ML ASSIGNMENT 1

TARANG PARIKH MT2019122
TARUN KUMAR RAI MT2019123
TUSHAR ANIL MASANE MT2019124

September 15, 2019

1 Introduction

Build a machine learning model to accurately classify whether or not the patients in the dataset have diabetes or not?

Some details about the dataset:

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. BloodPressure: Diastolic blood pressure (mm Hg)
4. SkinThickness: Triceps skinfold thickness (mm)
5. Insulin: 2-Hour serum insulin (mu U/ml)
6. BMI: Body mass index (weight in kg/(height in m squared))
7. Diabetes Pedigree Function: It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gives an idea of the hereditary risk one might have with the onset of diabetes mellitus

2 Data Exploration

On basic description, we find a lot of impractical values in the data. Negatives and Zeroes in columns of Glucose, SkinThickness, Insulin and BMI can be replaced with NULL.

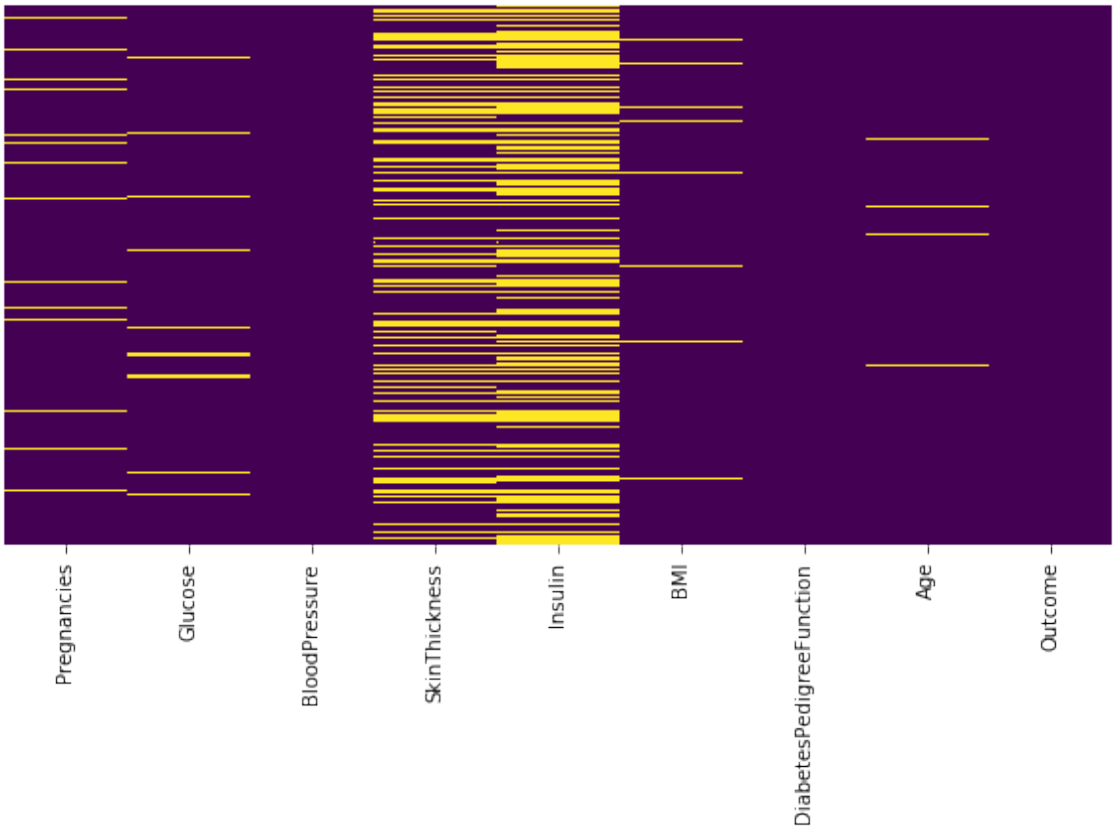


Figure 1: Visualising Missing Data

We find the correlation of every pair of features and the outcome variable.
In the generated heatmap as shown in Figure 2, brighter colors indicate more correlation. As we can see from the table and the heatmap, glucose levels, age, BMI and number of pregnancies all have significant correlation with the outcome variable.

