**Statistical Learning Lab**

**Assignment – 3**

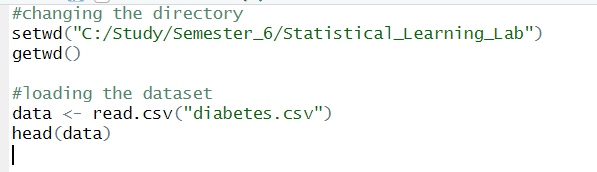
**LDA, QDA and KNN**

**Name: Semanti Ghosh**

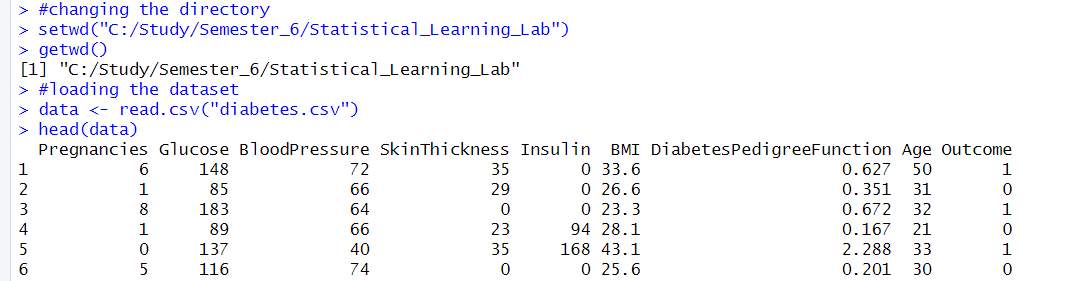
**Roll No.: 22IM10036**

**Loading the dataset**

Code snippet

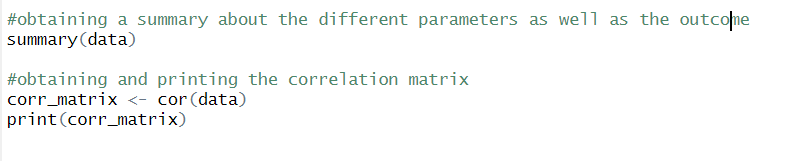


Output

