ML ASSIGNMENT 1

Lalitha Seelam (IMT2017027)
Ananya Appan (IMT2017004)
Swasti Shreya Mishra (IMT2017043)

September 2019

1 Data Preprocessing

In order to get an idea of how data is distributed, we first plotted graphs for each of the features. We have observed the following things for each feature

- **Glucose, BMI, BloodPressure :**
  It was observed that there were few negative values and few values were zero. So we considered the absolute values and converted all zeros to NaNs as their values can be computed later.

- **Pregnancies, Age :**
  Same data processing as above was applied but it doesn't make sense to have floating values for pregnancies and age. So all floating values were rounded.

- **SkinThickness, Insulin, DiabetesPedigreeFunction :**
  It was observed that there were few negative values and these were replaced with absolute

1.1 Missing Data

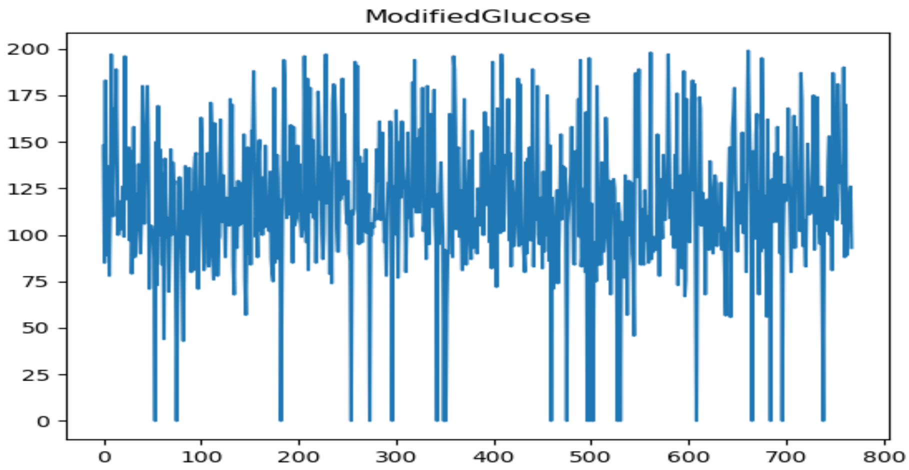
From the previous processing, we have removed inconsistent values. The average was computed ignoring all NaNs.It was computed as follows -

$$avg = \frac{total(feature)}{(numRows - numNA)}$$

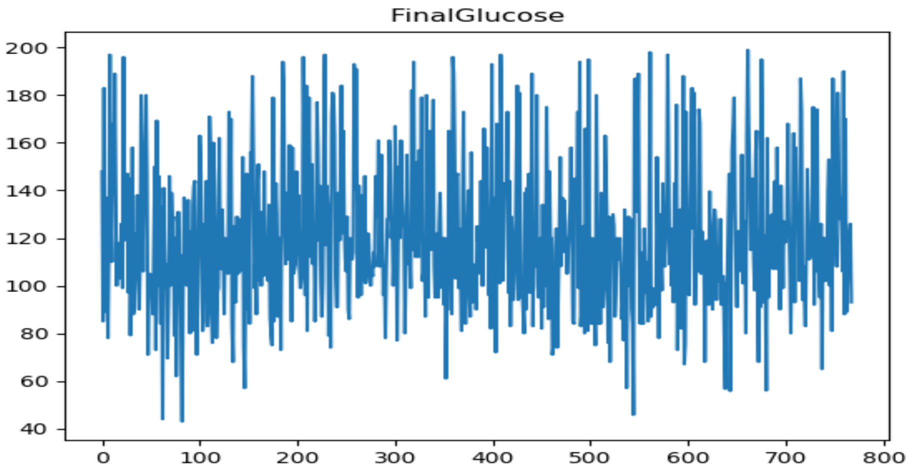
where ,
$total(feature)$ is the sum of values over all rows of a feature.
$numRows$ is the number of data points collected.
$numNA$ is the number of NaN values recorded for a particular feature.

