**Statistical Learning Lab**

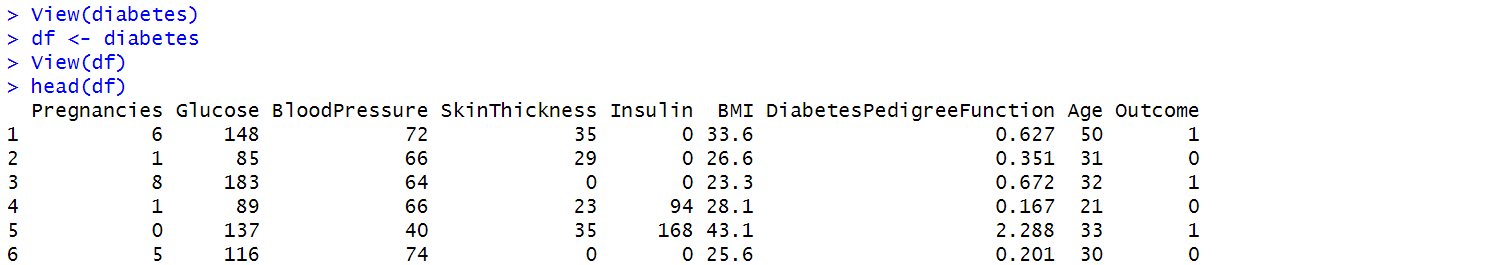
**Assignment - 3**

**LDA, QDA and KNN Assignment**

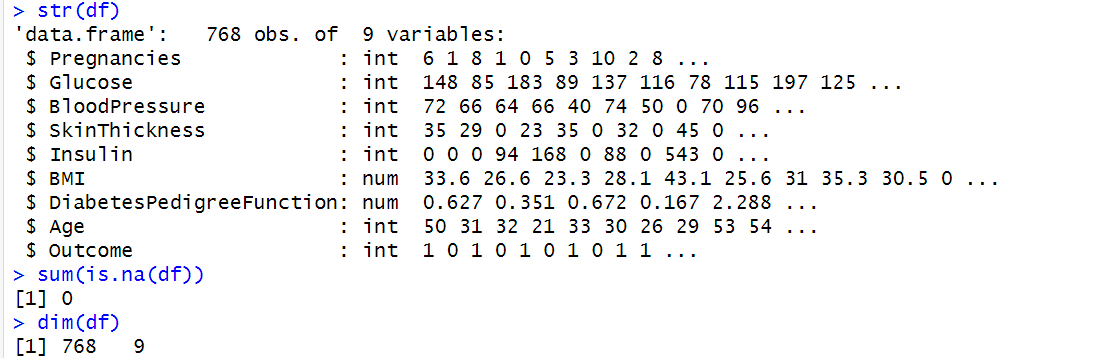
**Name : Sunny Kumar Roll No: 22IM10040**