1. The pregnancies are highly correlated with Age as seen from correlation heat map. Thus, missing values in Pregnancies can be filled by imputing them from average Pregnancy values depending on Age

2. SkinThickness is highly correlated with BMI. The missing values can be imputed with average SkinThickness based on different BMI levels

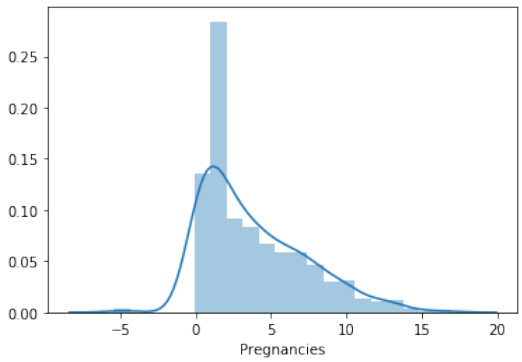
3. There is a high correlation of Insulin with Glucose. The missing values can be replaced based on different values of Glucose

Figure 2: Correlation Heatmap

3 Data Visualisation

We explore different columns and conclude various observations

1. Pregnancies



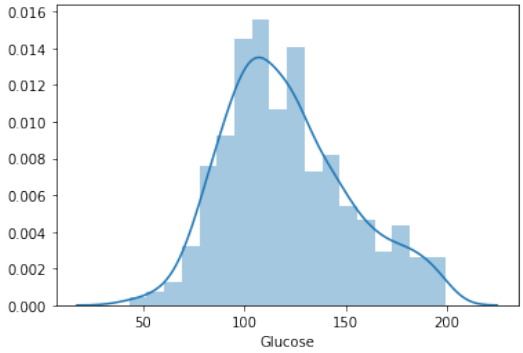
Distribution of Pregnancies

The distribution of pregnancies appears normal since the no of pregnancies is maximum for for lower values like 0 to 5 whereas number decreases as it approaches higher values

The data needs to be cleaned due to presence of negative values which is practically not possible

Thus, missing values in Pregnancies can be filled by imputing them from average Pregnancy values depending on Age

2. Glucose



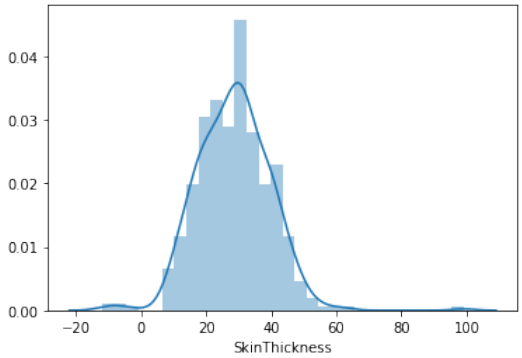
Distribution of Glucose

The distribution of Glucose values appears normal as the Glucose levels are in the valid range

The data needs to be cleaned for negatives

Since a very less number of values are missing, they can be imputed with means.

3. SkinThickness



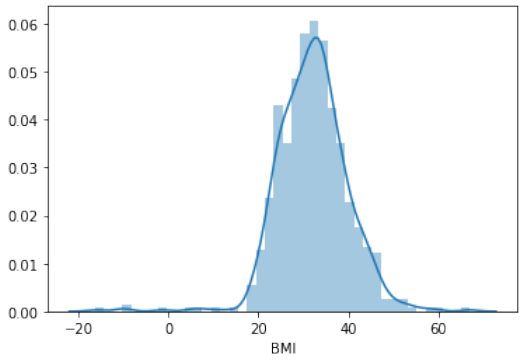
Distribution of SkinThickness

The data appears abnormal as large number of datapoints have value 0 which is practically not possible.

The data needs to be cleaned for negatives as well as zero

The missing values can be imputed with average SkinThickness based on different BMI levels

4. BMI



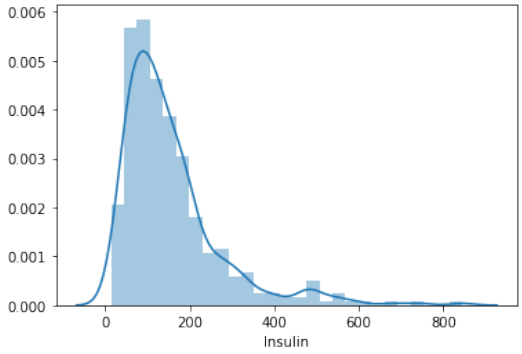
Distribution of BMI

The data appears normal as BMI is in valid range

The data needs to be cleaned for negatives and zero

Since very less values are missing, it can be imputed with mean value.

