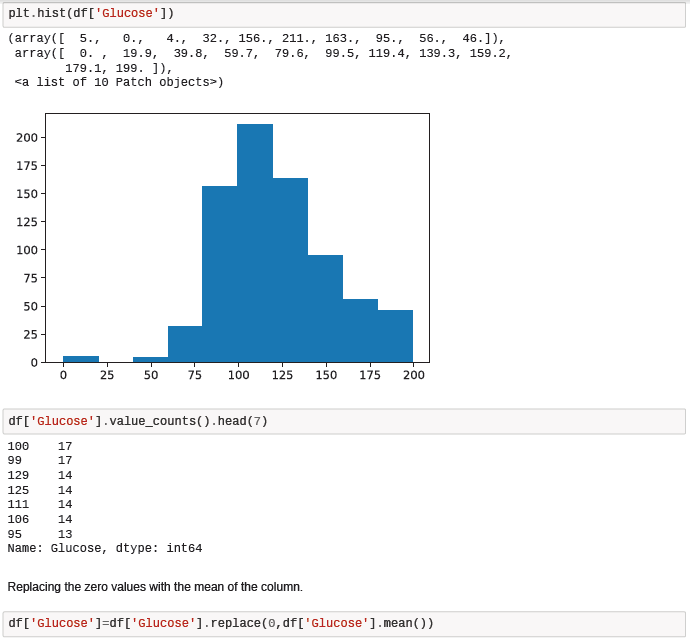


**Insights from Descriptive Analysis**

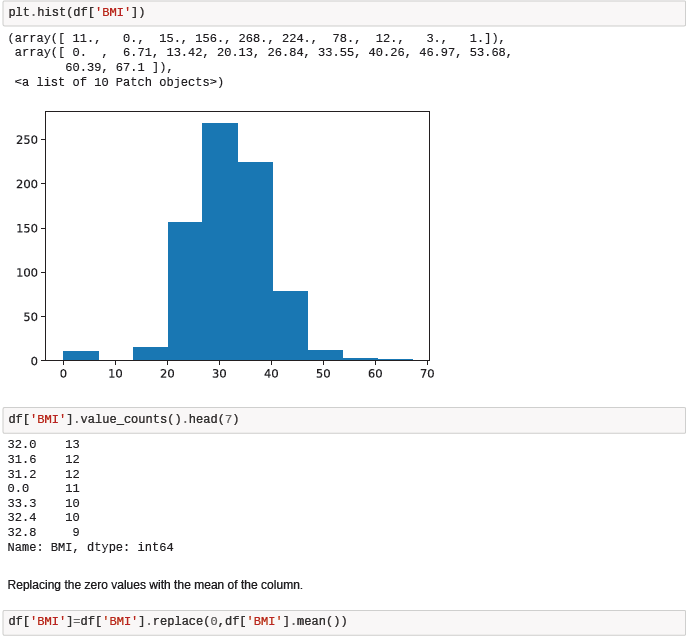
There are 768 observations of 9 variables. Independent variables are Pregnancies, Glucose, Blood Pressure, Insulin, BMI and Diabetes Pedigree Function. Age is Outcome Variable. Average Age of Patients are 33.24 with minimum being 21 and maximum 81. Avg. value of independent variables are Preg = 3.845052, Glucose = 120.894531, BP = 69.105469, ST=20.536458, Insulin = 79.799479, BMI = 31.992578, DPF = 0.471876.

1. **Visually explore these variables using histograms. Treat the missing values accordingly.**