1.1.1 Glucose

The following graph was plotted replacing all NaNs with zeros. It's clear from the graph that zeros as a value doesn't make sense and thus we replaced it with the average value.

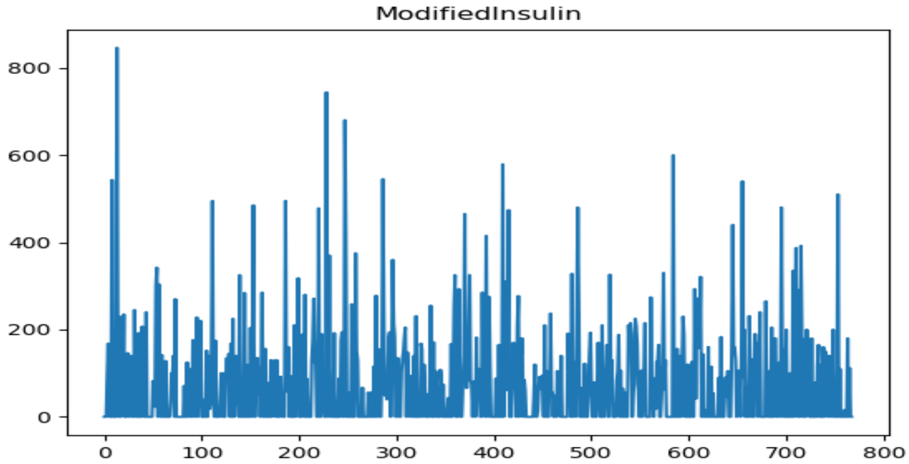


The below graph was plotted after replacing all zeros and NaNs with the average value.

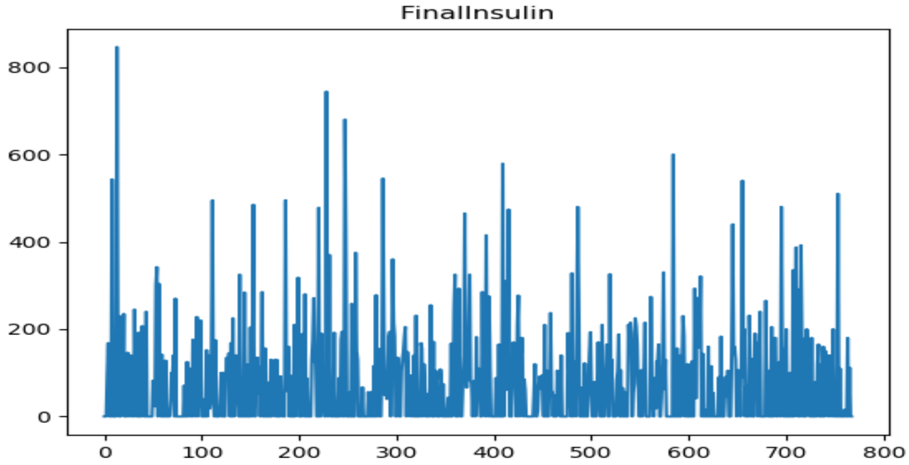


1.1.2 Insulin

The following graph was plotted replacing all NaNs with zeros.

