# Assignment 2

Evan Jones Boddu, ejb180003@utdallas.edu

In this assignment, I have implemented three supervised learning methods which are Support Vector Machine (SVM), Decision Tree and Boosting on two different datasets. The main goal is to analyze and understand the behavior of the above-mentioned models on different datasets. The algorithms that are performed in this assignment are done using Sckitlearn library in Python Language (Jupyter Notebook).

## Dataset 1: Appliances Energy Prediction Dataset

The dataset which has been used in this assignment is taken from the previous assignment and the link to the dataset can be found here. Since I have already worked on this assignment, I will go ahead and show the implementations that have been performed using this dataset. The dataset has a continuous variable as its target variable (Y), hence the following preliminary steps have been performed on the dataset before building and running the models:

- Converted the target variable ( which is "Appliances") which is a continuous variable into a binary classification problem with values 0 and 1.
- I have put a threshold for the feature Appliances to convert it into a binary classification model. If Appliances is greater than 85, it is considered as 1 or else, if Appliances is less than 85, it is considered as 0.
- Split the dataset into 80:20 ratio.
- Trained the dataset on different supervised models to observe and pick the best model.

Further, I have selected all the features from the dataset so that the data will be as raw as possible and to get the best results when the dataset has been tweaked very little. I have checked for null values and assigned the target variable to 'y' and all the other independent variables to 'x'. The dataset consists of 19735 observations and 28 features.

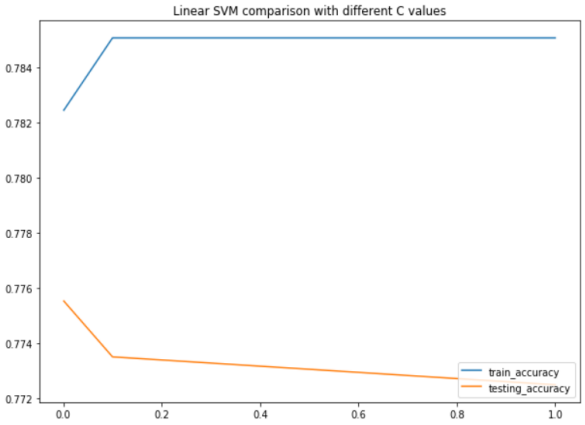
## Support Vector Machine

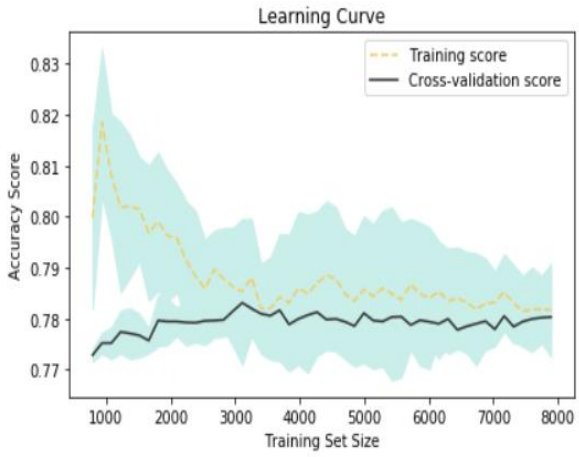
In this report for both the datasets, I have performed three different kernels on the datasets during the implementation and have selected the best kernel based on their results.

1.) **Linear SVM Kernel**:

The model was trained for different values of C [0.001,0.1,1] and the best accuracy for the test data was obtained at C =1 and plotted the learning curve and accuracy curve for training and testing set. Please follow the tables for measures;

| Linear SVM Kernel | C = 0.001 | C = 0.1 | C = 1 |
|---|---|---|---|
| Test Accuracy | 0.7907 | 0.7940 | 0.7948 |
| Confusion Matrix | [[2523 196]<br>[ 630 598]] | [[2474 245]<br>[ 568 660]] | [[2463 256]<br>[ 554 674]] |