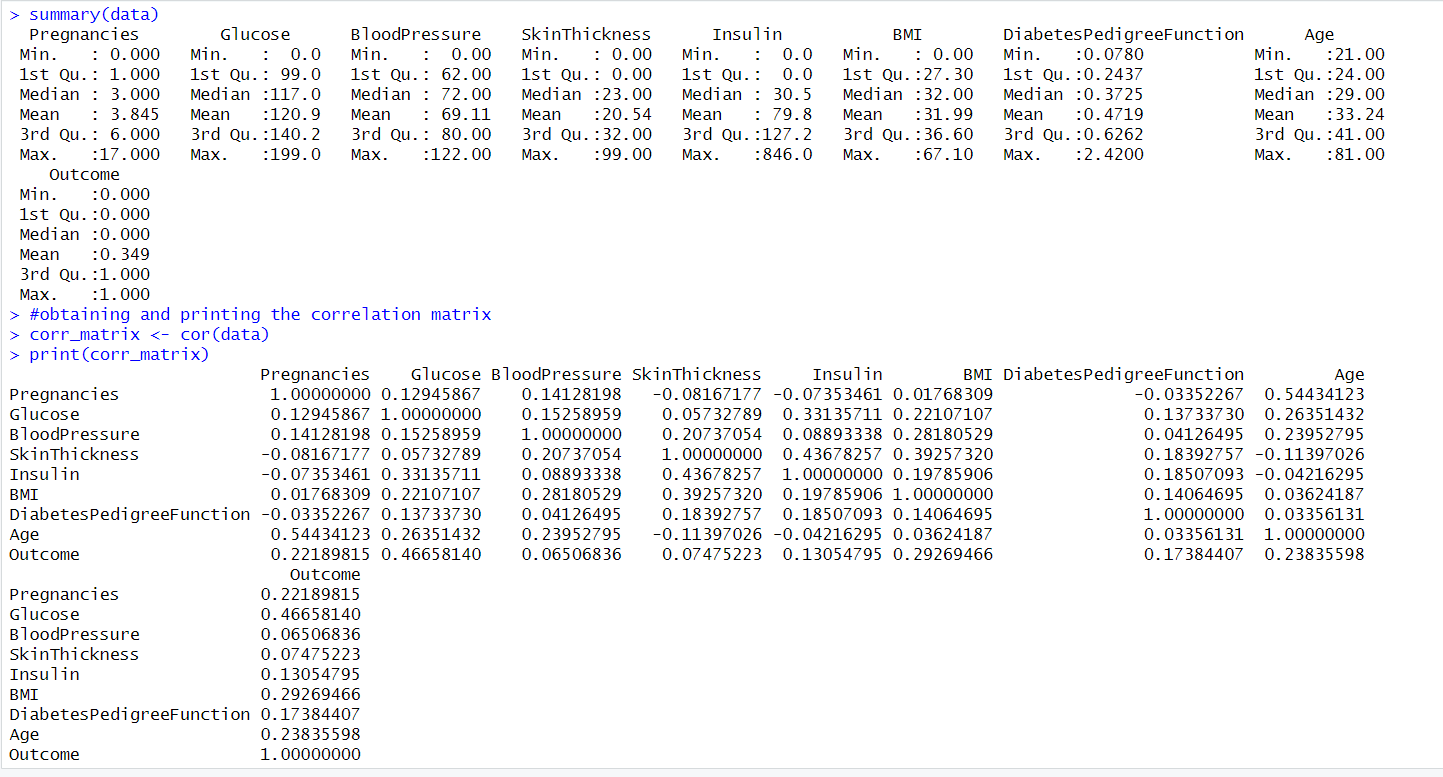


**Analysing the data (and finding the correlation between the parameters)**

Code snippet

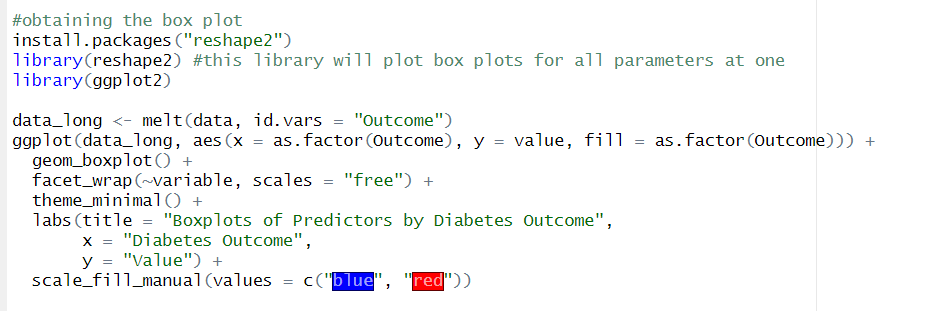


Output

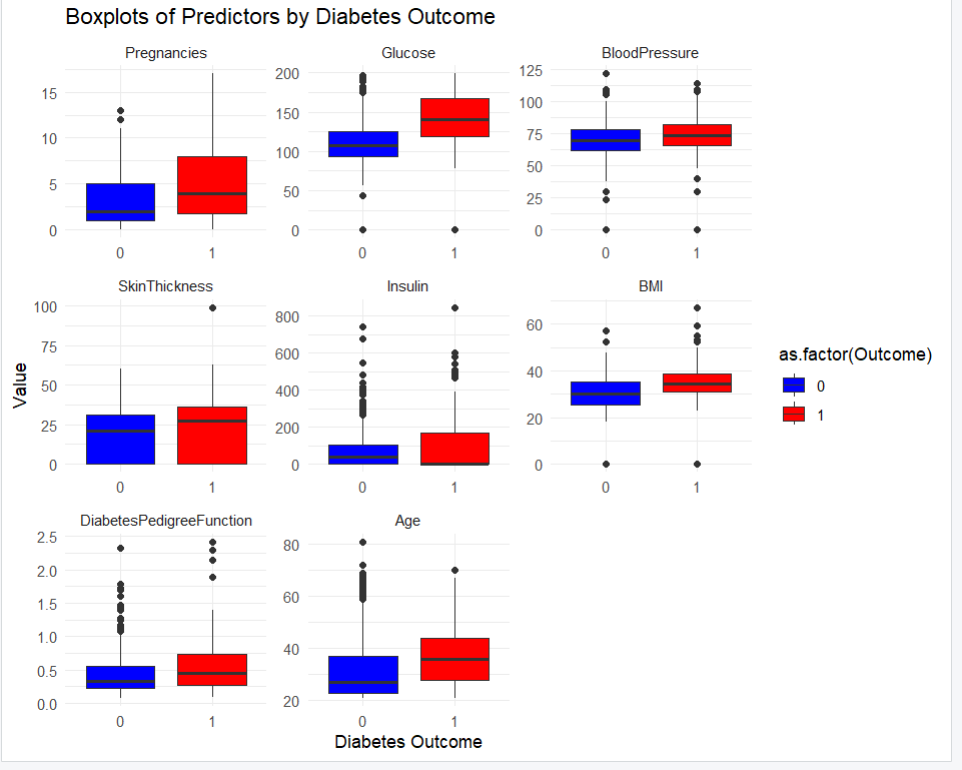


**Obtaining the boxplot**

Code snippet

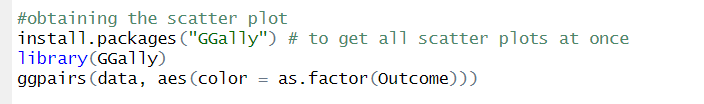


