Lovemore Tenha
Project8 Report

In this report, we explore the performance of various ensemble approaches to conduct classification. We applied the algorithms on 2 different datasets; the digits dataset and the mammographic_masses dataset from the UCI repository. We use the ensemble approaches provided by the scikit-learn libraries in python. These approaches are Random Forest, Bagging and Adaboost. We use the Decision Tree Classifier as the base classifier for Bagging and Adaboost. The performance accuracy and the running times of these ensemble approaches are analysed and compared to the results obtained from using the Decision Tree Classifier only. The table below shows the results obtained after applying the algorithms on the digits dataset.

Ensemble Approach	**Classification Accuracy(testing data)**	**Runtime/s**
Random Forest	0.98	0.25
Bagging	0.96	1.31
Adaboost	0.76	0.35
Decision Tree Classifier	0.86	-

Performance and Runtime for each ensemble algorithm on digits dataset

It can be seen from the table above that the classification accuracy of the ensemble approaches is better than that of the Decision Tree Classifier by itself. Random Forest and the Bagging approaches have the highest accuracy while Adaboost has the lowest accuracy. These results are as expected since the ensemble approach aggregates the results from various models and is therefore, likely to build the best model. The running times of the ensemble algorithms are also higher, as expected, since the ensemble approach runs the base classifier several times.

The following results were obtained after applying the ensemble classification algorithms on the mammographic_masses dataset

Ensemble Approach	**Classification Accuracy(testing data)**	**Runtime/s**
Random Forest	0.80	0.10
Bagging	0.81	0.13
Adaboost	0.82	0.11
Decision Tree Classifier	0.74	-

Unlike in the digits datasets, Adaboost has the highest classification accuracy on the test dataset. However, the accuracies for the ensemble approaches are relatively comparable with the Random Forest and Bagging with accuracies 0.80 and 0.81 respectively. The accuracy for the Decision Tree Classifier alone is the lowest at 0.74, as expected. The running times for the ensemble approaches are also relatives similar.

One of the most important parameters for the ensemble approaches is the n_estimators, which determines the number of models that the ensemble algorithm builds using the base classifier. Altering the n_estimators affects the accuracy of the algorithm. Lowering the number of estimators lowers the accuracy and the converse is true. However, above a certain number of estimators, the accuracy does not seem to increase significantly. For the Random forest, bagging algorithm, lowering number of estimators from 100 to 10 affects the accuracy but increasing the the estimators to 1000 does not seem to have significant effect on the performance .