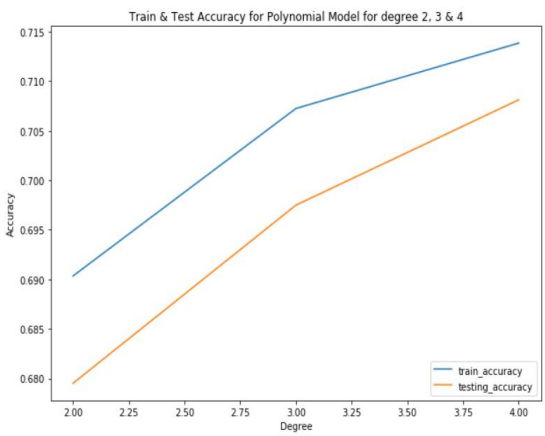

From the above figure, we can say that the training and testing set diverge i.e, the training accuracy and testing accuracy are opposite to each other and as the C values increases, the training accuracy increases while the testing accuracy decreases. For this dataset, the model with C = 1 gives the best accuracy of 79.48% with confusion matrix [[2463 256] [ 554 674]].

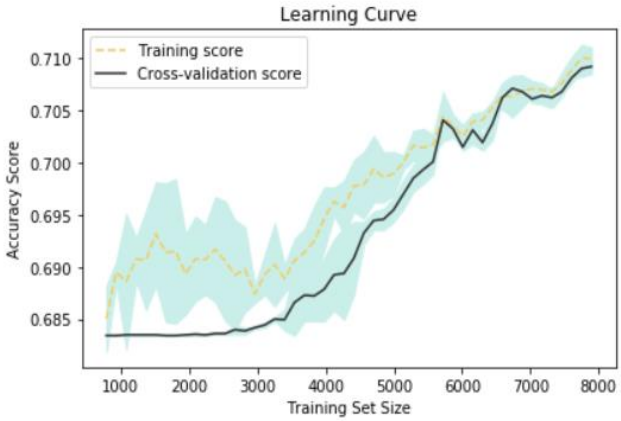
2.) **Polynomial SVM Kernel**:

The model was trained for different values of degree of polynomial and plotted the learning curve and accuracy curve for training and testing set. I have performed accuracy for this model for degree 2,3 and 4.

| Polynomial SVM Kernel | D.O.P = 2 | D.O.P = 3 | D.O.P = 4 |
|---|---|---|---|
| Test Accuracy | 0.6977 | 0.7173 | 0.7236 |
| Confusion Matrix | [[2706 13]<br>[1180 48]] | [[2627 92]<br>[1024 204]] | [[2581 138]<br>[ 953 275]] |

From the figure below, it is clear that the accuracy of both the training and testing dataset increases as the degree of polynomial increases from 2,3 and 4. For this dataset, the model with degree of polynomial 4 gives the best accuracy and it generalizes it well. The accuracy that is obtained is 72.36% with confusion matrix [[2581 138] [ 953 275]].



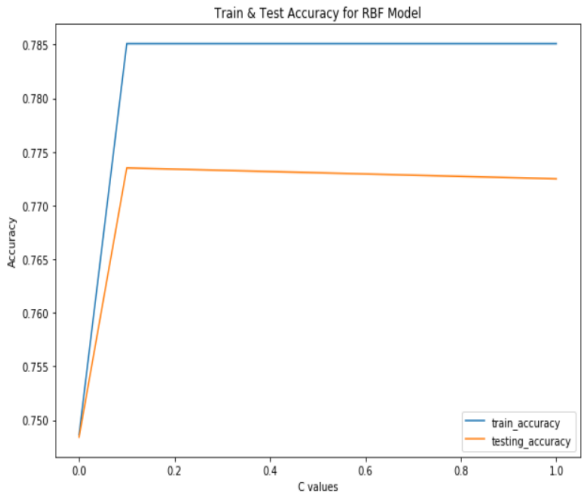



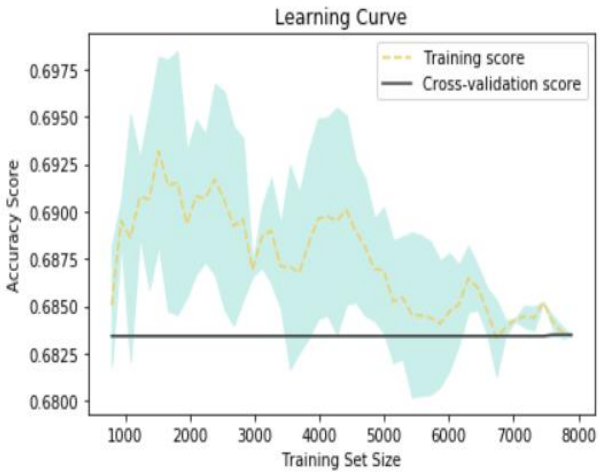

3.) **Radial Basis Function SVM Kernel**:

The model was implemented for different values of C [ 0.001,0.01,1] and plotted the corresponding train and test accuracy for different values of C. There is not much of difference in the accuracy of the train and test set until 0.770 and then it diverges.

| RBF SVM Kernel | C = 0.001 | C = 0.01 | C = 1 |
|---|---|---|---|
| Test Accuracy | 0.6888 | 0.6888 | 0.6922 |
| Confusion Matrix | [[2719 0]<br>[1228 0]] | [[2719 0]<br>[1228 0]] | [[2716 3]<br>[1212 16]] |

From the figure below we can say that, the test and train accuracy has no (or little) difference until 0.770 and then it diverges. This model performs best on the dataset with an accuracy of 69.22% and confusion matrix [[2716 3] [1212 16]] when C = 1.

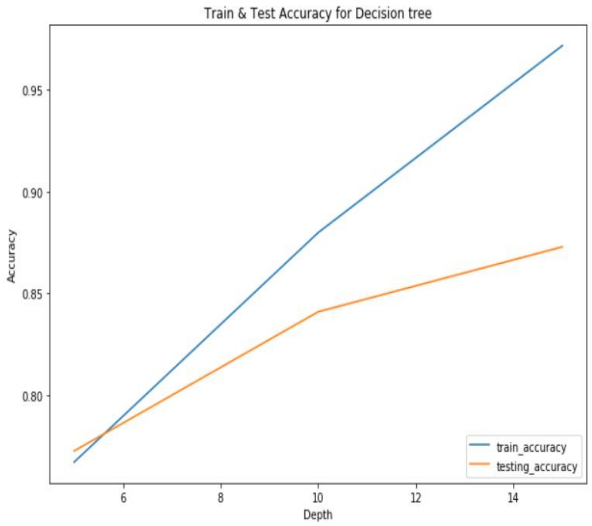


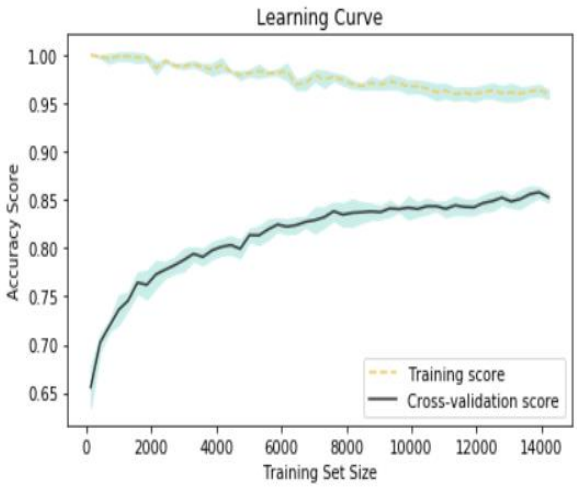



## Decision Tree

Decision tree is another algorithm which helps to overcome the amount of time that is taken by the other models for computing the values . While training the data, the tree grows to its extreme which will in turn result in high variance. In order to avoid all this, pruning is used in the model which will help to decrease the variance and makes it perform better when subjected to unknown values.

| Decision Tree | Depth = 5 | Depth = 10 | Depth = 15 |
|---|---|---|---|
| Test Accuracy | 0.7494 | 0.8305 | 0.8637 |
| Confusion Matrix | [[2421 298]<br>[ 691 537]] | [[2478 241]<br>[ 428 800]] | [[1916 832]<br>[ 841 358]] |