Output

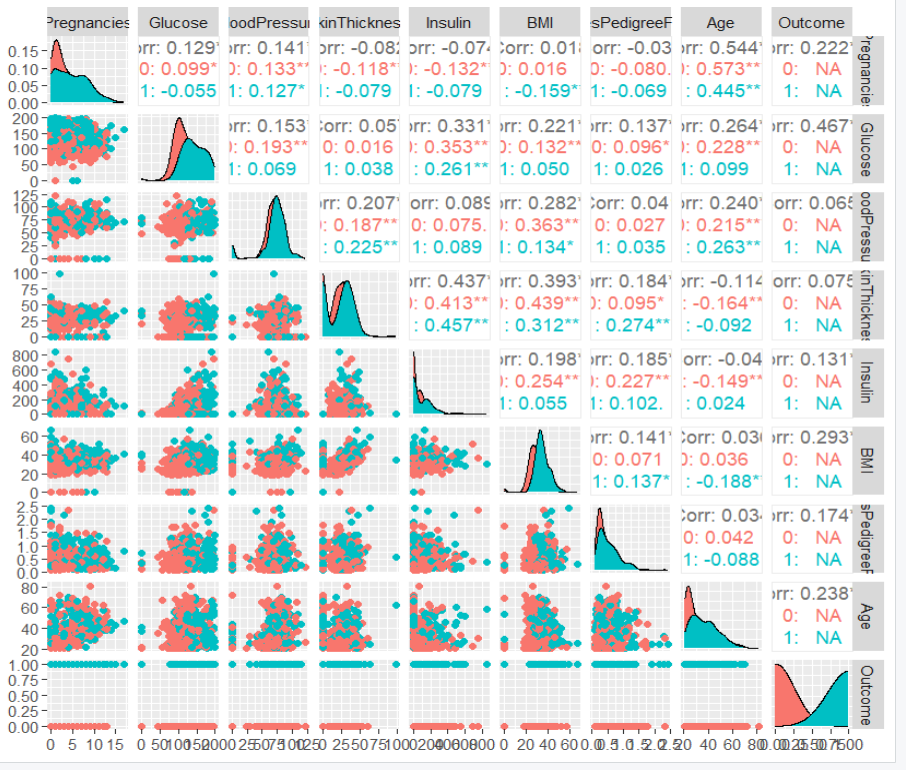


**Obtaining the scatter plot**

Code snippet

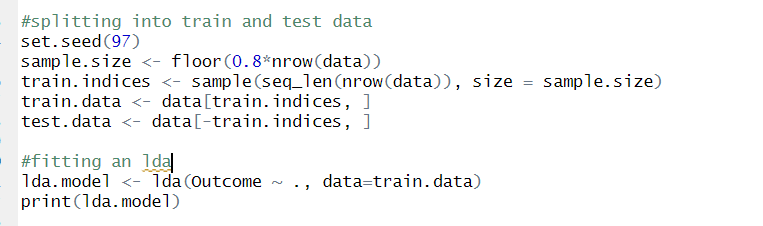


Output

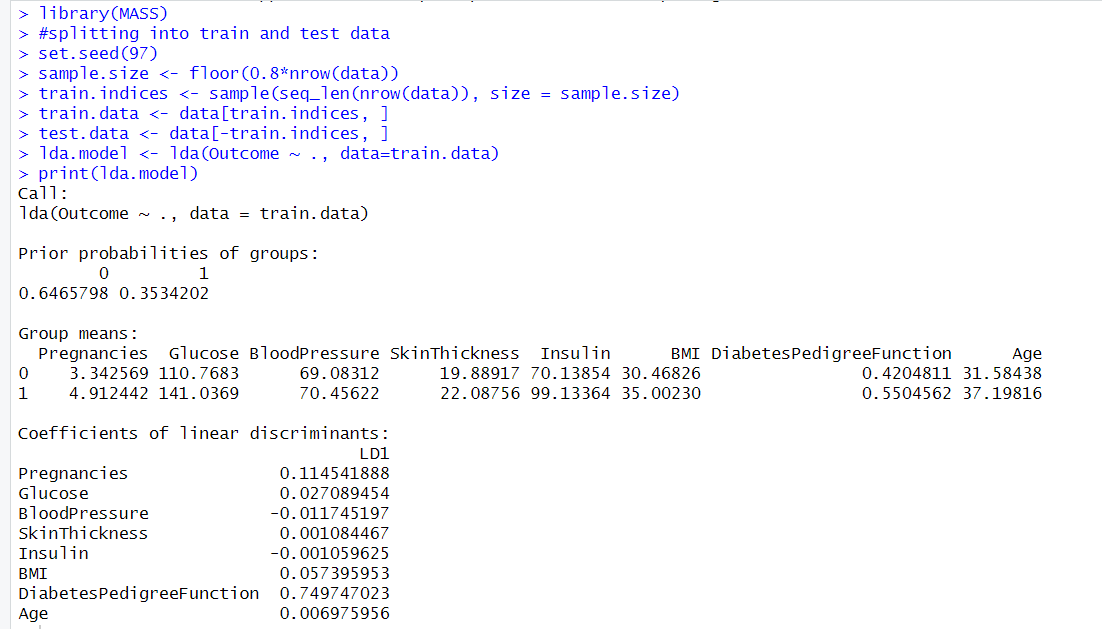


**Splitting into train and test set and fitting an LDA**

Code snippet