**Show the code snippets and the corresponding output for the following:**

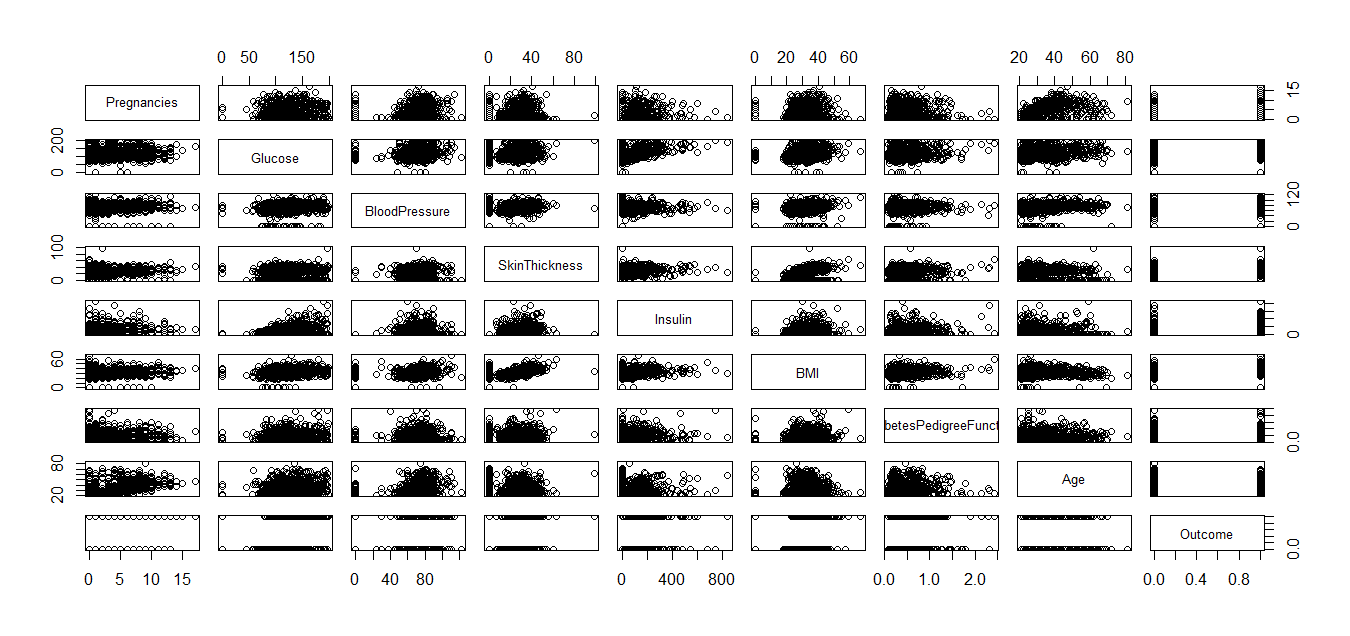
1. **Load the dataset “diabetes.csv”. Display first few rows of the dataset.**



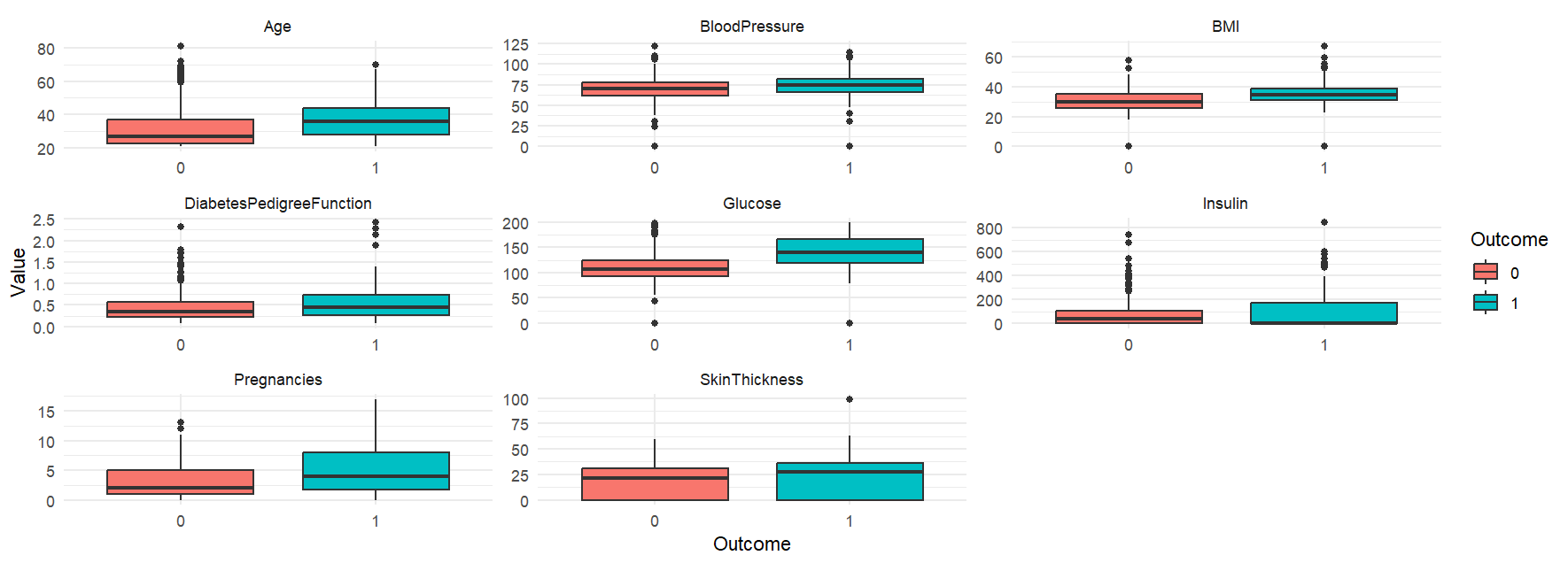
1. **Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the “Outcome” variable.**