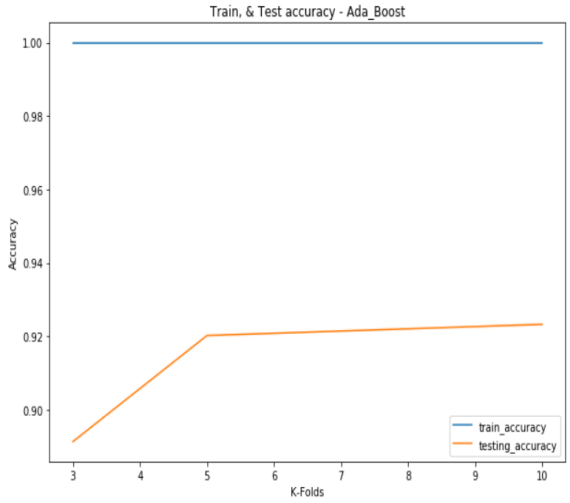


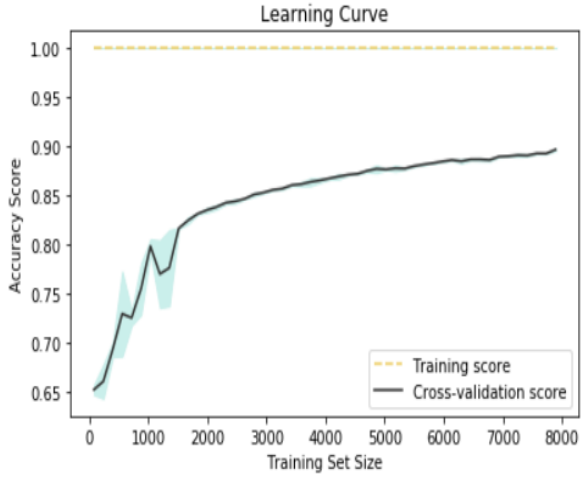




I have implemented the entropy approach for the training data by pruning the trees with depth as 5, 10 and 15. The figure below shows the linear increase in the accuracy of the training dataset as the dept increases while the test accuracy increases slowly after depth 10 . The best accuracy is obtained for the tree with maximum depth 15 showing an accuracy of 86.37% and confusion matrix is [[1916 832] [ 841 358]].

## AdaBoost Algorithm

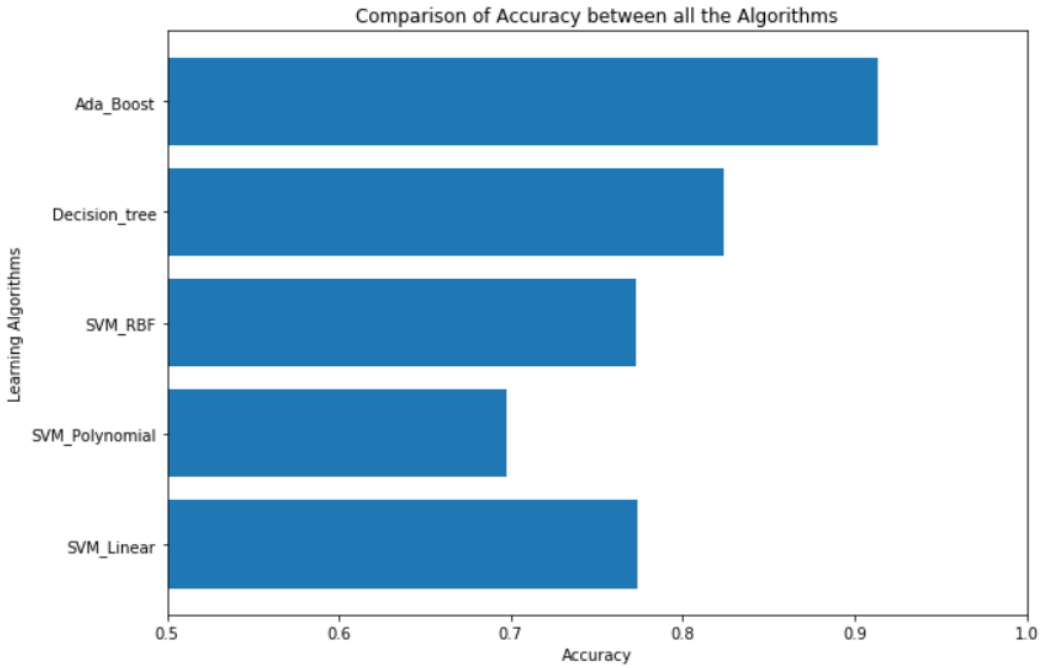
AdaBoost Algorithm uses a base learner which is a Decision Tree classifier and if we do not prune the dataset then the model will have high variance and may result in 100% accuracy. In order to avoid this issue, we prune the tree to get the best results possible. Pruning is done with three values 5,10 and 15. The best results are obtained for the model with the pruned value as 15.