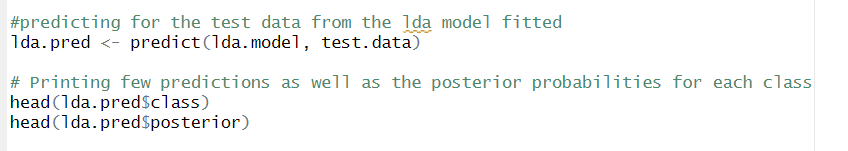


Output

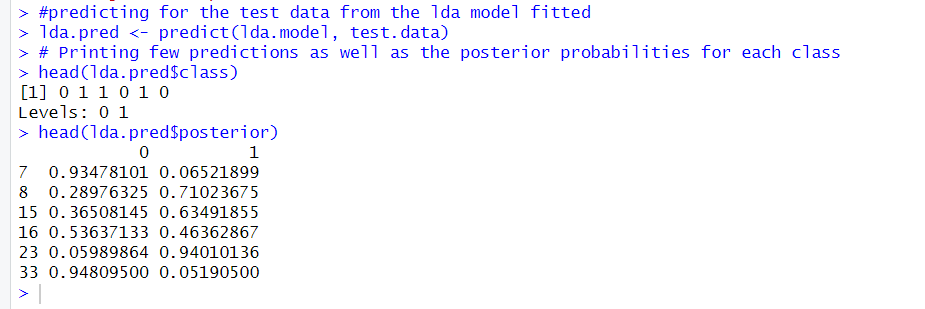


**Predicting the labels for the test data using the model fitted**

Code snippet

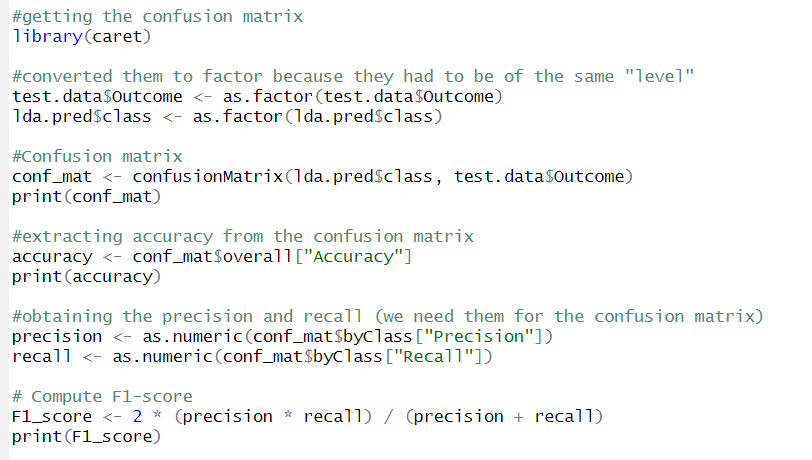


Output

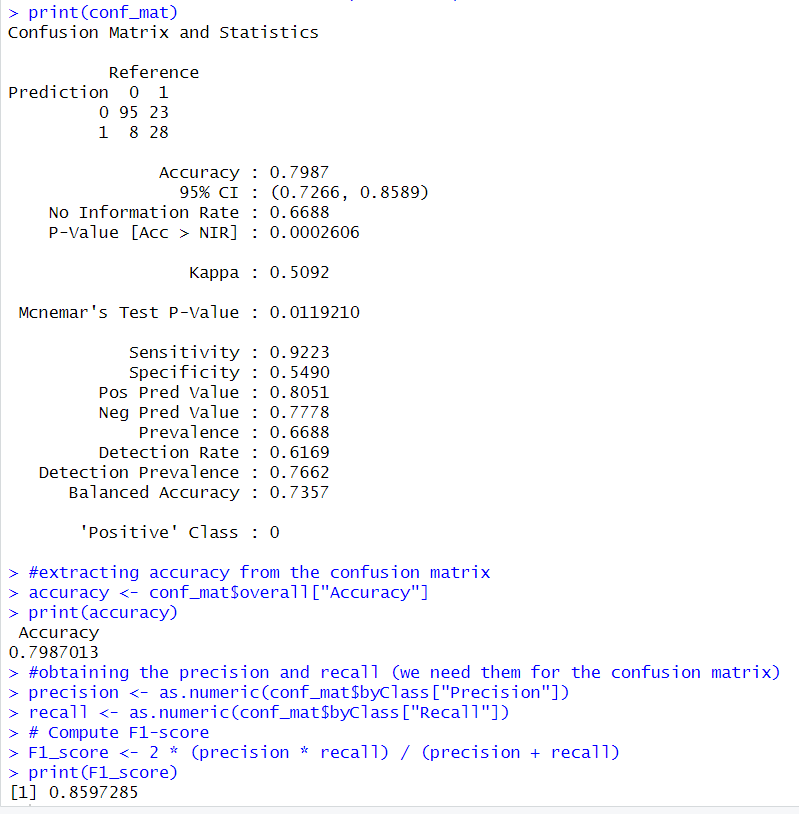