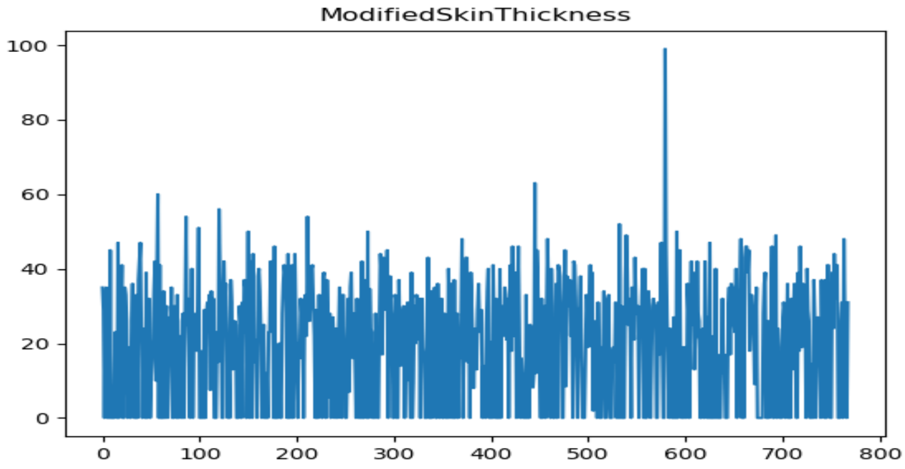


The below graph was plotted after replacing all NaNs with average value.

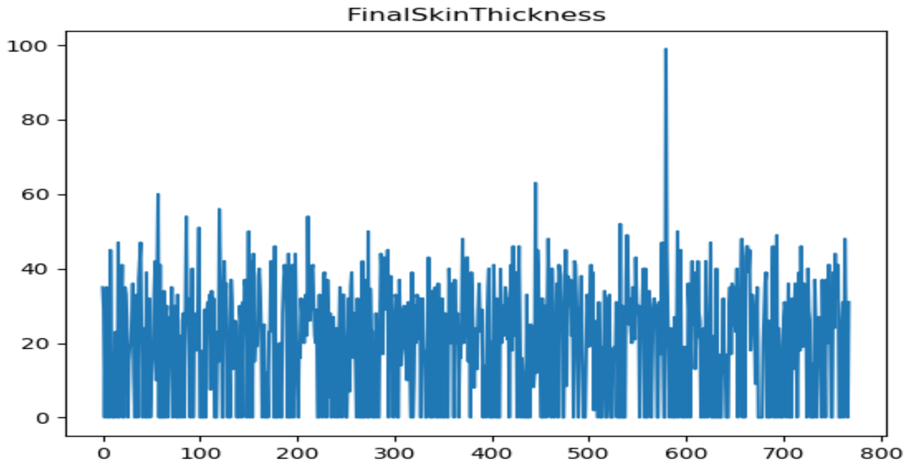


1.1.3 SkinThickness

The following graph was plotted replacing all NaNs with zeros.

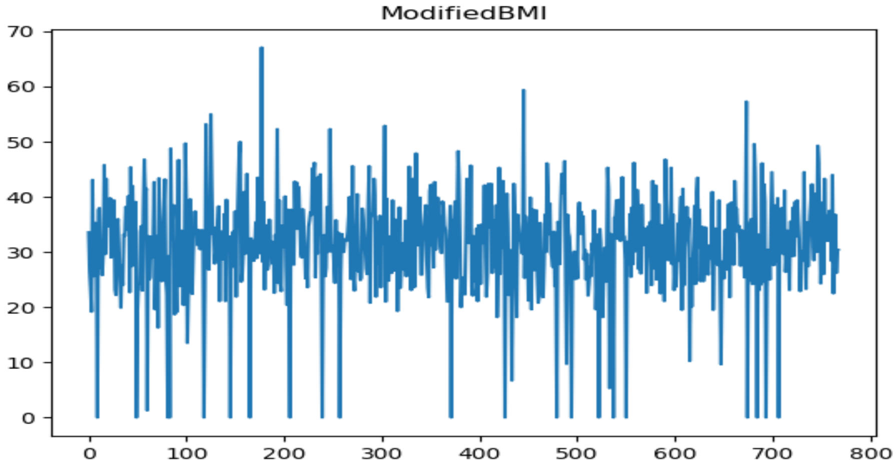


The below graph was plotted after replacing all NaNs with average value.

