Assignment-1 Applied AI

Karthik Upadhyayula

2021626620

In this assignment I have done three classification problems which are part of supervised learning techniques. Namely,

1. K-Nearest Neighbour
2. Decision Tree
3. Random Forest

The libraries used for these techniques are:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import cross_validate
from sklearn.metrics import precision_score, recall_score, accuracy_score,
f1_score
from sklearn.metrics import confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
```

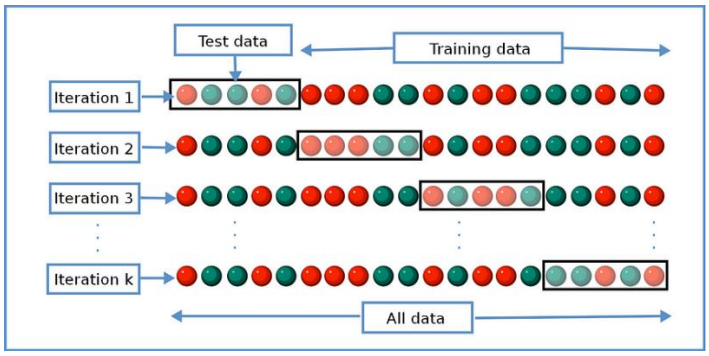
The classification method used for K-Nearest Neighbour is a library from scikit learn called kNeighboursClassifier which has hyperparameter as K value. The optimal K value is obtained using K-Fold cross validation method.

The Classification method used for Decision Tree is also based on Scikit learn library. The classifier used is DecisionTreeClassifier which has a hyperparameter as depth of the decision tree. The optimal depth for the decision tree is determined by using K fold cross validation method.

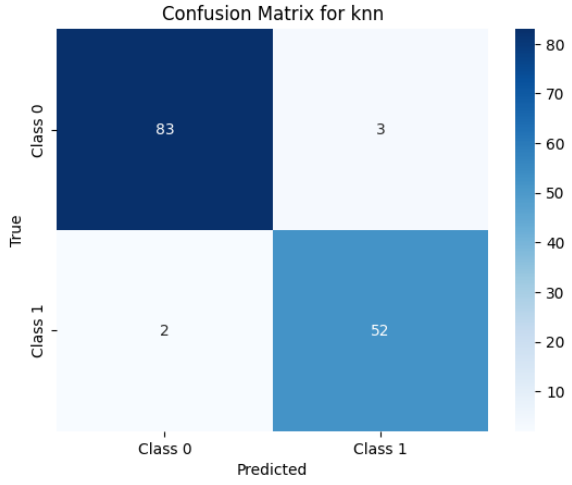
The Classification method used for Random Forest which is ensemble technique and is based on scikit learn library. The classifier used is RandomForestClassifier which has two hyperparameters namely estimators and depth. These can be obtained by using GridSearchCV which provides the optimal estimator and depth for the tree.

The training and test set are split into 80% and 20% respectively by using train_test_split function from scikit learn.

When using k fold cross validation or GridSearchCV we have taken fold_number as 5 in this case. The training set is split into 5 different subsets where any 4 are used a training set and the last set is used as a test set to determine the hyperparameters.



Then we retrain the model with obtained hyperparameters and then test according to the 80% and 20% training and test split respectively to obtain the result.



In the first column of the KNN Confusion Matrix, there are 83 true positives and 3 false positives, meaning that 83 samples were accurately identified as positive and 3 samples were misidentified as positive. Two samples are wrongly labelled as negative, while 52 samples are correctly classified as negative, as seen by the two false negatives and 52 genuine negatives in the second column.

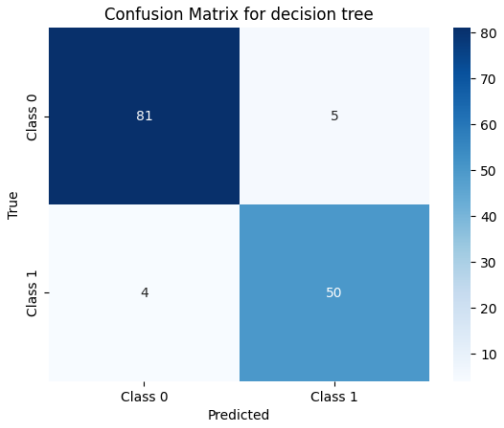
K value for KNN 5

Precision for KNN 0.945455

Recall for KNN 0.962963

Accuracy for KNN 0.964286

F1-score for KNN 0.954128



In the first column of the Decision Tree Confusion Matrix, there are 81 true positives and 5 false positives, meaning that 81 samples were accurately identified as positive and 5 samples were misidentified as positive. Four samples are wrongly labelled as negative, while 50 samples are correctly classified as negative, as seen by the two false negatives and 50 genuine negatives in the second column.

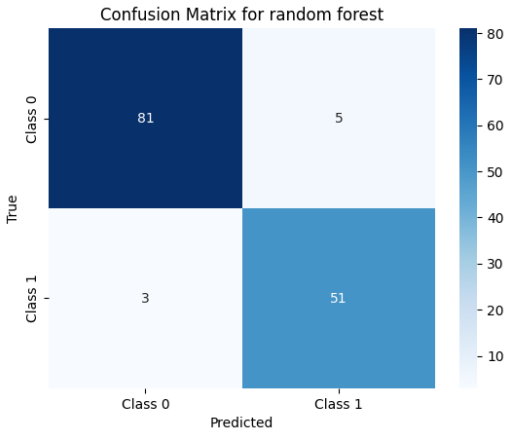
depth for decision tree 3

Precision for decision tree 0.909091

Recall for decision tree 0.925926

Accuracy for decision tree 0.935714

F1-score for decision tree 0.917431



In the first column of the Random forest Confusion Matrix, there are 81 true positives and 5 false positives, meaning that 81 samples were accurately identified as positive and 5 samples were misidentified as positive. Three samples are wrongly labelled as negative, while 51 samples are correctly classified as negative, as seen by the two false negatives and 51 genuine negatives in the second column.

depth for random forest 6.000000

estimator for random forest 50.000000

Precision for random forest 0.910714

Recall for random forest 0.944444

Accuracy for random forest 0.942857

F1-score for random forest 0.927273

Conclusion:

According to the evaluation measures, KNN outperforms the other two in terms of accuracy and recall. Then, depending on the estimators used, Random forest has accuracy that is second only to Decision tree in terms of computation time.

In general, the decision on which classification method to use is influenced by the particular issue at hand, the size and complexity of the dataset, and the resources at hand. Before choosing the best classifier for a given problem, it is crucial to test several of them and assess their performance using the right metrics.

From this Project, I have learned using supervised learning algorithms for different datasets. I would also like to implement naïve bayes, Logistic regression and SVM on the same dataset in this project.