**Deriving the confusion matrix, accuracy, F1-score on the test data (for LDA)**

Code snippet

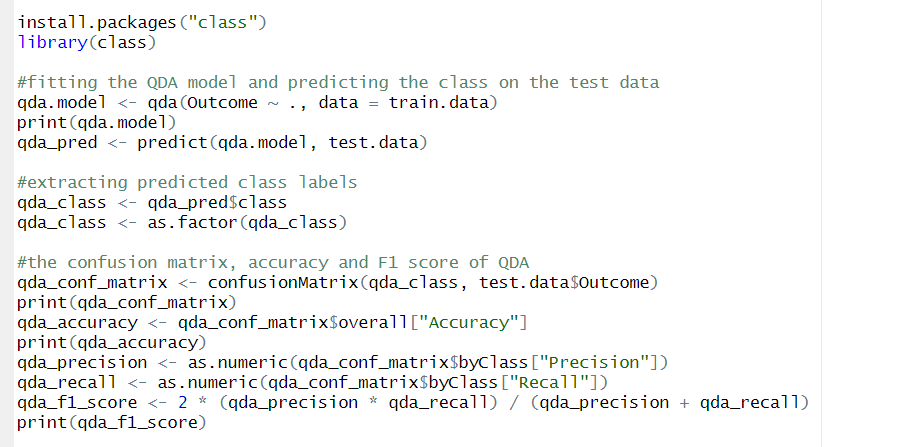


Output

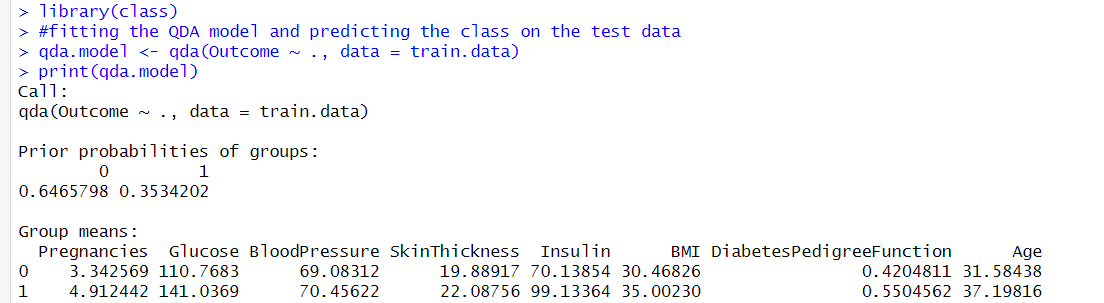


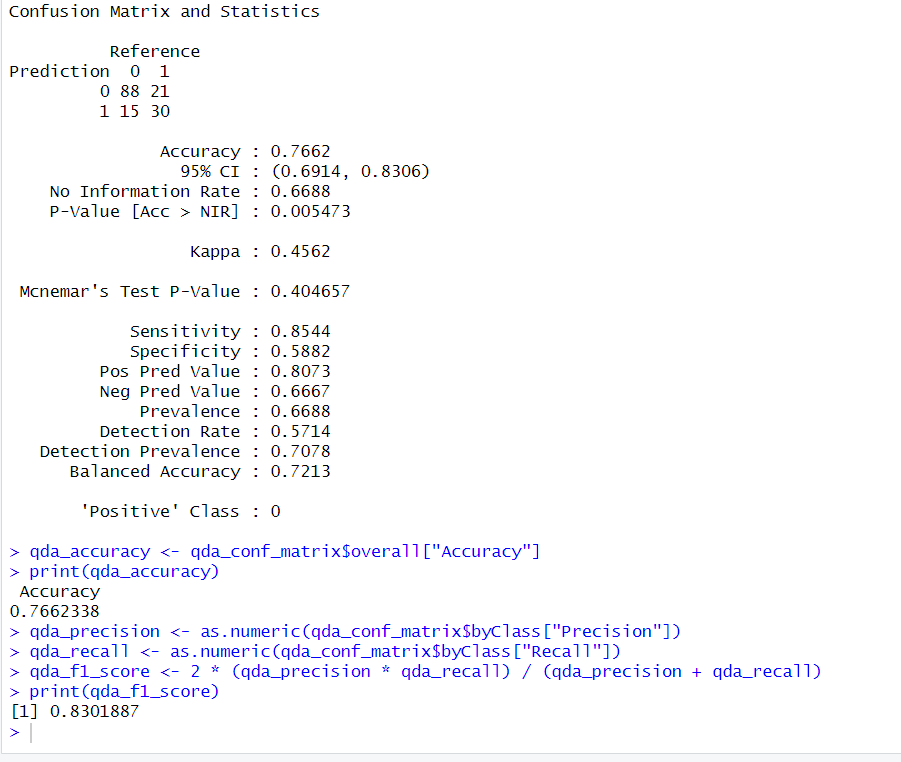
**Fitting a QDA model and printing its metrics (confusion matrix, accuracy and F1 score)**