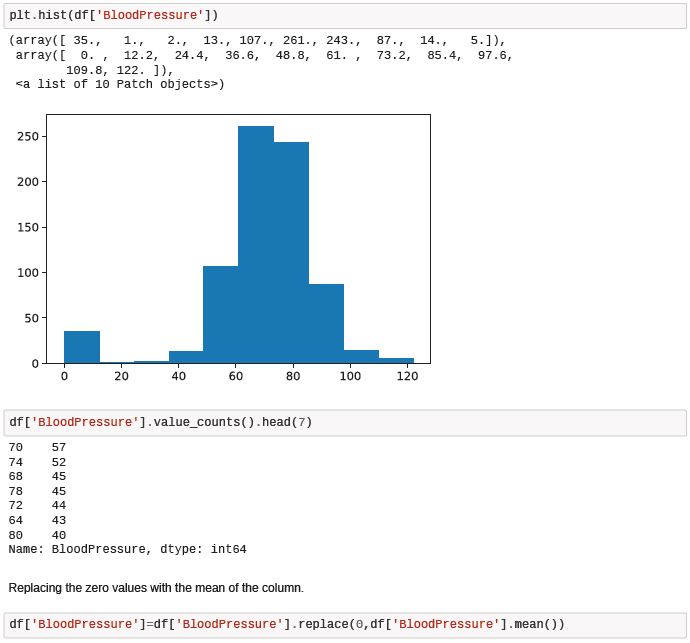
* **Glucose**



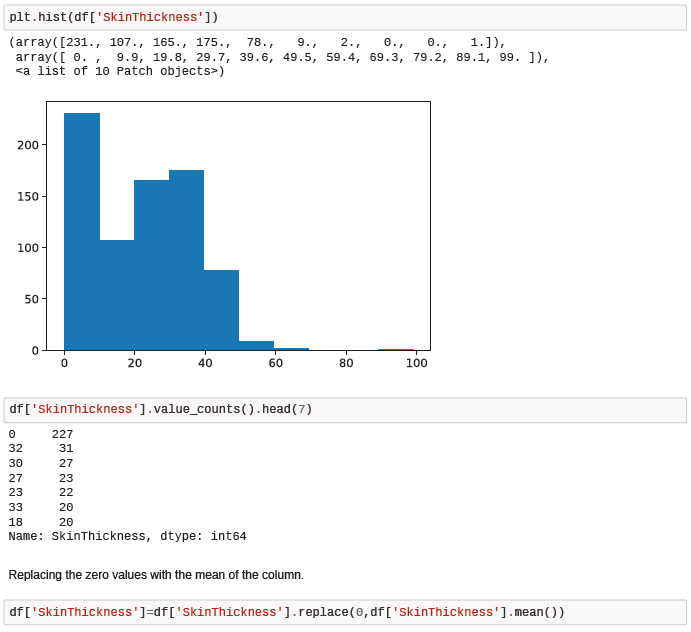
* **BMI**



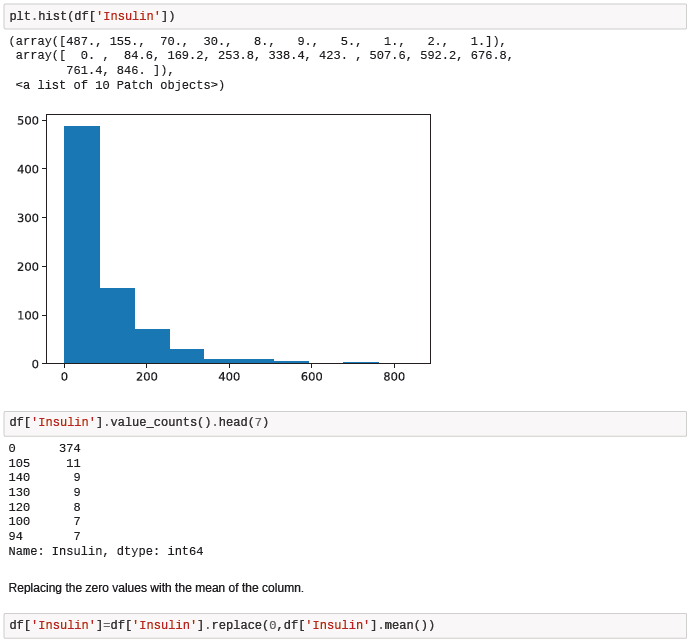
* **BloodPressure**



* **SkinThickness**



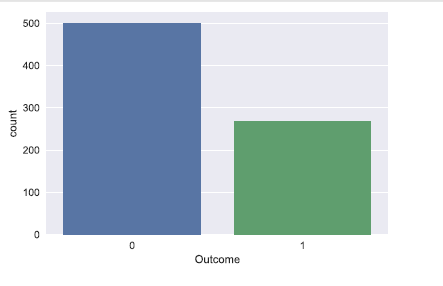
* **Insulin**



**Project Task: Week 2**

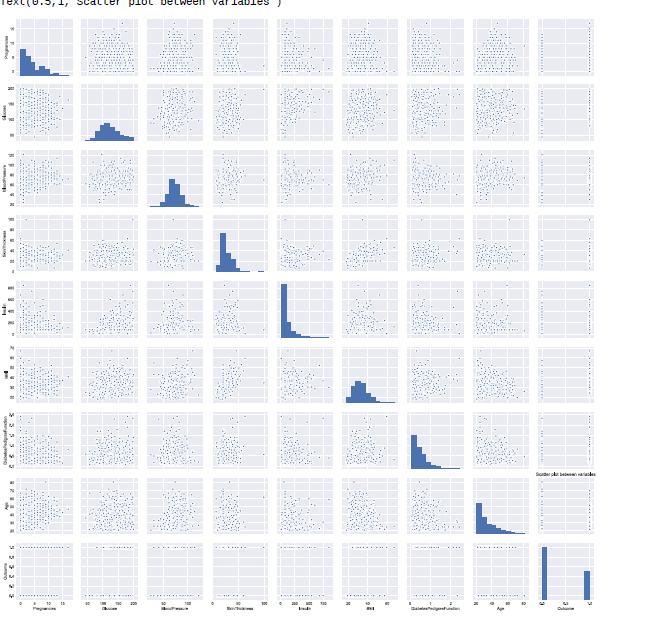
1. **Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.**

Checked the balance data by plotting the count of outcomes by their value. As both class is balanced, there is no need not to perform any sampling method to maintain the balance between both classes. So this data is directly used in training and testing. Meanwhile during Model Validation, we also need not worry about ROC Curve because data is not imbalanced.

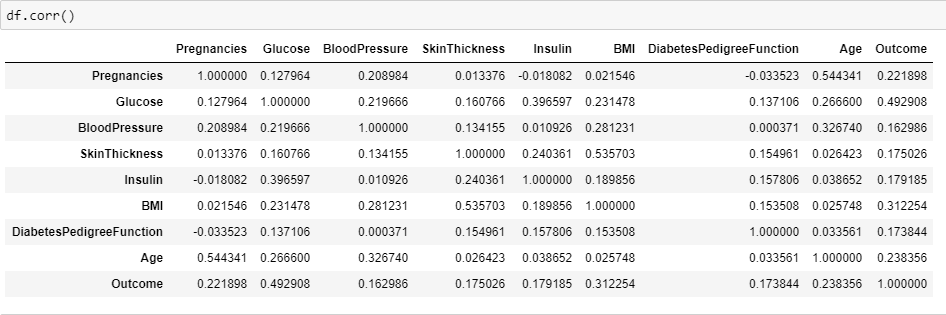


1. **Create scatter charts between the pair of variables to understand the relationships. Describe your findings.**

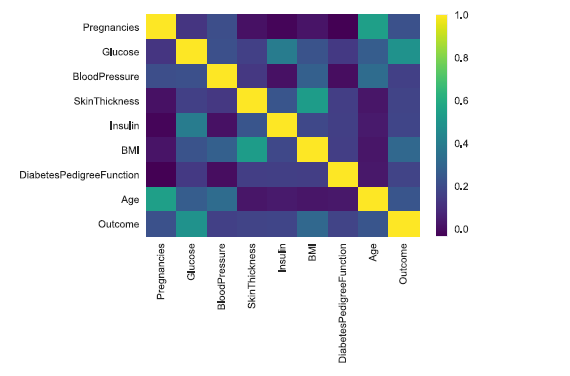
**The scatter plot showed no strong multi collinearity among features, but between skin thickness and BMI, Pregnancies and age it looks like there is small chance of positive correlation.**



1. **Perform correlation analysis. Visually explore it using a heat map**



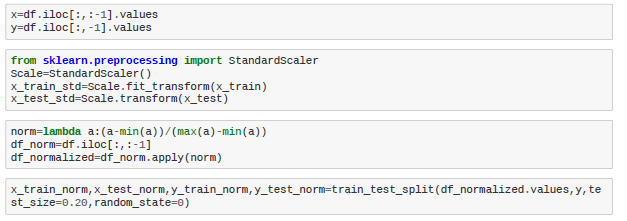
**Glucose and BMI has good impact on outcome. There is a strong positive correlation between BMI and Skinthickness or Pregnancies and age.**



**Project Task: Week 3**

1. **Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.**

Since the data has its outcome or target variable in categorical form (1 0r 0 = diabetic or non-diabetic) and categorical values use classification. The classification models used in this project are Logistic Regression, KNN, Random Forest. I will be also considering ROC Curve and ROC AUC Score to make sure Type 2 Error will not occur for Positive class, that is 1.