Scatter plot between different variables is given below:

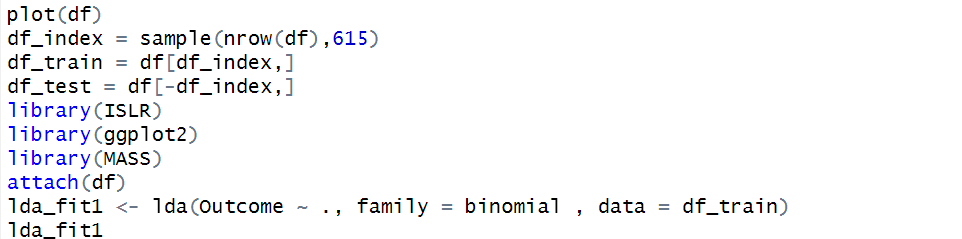


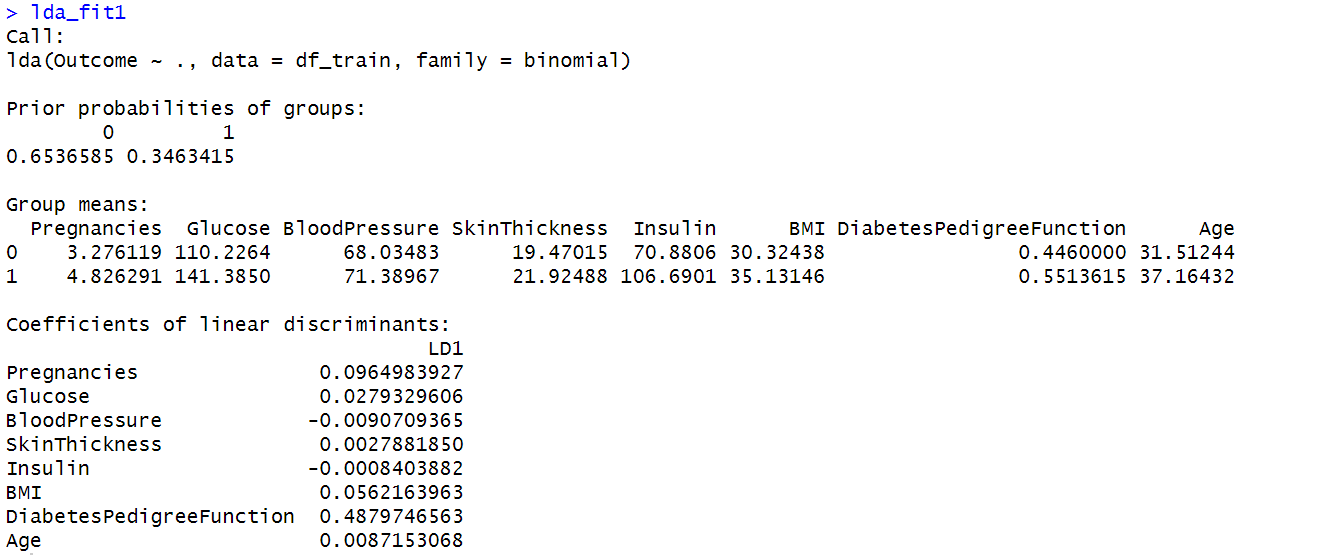
My Inference : The median BMI for individuals with Outcome = 1 (diabetic) is higher than for individuals with Outcome = 0 (non-diabetic). This suggests that diabetics tend to have a higher BMI on average. BMI appears to have a relationship with the Outcome variable. Higher BMI values are more associated with diabetes (Outcome = 1). Similarly, All the plots are shown below.