Code snippet



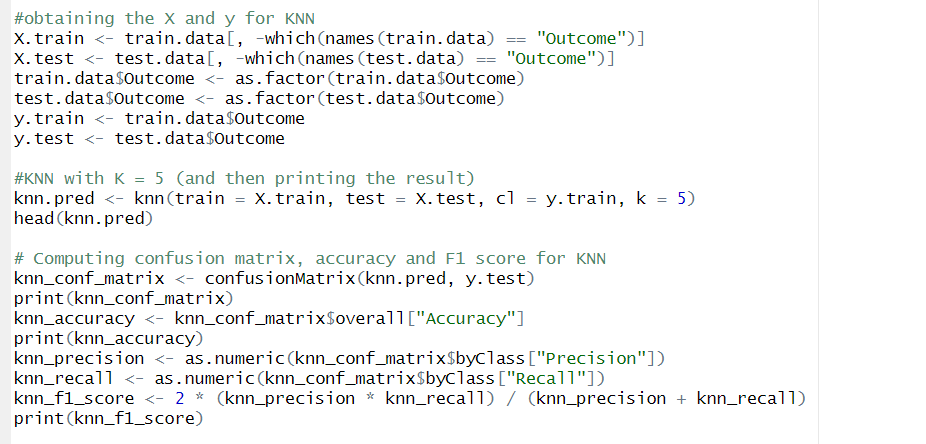
Output



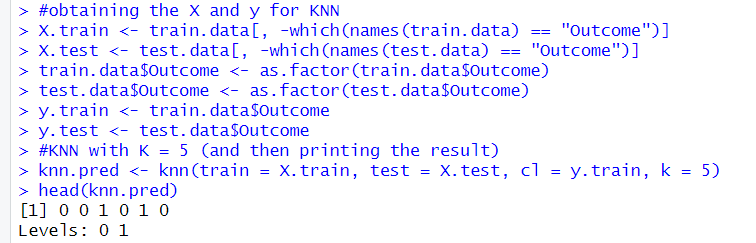


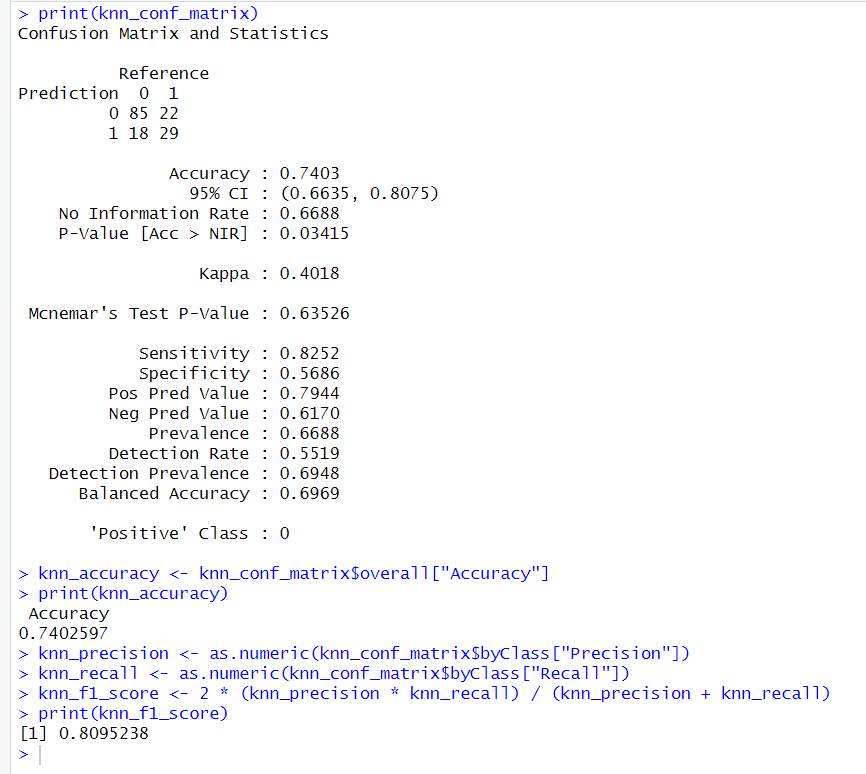
**Fitting a KNN model with K=5 and printing the necessary metrics**