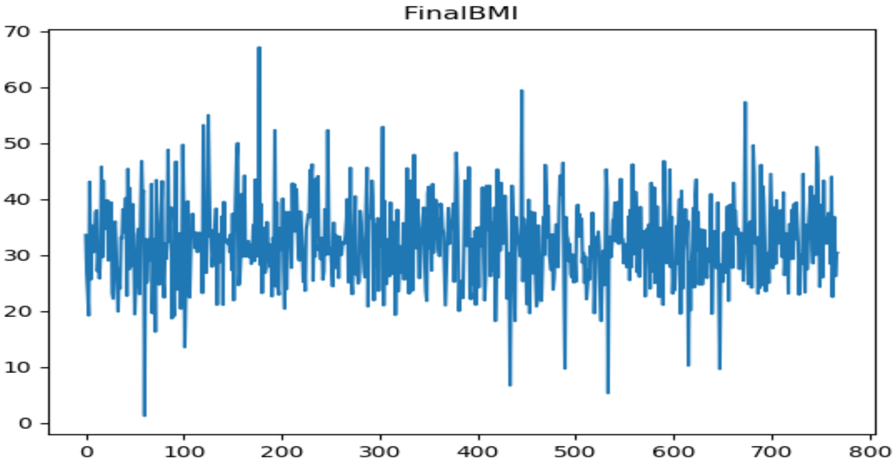


1.1.4 BMI

The following graph was plotted replacing all NaNs with zeros. It's clear from the graph that zeros as a value doesn't make sense and thus we replaced it with the average value.

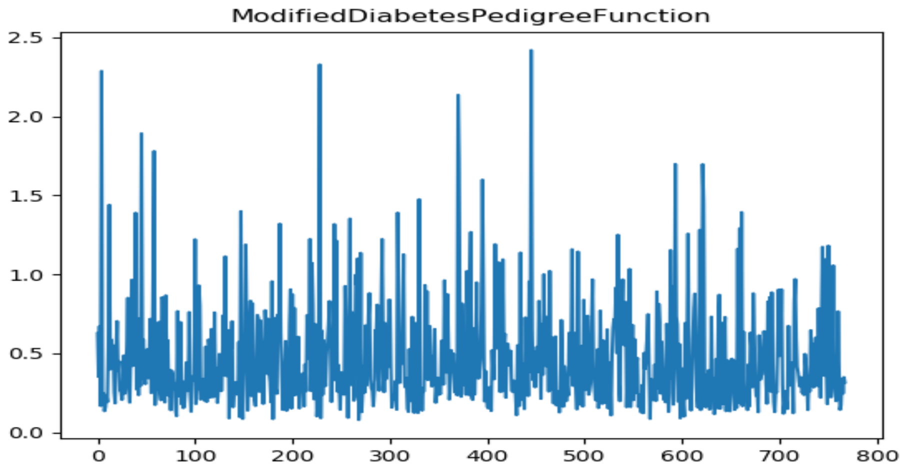


The below graph was plotted after replacing all zeros and NaNs with the average value.

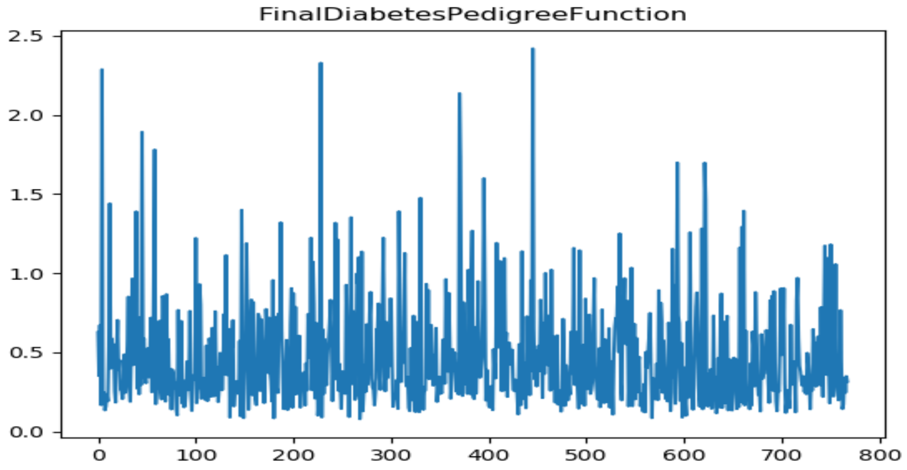


1.1.5 DiabetesPedigreeFunction

The following graph was plotted replacing all NaNs with zeros.