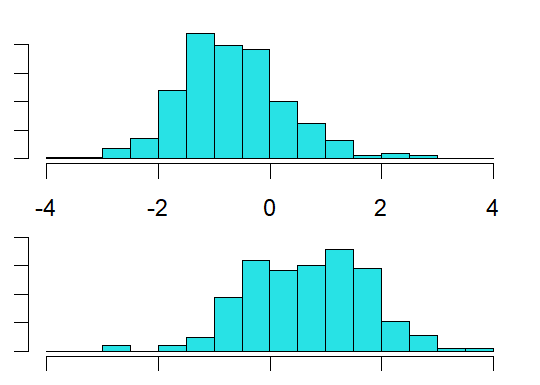
1. **Randomly sample 80% of the data as training data and rest as test data. Fit a LDA model and interpret the result.**

Code and corresponding output :





> plot(lda\_fit1)



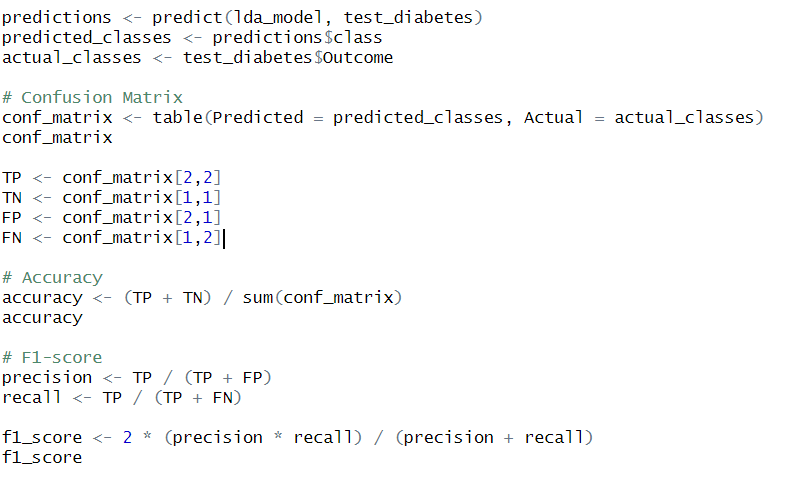
Interpretation from the result :

Diabetics (Outcome = 1) tend to have:

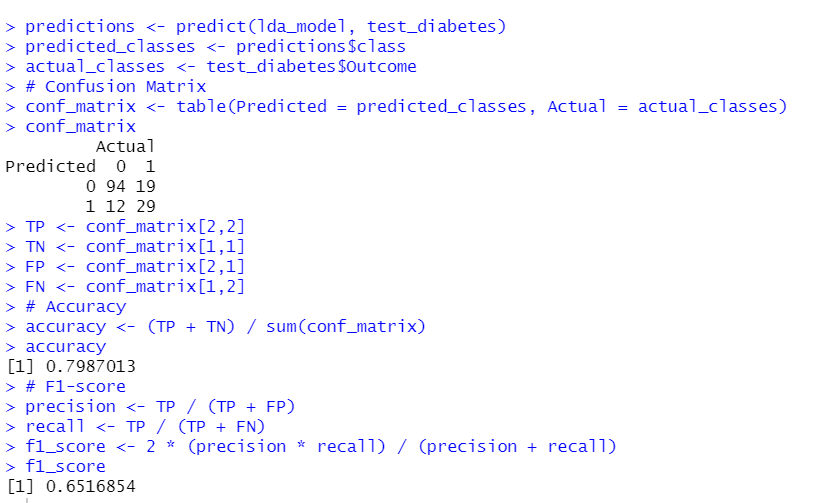
* Higher **Pregnancies**
* Higher **Glucose levels** (a strong indicator)
* Higher **BMI**

1. From the model fitted in problem 3, derive confusion matrix, accuracy, and F1-score on test data.

Code:

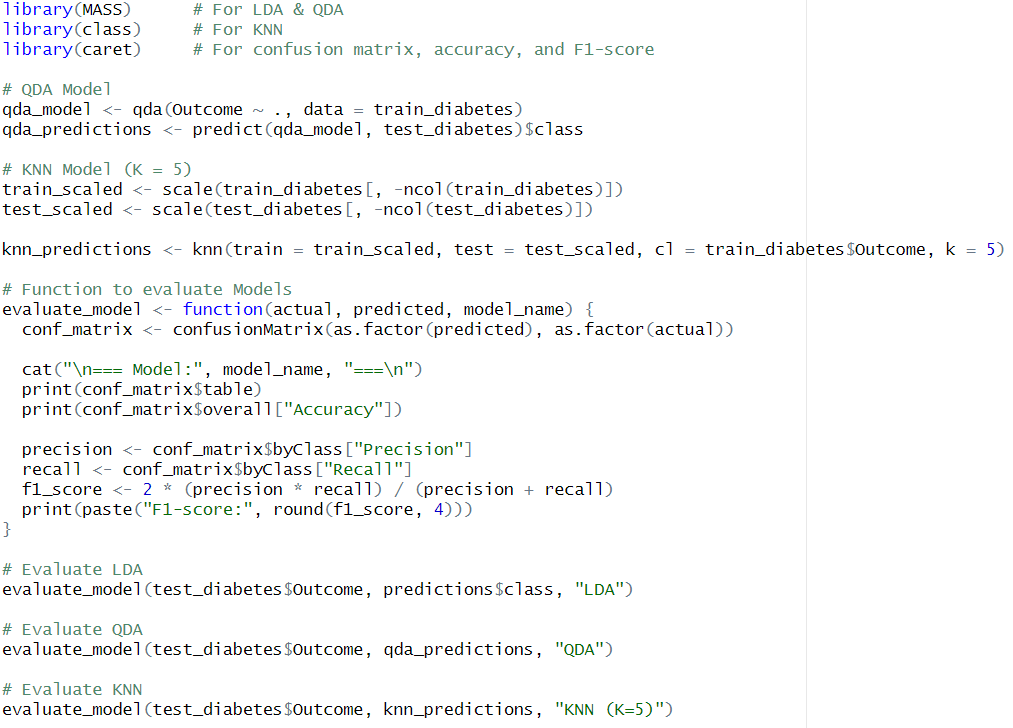


Output:

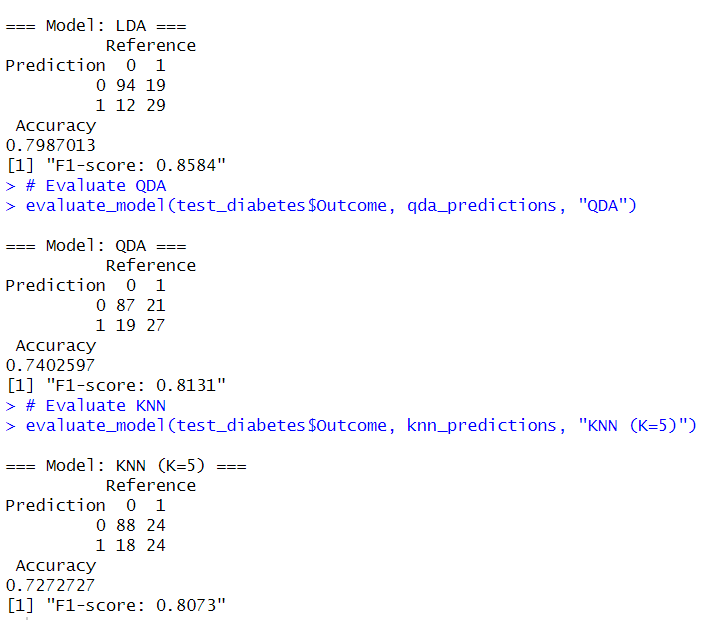


1. Fit QDA and KNN (K = 5) models on training data. Compare the metrics in problem 4 for LDA, QDA and KNN models for test data and discuss the results.

Code :

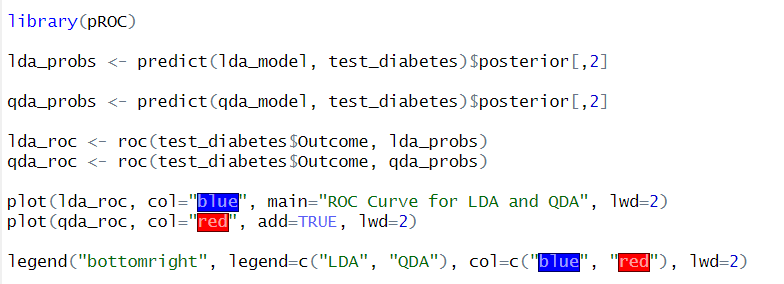
****

Output:

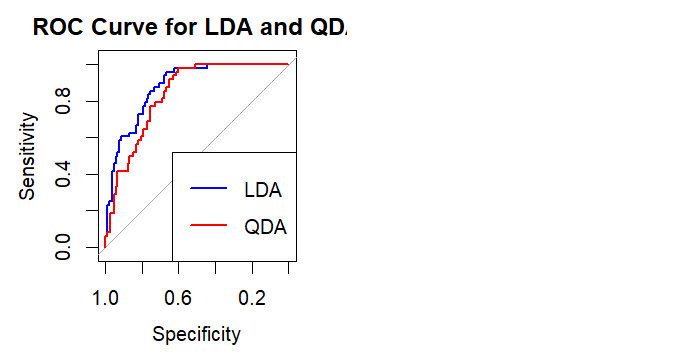
****

1. Plot ROC curve for LDA and QDA models using the test data.

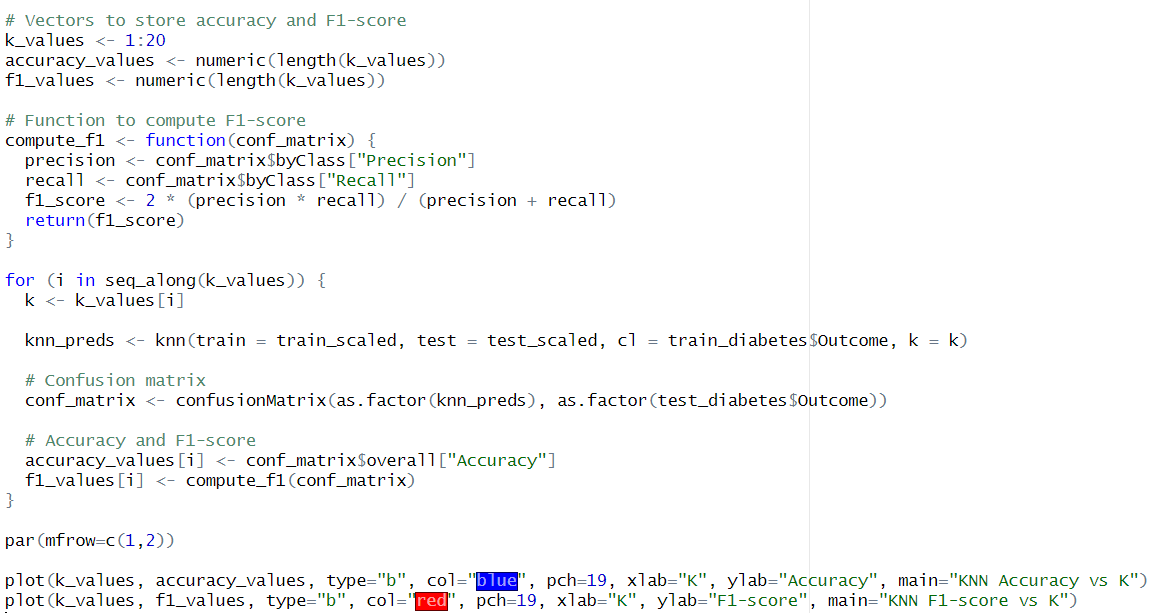
Code:

****

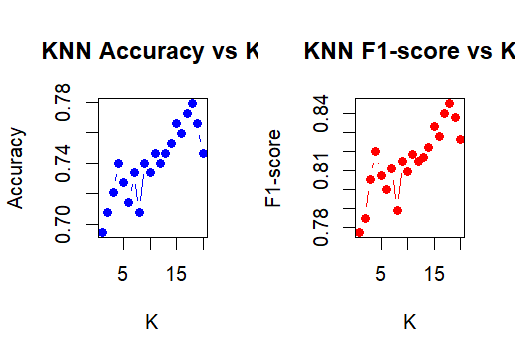
Output:

****

1. Plot accuracy and f1-score by varying the neighbourhood size from K=1 to K=20 and interpret the results.

****

Output:



**Interpretation:**

* For smaller K values (e.g., K = 1 to 5), the model overfits to the training data, resulting in lower accuracy and F1-score. Also, predictions are more sensitive to noise.
* For higher K values (e.g., K = 15 to 20), accuracy and F1-score stabilize, indicating that the model generalizes well. The performance is better balanced, meaning less overfitting and better predictions.
* The best value of K is likely between 15 and 20, where both accuracy and F1-score are highest.
* Choosing K too high (e.g., K > 20) may start to decrease performance due to underfitting.

Data can be downloaded from: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

Description of the study:

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261). American Medical Informatics Association.