Code snippet



Output



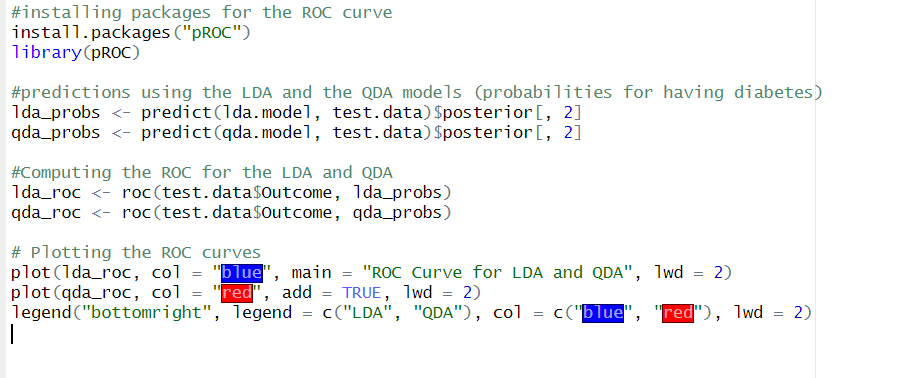


**Interpretation from the different classifiers fitted (LDA, QDA and KNN)**

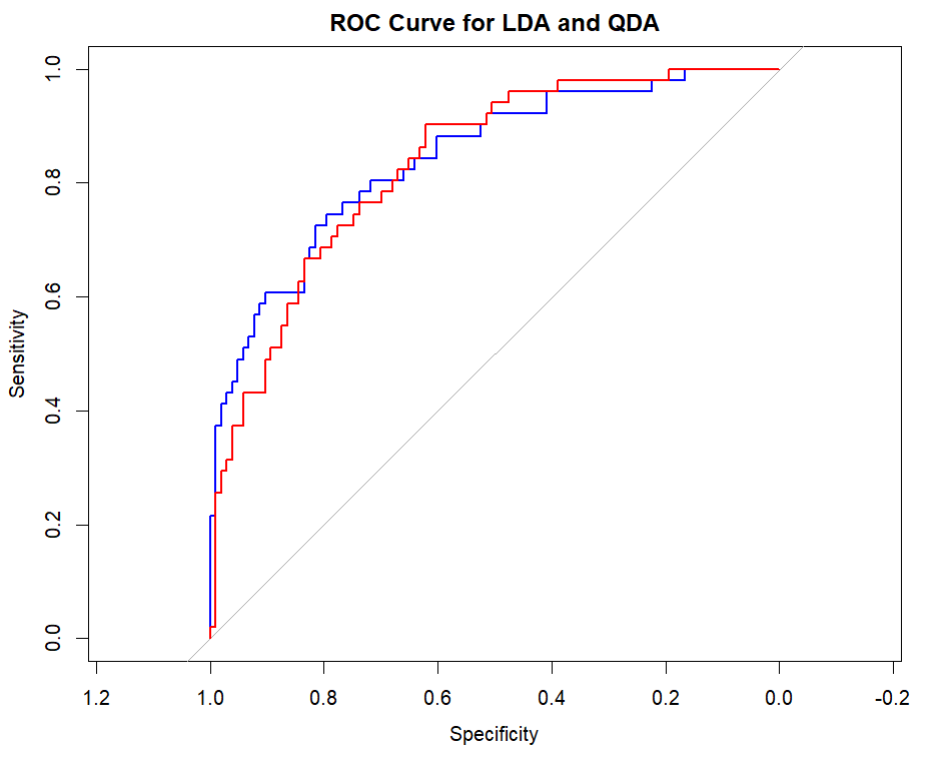
Both the accuracy and F1 score are maximum in case of LDA and minimum in case of KNN. From this, we can say that the separation boundary is more linear than curved (that’s why LDA performs better than QDA). Also, we can say that since KNN performs poorly, the data does not need a highly flexible, non-linear decision boundary. Besides, there are too many parameters for KNN to perform well.