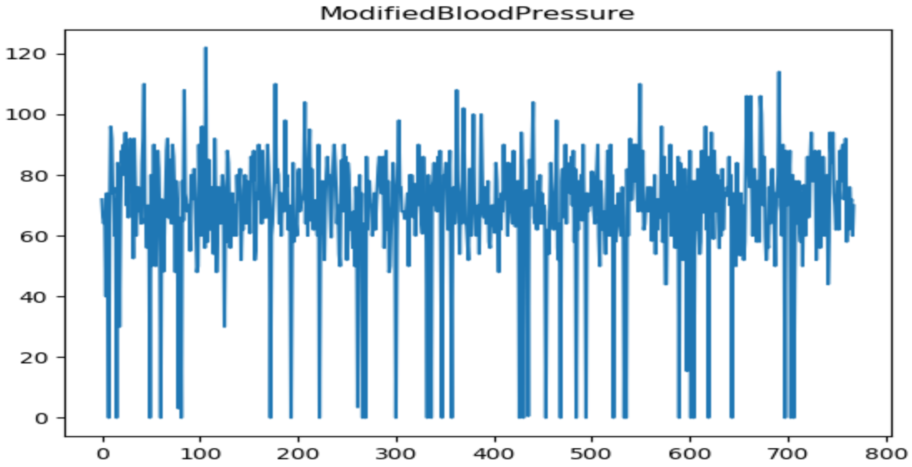


The below graph was plotted after replacing all NaNs with average value.

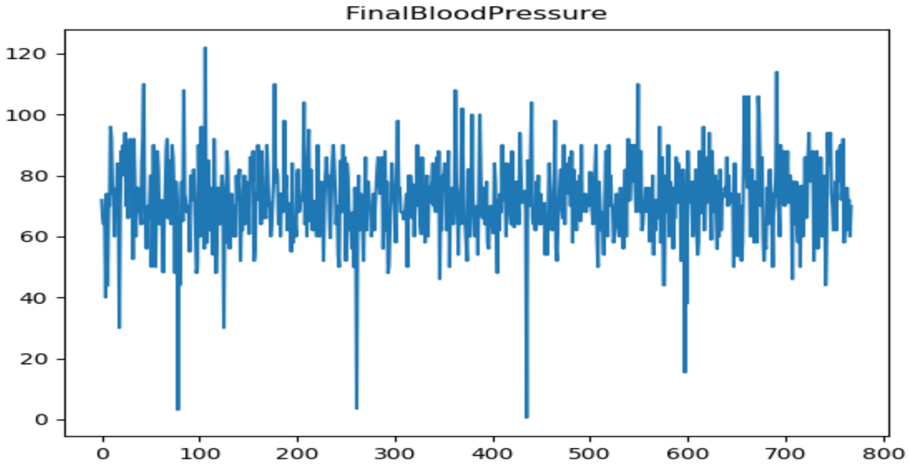


1.1.6 BloodPressure

The following graph was plotted replacing all NaNs with zeros. It's clear from the graph that zeros as a value doesn't make sense and thus we replaced it with the average value.

