COMP534: Applied Artificial Intelligence

First Assignment

A Report

on

"Performance analysis of multiple supervised learning methods for solving binary classification problem"

Submitted to: **Dr. Pamela Bezerra**

Submitted by: **Ananya Swami**

Rajesh Adepu

Introduction

The performance analysis has been executed using python programming language. A code.py file has been submitted along with this report containing the code for performance analysis.

The code begins with importing libraries. Various libraries are used within the code namely; numpy, matplotlib, pandas, scikit-learn (sklearn).

Loading the dataset:

To read the "diabetes.csv" file, we use pandas library and divide the data into input X and output Y.

Splitting the dataset:

Here, we have split the data into 3:1 ratio using train_test_split from sklearn library i.e. train(75%) and test(25%) data.

Cleaning the dataset:

To clean and preprocess the data we have used StandaradScalar from sklearn library.

Classification:

The classification has been done using three different supervised learning methods namely; **k-nearest neighbors**, **random forest** and **logistic regression**.

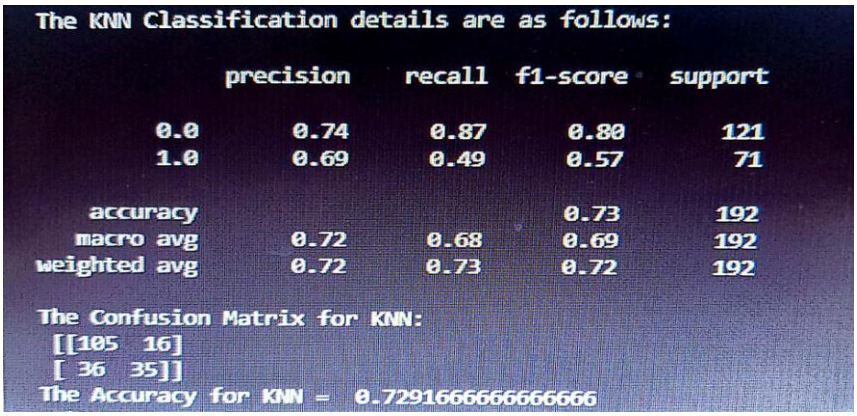
To get this classification done we have directly used KNearestClassifier(), RandomForestClassifier() and LogisticRegression() functions from sklearn library.

Analysis:

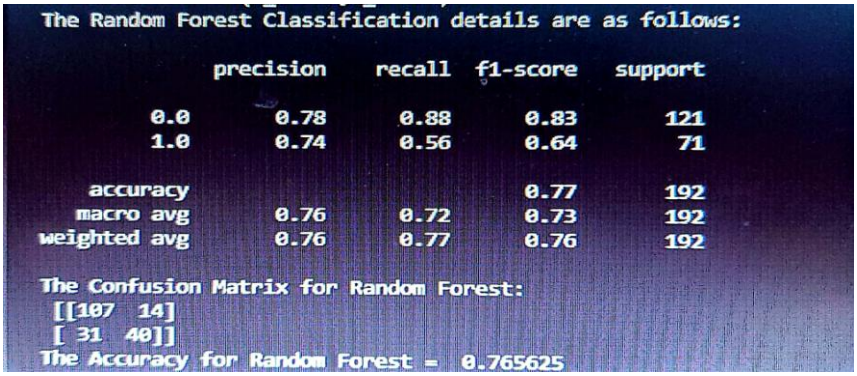
The analysis is done by creating tables for each classification method and graph comparing accuracies of all three methods, out of which "*logistic regression was found to be the best*".