The figure above shows that the accuracy increases as the depth increases in the case of testing set. We may say that this is the case of high variance of the model. When the pruned value is at 15, the accuracy of the model is 91.95%.

## Overall Performance Comparison




The figure above shows the accuracy of all the models and algorithms that have been used in this assignment on the test data. As you can seem AdaBoost Algorithm on Decision Tree has the highest accuracy which is followed by Decision Tree and Linear SVM kernel & RBF SVM kernel and the least accuracy model available in this assignment is Polynomial SVM kernel. For the dataset 2, the best model is RBF SVM kernel with a degree of polynomial 59.35%.

## Dataset 2: Heart Disease Dataset

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. You can find the link to the dataset here.

I chose this dataset because heart diseases occur very frequently everywhere, and I would like to predict and get to know what factors affect heart. Performed the same models and algorithm which are used in dataset 1.

## Support Vector Machine

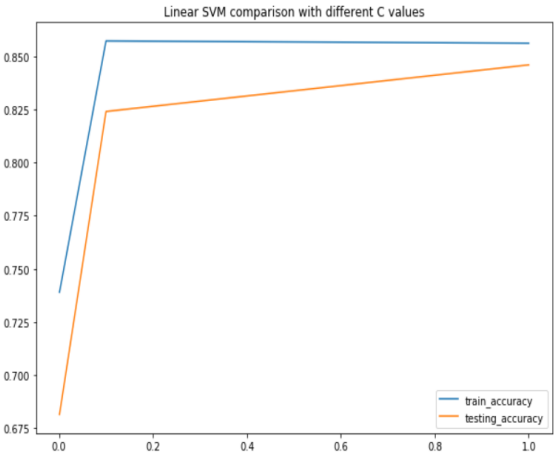
A Support Vector Machine (SVM) is a discriminative classifier which is formally defined by a separating hyperplane. The SVM classifies the data using the support vectors which are the data points which are located close to the classifying line.

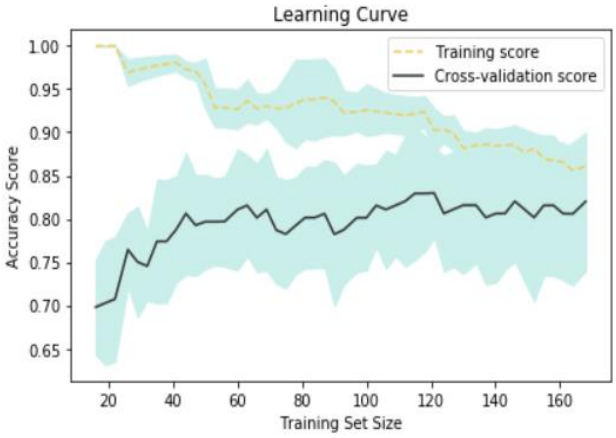
1.) **Linear SVM Kernel:**

For linear model, the data has been trained for three different values of C [0.001,0.1,1]. For better understanding of the increase in the accuracy, a plot of training and testing values has been plotted for different values of the hyperparameter C.

| Linear SVM Kernel | C = 0.001 | C = 0.1 | C = 1 |
|---|---|---|---|
| Test Accuracy | 0.6813 | 0.8242 | 0.8462 |
| Confusion Matrix | [[25 17]<br>[12 37]] | [[30 12]<br>[ 4 45]] | [[31 11]<br>[ 3 46]] |

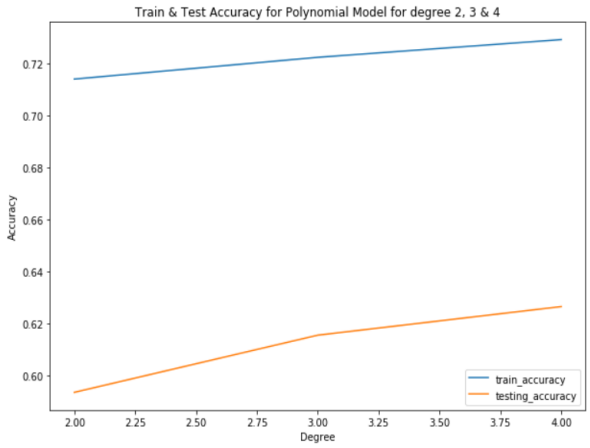
From the figure below, we can say that as the value of hyperparameter C = 1 increases, the accuracy of both testing and training set increases. For the value of C = 1, the model performs best with an accuracy of 84.62% with confusion matrix [[31 11] [ 3 46]].

