MACHINE LEARNING
Assignment 4

Surabhi S Nath
2016271

A. GridSearchCV was used to find the best decision tree and random forest.
Parameters tunes: min_samples_leaf and max_depth
Cross validation = 10
Min_samples_leaf values = [1,2,5,8,10,12,15,20,25,50]
Max_depth values = [1,2,3,4,5,6,7,8,9,10]

- Decision Tree results
  Best parameters: {'max_depth': 5, 'min_samples_leaf': 8}
  Train Accuracy = 0.826
  Test Accuracy = 0.684

- Random Forest results
  Best parameters: {'max_depth': 10, 'min_samples_leaf': 5}
  Train Accuracy = 0.81
  Test Accuracy = 0.71

Random forest performs slightly better than decision tree. This could be because random forests take random bootstrap samples and train several decision trees using a subset of features and then takes a majority. We can also see that train accuracy for decision tree is higher than random forest. Decision trees are more prone to overfitting and perhaps therefore cannot generalize to new unseen samples.

The major differences between Decision Tree and Random Forest are the following:

Decision Tree	**Random Forest**
Higher variance	Lower variance
Faster evaluation	Slower evaluation
Easier to know decision rules	Harder to know decision rules

B. Hyperparameters used:

- Decision Tree: max_depth = 5, min_samples_leaf = 8

  This can be verified by printing the validation errors for every combination of possible parameters.

  [[0.714 0.714 0.714 0.714 0.714 0.714 0.714 0.714 0.714 0.714]
  [0.718 0.718 0.718 0.718 0.718 0.718 0.718 0.718 0.718 0.722]
  [0.728 0.732 0.734 0.734 0.73 0.73 0.736 0.746 0.734 0.736]
  [0.748 0.742 0.75 0.742 0.734 0.736 0.736 0.74 0.732 0.736]
  [0.73 0.732 0.734 **0.754** 0.74 0.73 0.736 0.746 0.734 0.736]
  [0.736 0.746 0.732 0.754 0.73 0.726 0.726 0.732 0.734 0.736]
  [0.718 0.716 0.714 0.744 0.738 0.724 0.724 0.73 0.736 0.736]
  [0.706 0.692 0.712 0.726 0.736 0.724 0.724 0.73 0.734 0.736]
  [0.704 0.686 0.71 0.738 0.734 0.718 0.724 0.726 0.734 0.736]
  [0.702 0.7 0.712 0.742 0.738 0.718 0.724 0.73 0.736 0.736]]

  It is seen the maximum validation accuracy indicates the choice of parameters. Since the parameters chosen maximize the validation accuracy, they are correctly chosen.

  max x index: 4, max y index: 3
  min_samples_leaf: 8, max_depth : 5

- Random Forest: max_depth = 10, min_samples_leaf = 5

  This can be verified by printing the validation errors for every combination of possible parameters.

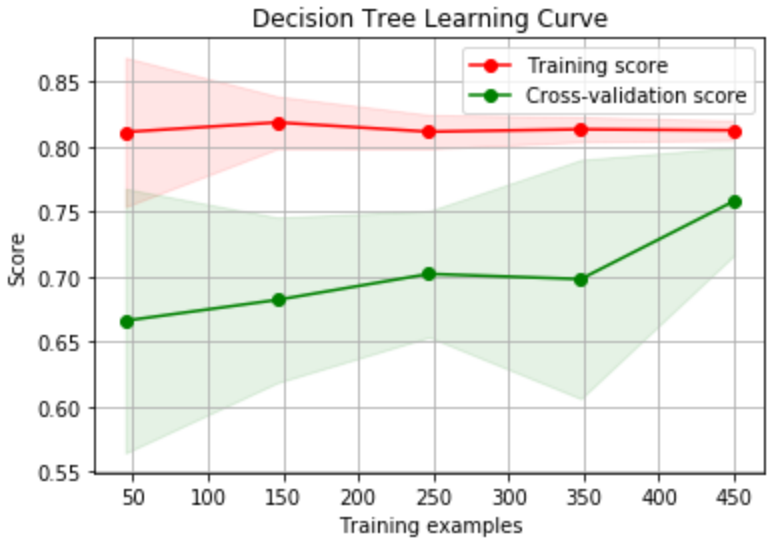
  [[0.712 0.718 0.712 0.718 0.712 0.714 0.714 0.714 0.714 0.714]
  [0.724 0.712 0.718 0.716 0.73 0.722 0.714 0.72 0.714 0.718]
  [0.734 0.728 0.734 0.748 0.734 0.73 0.728 0.726 0.718 0.714]
  [0.756 0.752 0.75 0.728 0.744 0.75 0.752 0.722 0.728 0.714]
  [0.758 0.762 0.758 0.742 0.758 0.736 0.736 0.742 0.736 0.714]
  [0.758 0.764 0.758 0.75 0.758 0.77 0.75 0.74 0.726 0.714]
  [0.742 0.746 0.774 0.756 0.748 0.746 0.748 0.73 0.74 0.716]
  [0.768 0.75 0.752 0.738 0.752 0.74 0.742 0.744 0.724 0.714]
  [0.746 0.748 0.778 0.768 0.77 0.75 0.744 0.744 0.726 0.714]
  [0.744 0.772 **0.784** 0.766 0.758 0.748 0.734 0.728 0.722 0.714]]

  It is seen the maximum validation accuracy indicates the choice of parameters. Since the parameters chosen maximize the validation accuracy, they are correctly chosen.