**Plotting the ROC curve for LDA and QDA using the test data**

Code Snippet

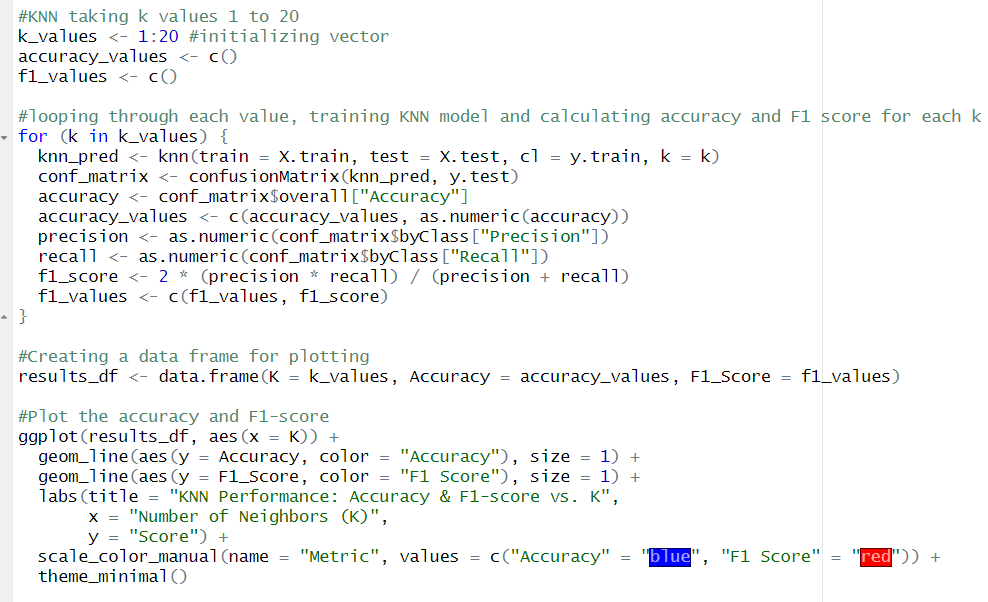


Output

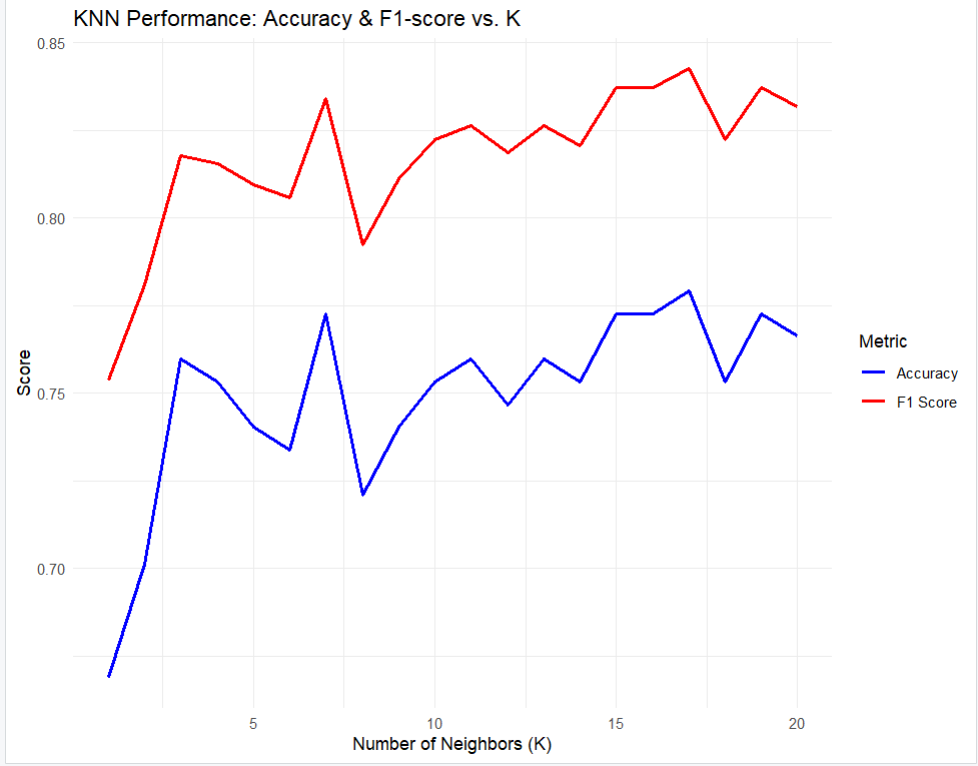


**Fitting KNN models for k=1 to k=20 and then plotting the accuracy and F1-scores**

Code snippet



Output



**Interpreting the results:**

* Small K (K=1 to K=3): high variance due to overfitting, sensitive to noise.
* Moderate K (K=5 to K=10): usually a good balance between bias and variance and peak accuracy and F1-score often occur in this range
* Large K (e.g., K>15): Predictions become less flexible, can underfit if they become too generalised