1. **Apply an appropriate classification algorithm to build a model. Compare various models**

**with the results from KNN algorithm.**

The classification models used in this project are Logistic Regression, KNN, Random Forest.

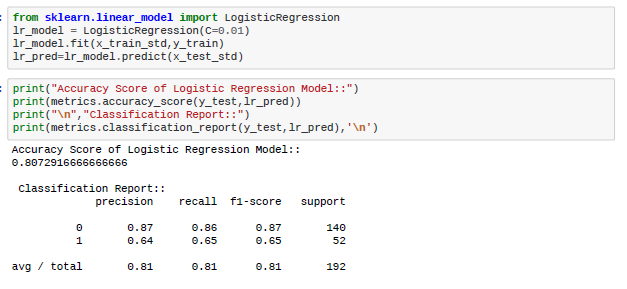
**Project Task: Week 4**

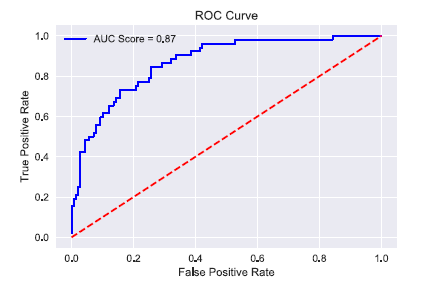
1. **Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.**

**Please be descriptive to explain what values of these parameter you have used.**

**Logistic Regression**:

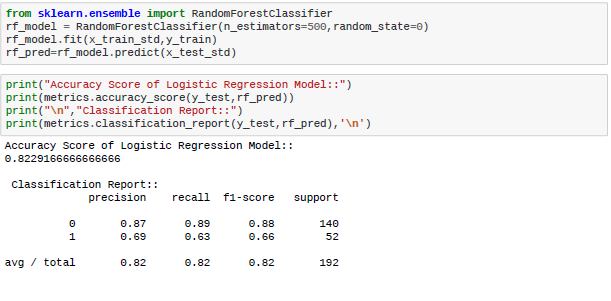
On applying Logistic Regression model, received an accuracy of 80% and Auc score is 87%. Classification report is applied and had f1 score and these scores are used to determine the better model.

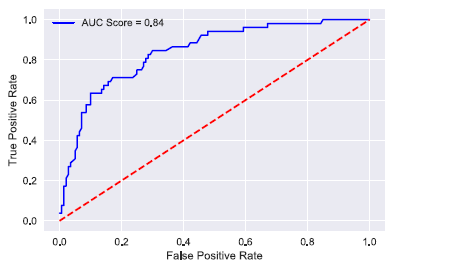




**Random Forest:**

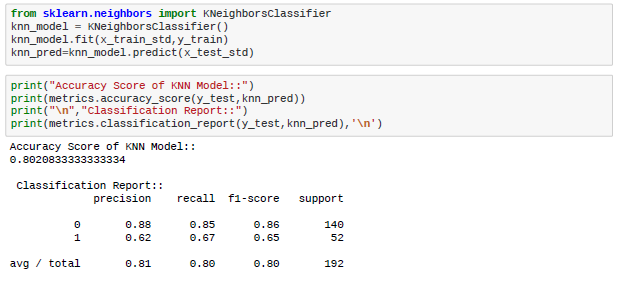
On applying the model, received an accuracy of 82% and auc score is 87%. Classification report is applied and had f1 score and these scores are used to determine the better model. Parameters used are no\_estmtors and

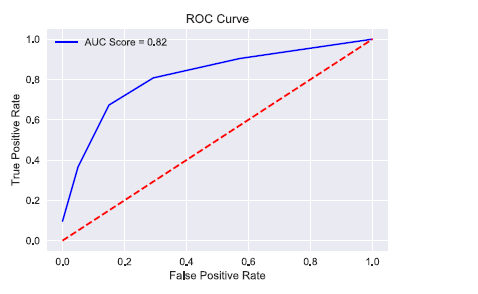




KNN:

On applying the model, received an accuracy of 80% and auc score is 82%. Classification report is applied and had f1 score and these scores are used to determine the better model





**Considering accuracy of the model, sensitivity, specificity and AUC (ROC curve), Random Forest model is the one that is better model compared to other two models.**

**Project Task: Week 4**

**Data Reporting:**

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

a. Pie chart to describe the diabetic or non-diabetic population

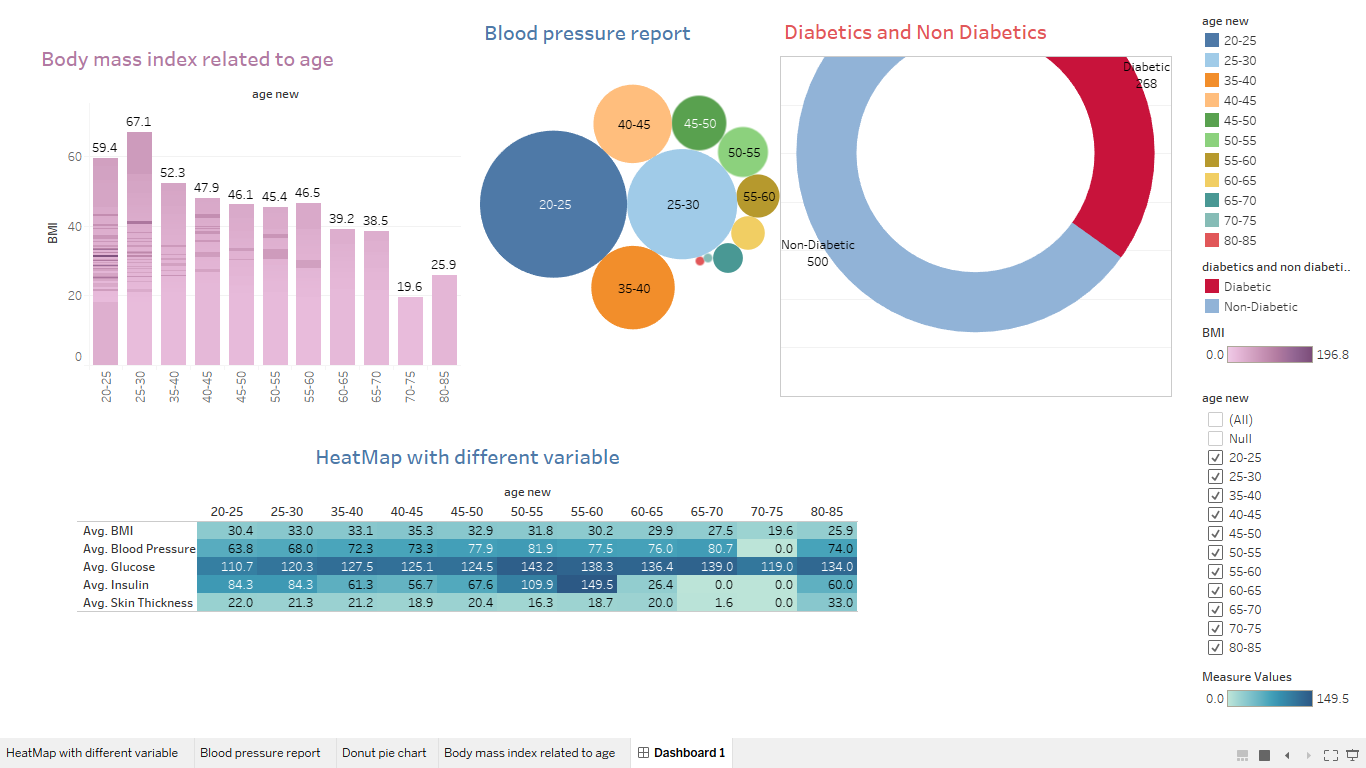
b. Scatter charts between relevant variables to analyze the relationships

c. Histogram or frequency charts to analyze the distribution of the data

d. Heatmap of correlation analysis among the relevant variables

e. Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

**Find the dashboard in the tableau in the link mentioned below:** [**https://public.tableau.com/profile/keerthana.sudina#!/vizhome/capstone-dashboard/Dashboard1?publish=yes**](https://public.tableau.com/profile/keerthana.sudina#!/vizhome/capstone-dashboard/Dashboard1?publish=yes)

****