Evaluation

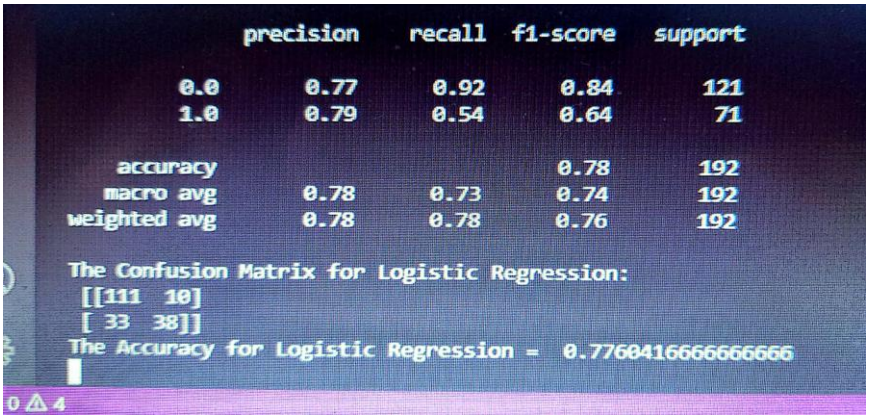
k-Nearest Neighbors Classifier: (with k=10)



Random Forest Classifier: (with 50 estimators)

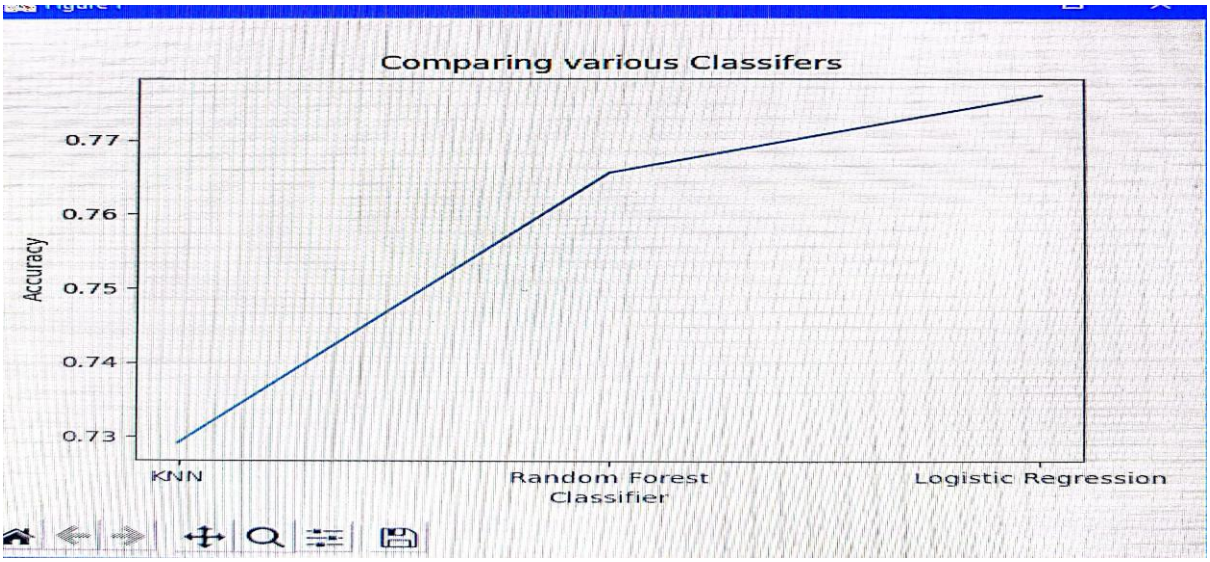


Logistic Regression Classifier: (best among all three)



Results and Conclusion

As mentioned, a graph was also plotted using matplotlib library, comparing the accuracies of all three methods and it was found that logistic regression performed best among three as the number of instances were very large as compared to the number of features which were just 8. It generally does not makes any assumptions in distributing classes in a feature spce. The graph can be seen as follows:



Challenges faced in creating this project are as follows:

- There was a conflict of ideas wherein choosing three supervised learning methods out of the list of methods.
- Also, finding the best version for each method which includes finding appropriate k value and estimators was tedious.
- Time management was also one of the problems as we had to put in equal efforts in taking the assignment to its accomplishment stage.

Task Allocation:

Choosing 3 methods was done on a mutual agreement basis.

Programming and Code Creation: Ananya

Report Preparation: Rajesh