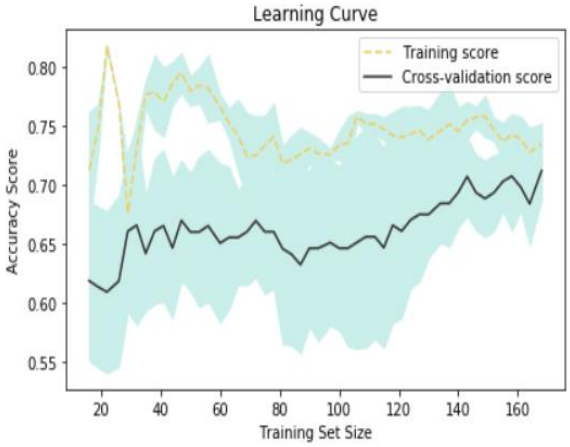


2.) **Polynomial SVM Kernel**:

This model has been implemented and trained over different degree of polynomial [5, 10, 15] and the accuracy is reported for the best performing model along with the confusion matrix, and also the learning curve for Polynomial SVM kernel is also reported with both testing set and training set with accuracy on y axis and training size on x axis.

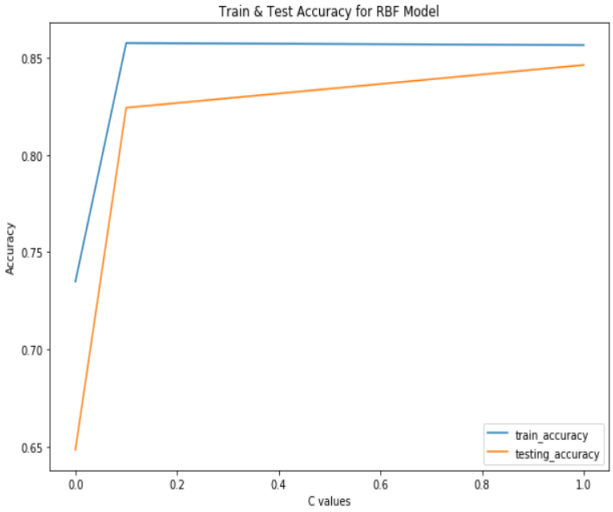


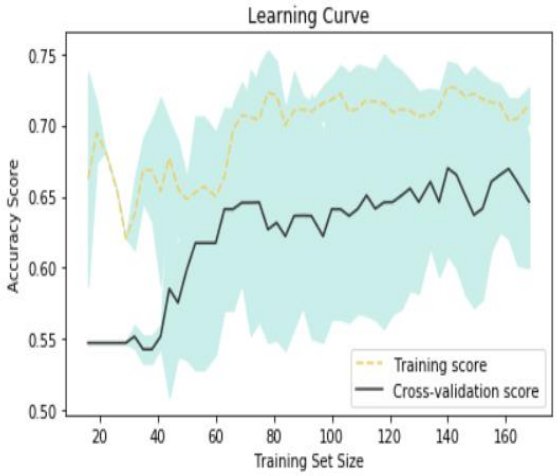




From the plot, which is reported above, we can see that there is high variance in the data and as the degree of polynomial increases, the accuracy also increases. There was a steady increase in the testing curve until degree of polynomial 3 and then it increased slightly when the D.O.P increases.

3.) **Radial Basis Function SVM Kernel**:

RDF SVM Kernel is the best accuracy for the dataset in all aspects in this dataset. This model has been trained over different values of C [ 0.001, 0.1, 1]. With the increase in values of C, the accuracy increases.

