Statistical Learning Lab

Assignment - 3

LDA, QDA and KNN Assignment

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

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

2. 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.

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

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

5. 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.

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

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

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.