# Machine Learning

## Assignment 4

Ishaan Bassi, 2016238

a) min_leaf = [8,20,30,45,75,90,100]
min_split = [45,75,90,100]
depth = [3,5,8,10,20]
estimators = [15,25,40,50,70,90] (For random forest)
criterion = ['gini','entropy']

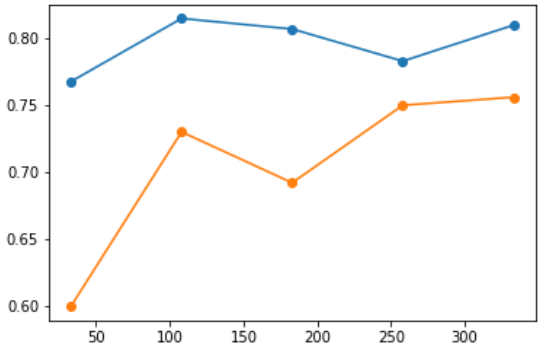
DecisionTreeClassifier
Train Accuracy – 80%
Test Accuracy – 68%
Best Parameters - max_depth = 3, min_samples_split = 90 , min_samples_leaf = 8 , criterion = gini index

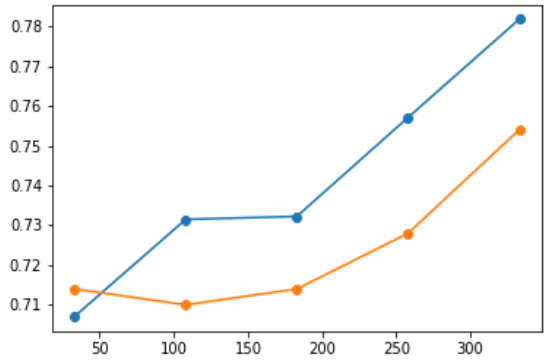
RandomForestClassifier

Train Accuracy – 78%
Test Accuracy – 72%
Best Parameters - max_depth = 10, min_samples_split = 45, min_samples_leaf = 8 , criterion = gini index, n_estimators = 15

From the above accuracies, we can observe that decision tree classifier tends to overfit on the training data as the accuracy on the training data in this case is 80% i.e it is much greater than test accuracy. On the other hand, random forest gives 78% accuracy (<80%) on training set and 72% on the test set i.e the gap between train and test accuracies is less as compared to decision tree classifier. The probable reason for this is the number of estimators or trees being used by the random forest which helps in generalizing more on the training data due to multiple hypotheses. Hence random forest is better.

b) The hyperparameters have been listed above. These have been found using a grid search with 3-fold cross validation. As grid search tries all possible combinations these hyperparameters give the best possible model. The grid search results have been attached.
c) The decision tree is an algorithm that has high chance of overfitting on the training data. This is the reason for a noticeable gap between train and test error in both cases.
However both models are learning well as can been shown from the following learning curves where the error on validation set decreases as no. of training samples is increased.




**Here the train accuracy is in blue and validation accuracy is in orange**.

d) The decision tree classifier shows a much higher variance as compared to random forest classifier. This is shows that random forest is trained better than decision tree and does not overfit on the training data.
5.8375352040639264e-05 -> Random Forest variance
0.000563848317370057 -> Decision Tree Variance

e) The classifier and grid search objects have been saved as pickle files.