5. Insulin



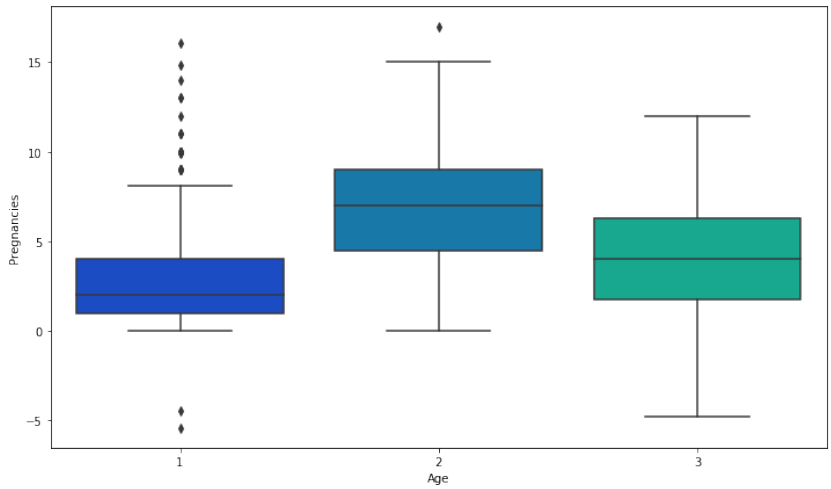
Distribution of Insulin

The distribution of Insulin appears normal

The missing values can be replaced based on different values of Glucose

4 Imputing Data

1. Imputing Pregnancies values using Age



Bucketting Age to find mean Pregnancies for a given range

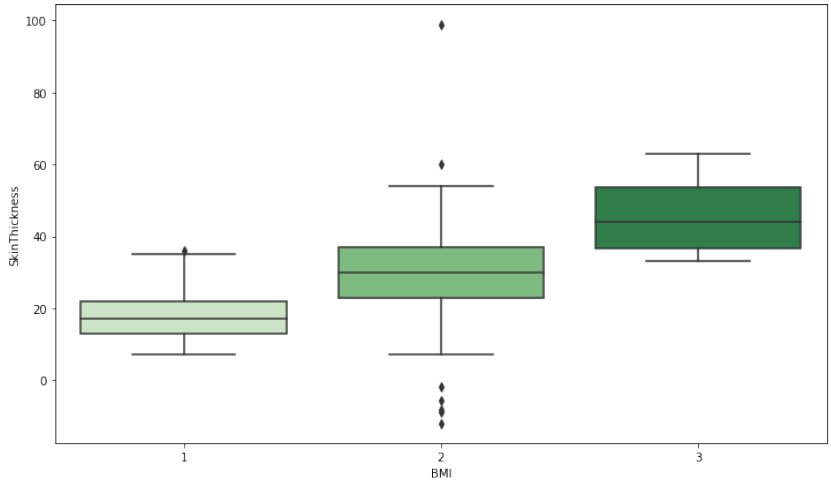
From the above boxplot, the average number of pregnancies

In age group 1 (Ages 20 to 40) is 2

In age group 2 (Ages 40 to 60) is 6

In age group 3 (Ages 60 to 100) is 4

2. Imputing SkinThickness values using BMI



Bucketting BMI values to find mean SkinThickness for a given range

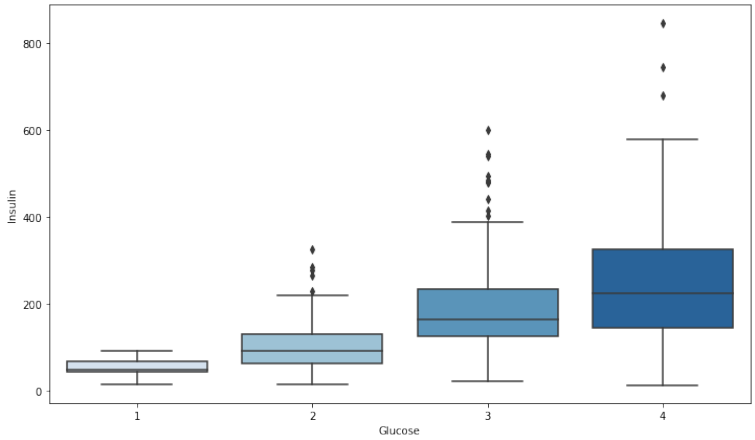
From the above boxplot, the average values of SkinThickness

In bmi class 1 (BMI values 0 to 25) is 33

In bmi class 2 (BMI values 25 to 50) is 38

In bmi class 3 (BMI values 50 to 75) is 40

3. Imputing Insulin values using Glucose



Bucketting Glucose values to find mean Insulin for a given range

From the above boxplot, the value of Insulin

in glucose class 1 (Glucose values 40 to 80) is 45

in glucose class 2 (Glucose values 80 to 120) is 100

in glucose class 3 (Glucose values 120 to 160) is 170

in glucose class 4 (Glucose values 160 to 200) is 200

5 Further Steps

5.1 Training and Predicting

The data is split in 70:30 ratio for training and testing the model.

As we have to predict whether the patients have diabetes or not, this is a classification problem.

There are only two classes "Yes" or "No" i.e. Binary Classification.

For Binary Classification, we can use Logistic Regression

5.2 Evaluation and Accuracy

Metrics like Confusion Matrix, Classification Report (Recoil and Precision) and Accuracy are used to evaluate the model

6 Conclusion

Thus, we were able to achieve 77 - 78 % of accuracy using Logistic Regression Model via Scikit Learn Library. We successfully were able to perform the following operations:

-Data Exploration.

-Data Visualization.

-Data Cleaning/Imputations.

-Model Training.

-Model Testing.

-Model Evaluation.