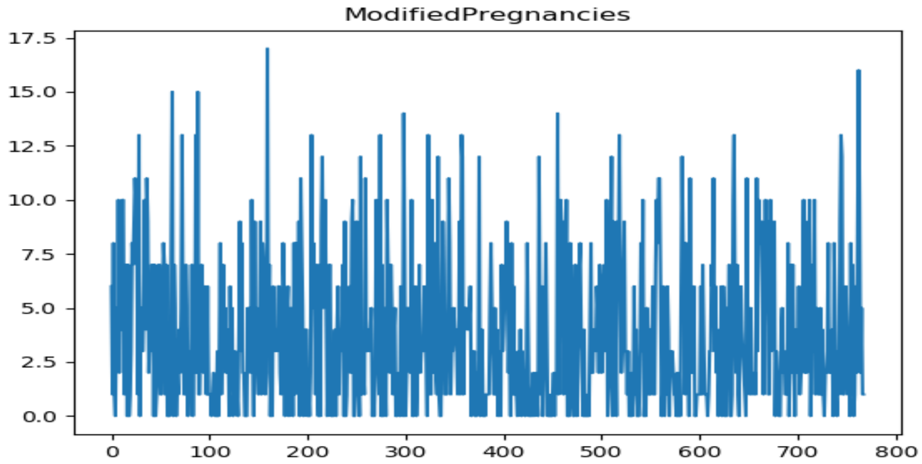


The below graph was plotted after replacing all zeros and NaNs with the average value.

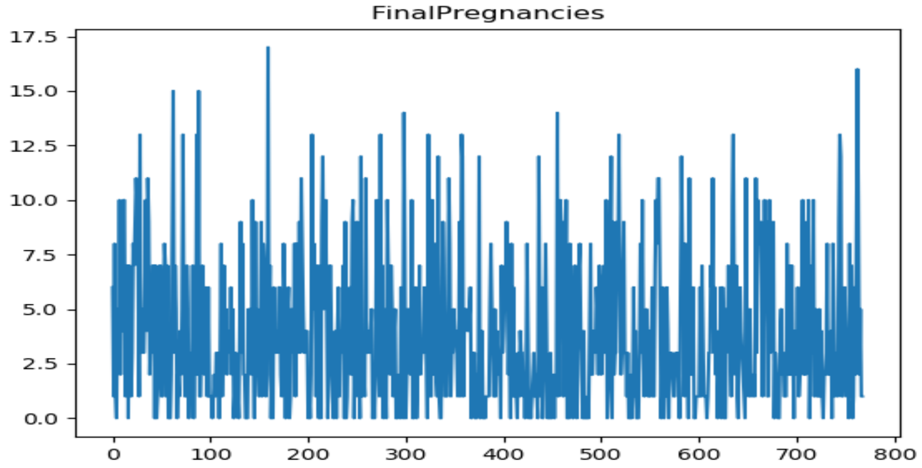


1.1.7 Pregnancies

The following graph was plotted replacing all NaNs with zeros.

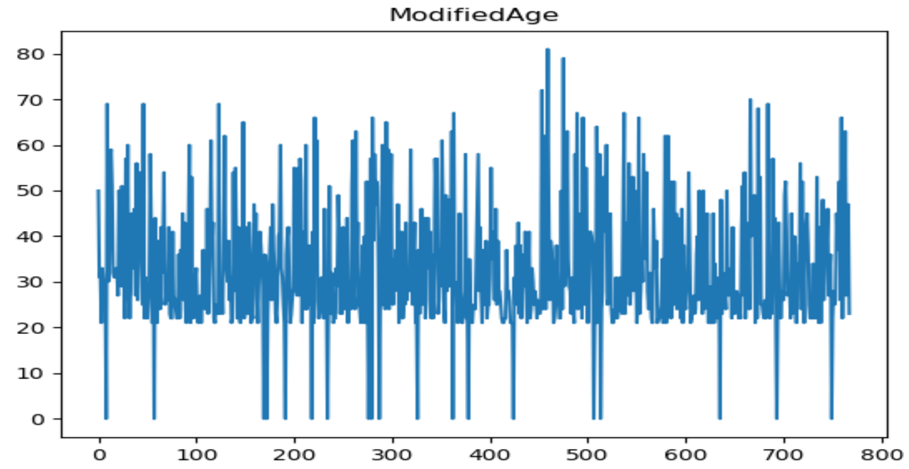


The below graph was plotted after replacing all NaNs with average value.