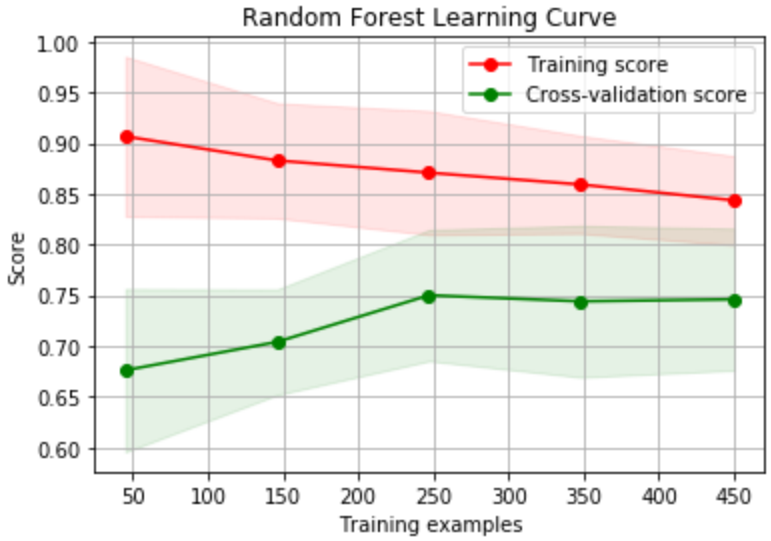
  max x index: 9, max y index: 2
  min_samples_leaf: 5, max_depth : 10

C. To show that the models are not overfitting or underfitting on the data, we can plot the learning curves.

- Decision Tree



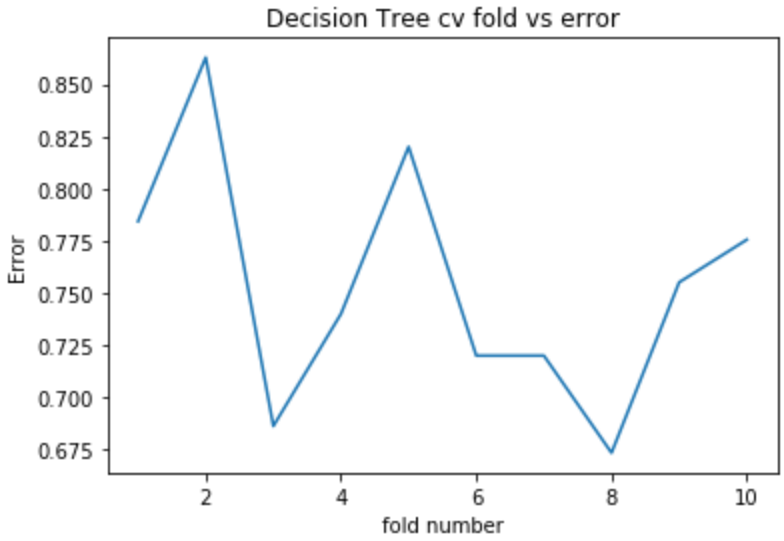
- Random Forest



We observe in both these cases that as number of training examples increase, the validation accuracy increases and training accuracy reduces. This means that the model is getting more and more generalized as the samples increase. Hence, for the chosen hyperparameter values, the model is not overfitting. Also, since training and test accuracies are both high, the models are neither underfitting.

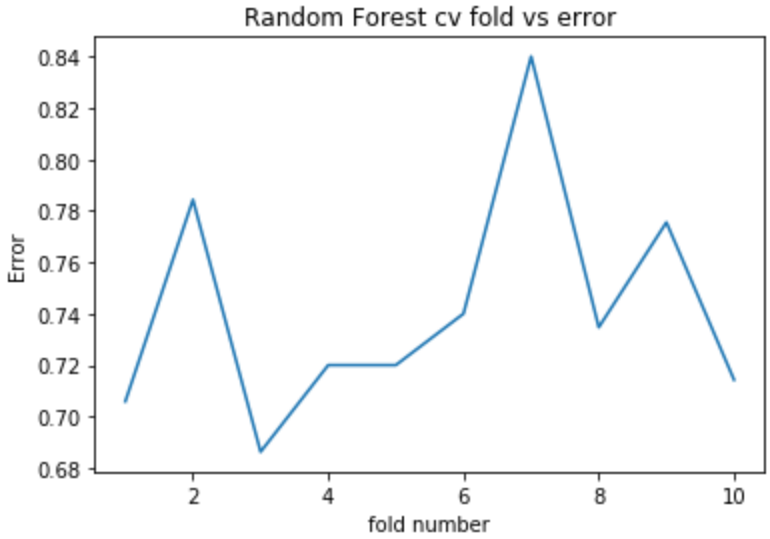
D. The validation errors for each of the 10 folds were calculated for both models.

- Decision Tree



The variance obtained was = 0.0031143574965000016

- Random Forest



The variance obtained was = 0.001872056865323159

Clearly, the variance of the validation errors for Decision Tree is greater than the variance for Random Forest. This is because the decision tree is more prone to overfitting thus has higher variance and lower bias.