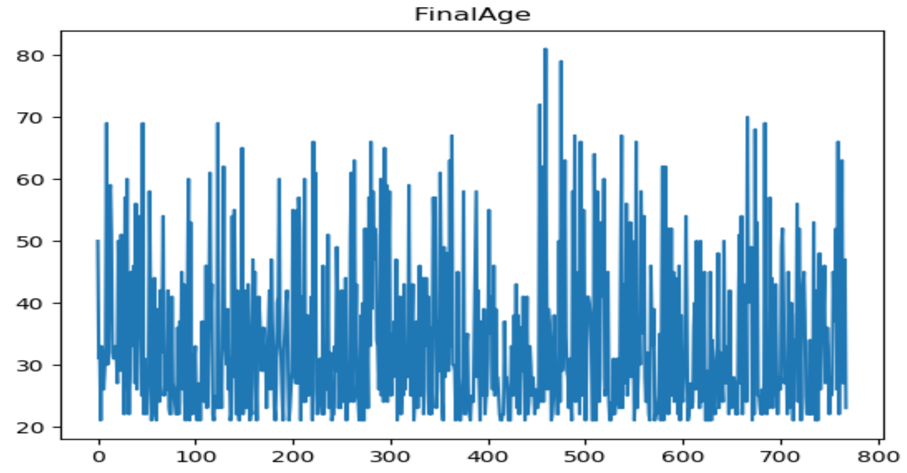


1.1.8 Age

The following graph was plotted replacing all NaNs with zeros.



The below graph was plotted after replacing all NaNs with average value.



1.2 Outlier Detection and Removal

1.2.1 Replacing Zero Outliers

It was observed that many values were recorded as zeros for Glucose, BMI and BloodPressure. This does not make sense logically. Even in the graphs plotted for the above features, it was obvious that zero as a data point was an outlier.Thus, all zeros were converted to NaNs, which were subsequently filled when we processed missing data.

1.2.2 Removal of other Outliers

The Mahalanobis distance (MD) is the distance between two points in multivariate space. This is used to compute the distance between points, even if they are correlated.
The Mahalanobis distance is computed as follows -

$$MD = (x - x_{mean})^T \cdot C^{-1} \cdot (x - x_{mean})$$

where ,
C is the co-variance matrix of the feature space.
x_{mean} is a vector consisting of the mean values recorded for each feature, repeated as many times as the number of rows.
x is the values of each feature after pre-processing.

We are removing outliers based on the 68-95-99.7 rule (*empirical rule*). This gives an idea of the distribution of values around the mean in a normal distribution, with a band of 2,4 and 6 standard deviations. 68% 95% and 99% of the values lie within one, two and three standard deviations of the mean, respectively.

In mathematical notation, these facts can be expressed as follows, where X is an observation from a normally distributed random variable, μ is the mean of the distribution, and σ is its standard deviation

$$Pr(\mu - \sigma <= X <= \mu + \sigma) \approx 68$$

$$Pr(\mu - 2\sigma <= X <= \mu + 2\sigma) \approx 95$$

$$Pr(\mu - 3\sigma <= X <= \mu + 3\sigma) \approx 99.7$$

2 Feature Extraction and Feature Selection

2.1 Normalization

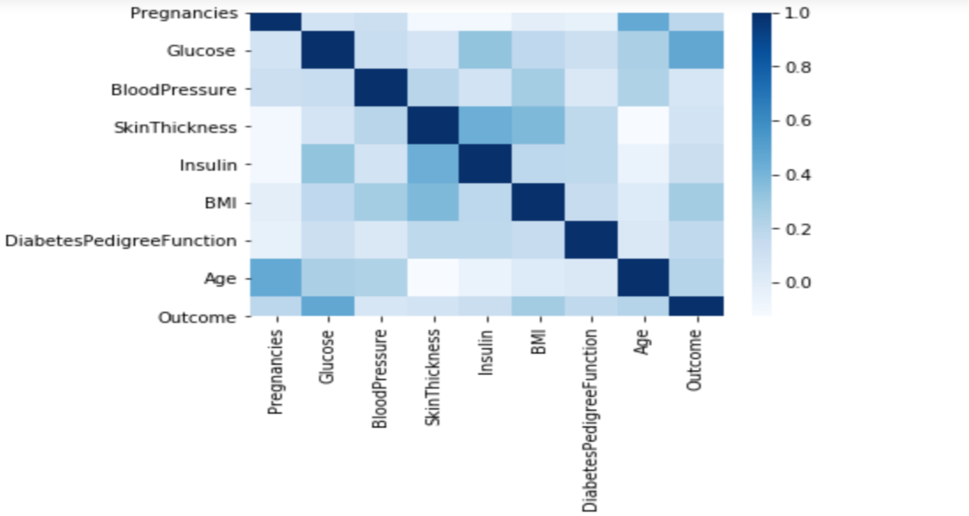
The data was normalized using the normalize function in NumPy, which scales each input feature to it's unit norm, by dividing it with it's magnitude.

2.2 PCA (Principal Component Analysis)

We tried to reducing the dimensionality by various dimensions and realised that the results were most accurate when we reduced it to seven dimensions. However, the overall accuracy of the model turned out to be better when we didn't apply PCA at all. Thus, we have used each feature vector as is after pre-processing.

2.3 Exploratory Data Analysis

The following is the correlation matrix of the data.
We can clearly see that the features aren't that correlated which is why PCA did not improve the accuracy of the model.



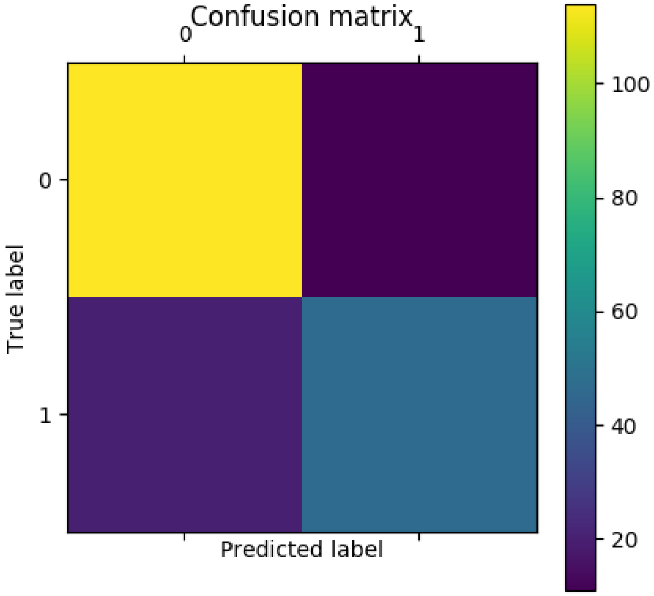
3 Model Building

Logistic Regression

The given problem of telling whether a patient has diabetes or not is a classification problem. Hence, we have used **Logistic Regression** as our Model to predict the output.

Our model produces:
Average Train set accuracy: 0.770
Average Test set accuracy: 0.758
Maximum Accuracy: 0.828

Confusion Matrix for the above model



4 References

- Mahalanobis Distance : https://stackoverflow.com/questions/46827580/multivariate-outlier-removal-with-mahalanobis-distance
- 68–95–99.7 rule : https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7$_r ule$
- PCA : https://www.geeksforgeeks.org/principal-component-analysis-with-python/
- Missing Values Handling : https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-1-dda69d4f88ca
- Correlation Matrix : https://nbviewer.jupyter.org/github/PBPatil/Exploratory$_D ata_A nalysis-Wine_Q uality_D ataset/blob/master/winequality_w hite.ipynb$