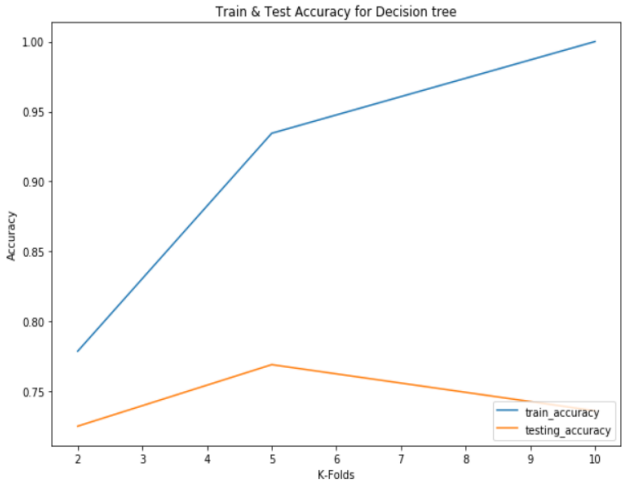


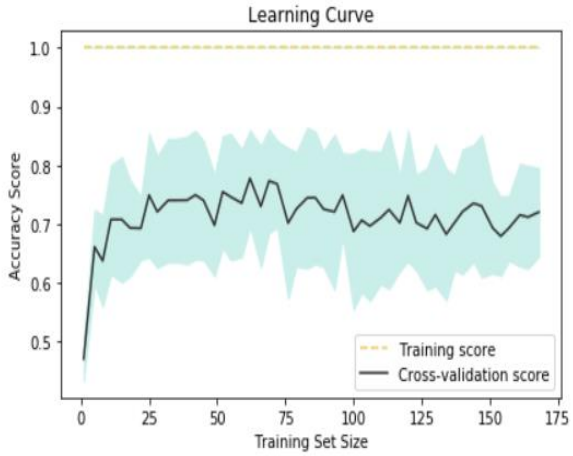



From the above plot we can say that as the values of hyperparameter C increases, the accuracy of the model also increases. The model has the best accuracy of 59.34% and confusion matrix [[16 26] [11 38]].

## Decision Tree

A decision tree is a decision support tool that uses a tree model of decisions and their possible consequences. The data was trained for Entropy with depth of different values 5,10,15, cause entropy model splits the data based on information gain results in better decision making. I have provided the accuracy plot as well as the learning curve for testing data and training data.

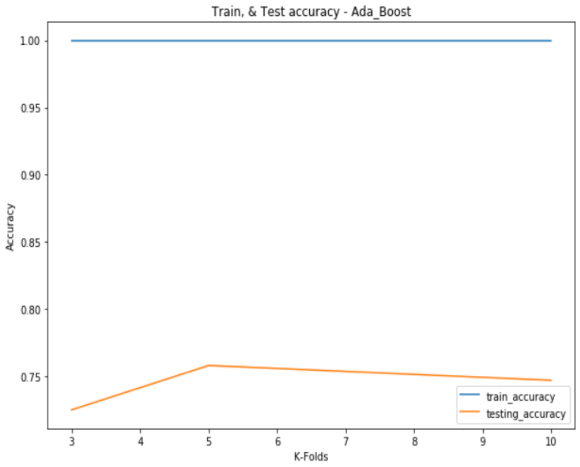


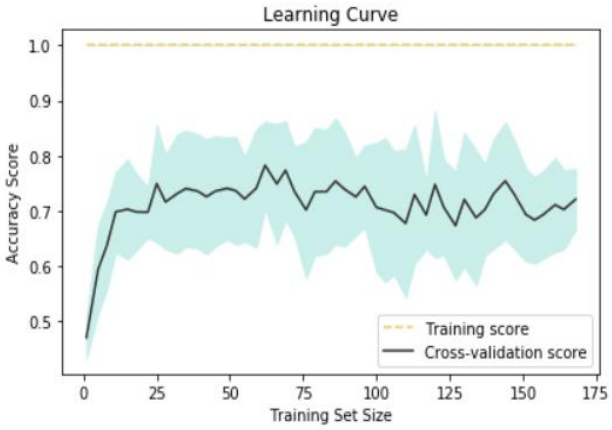




From the above graph, we can say that as the depth of the tree increases, the accuracy of the model also increases. There is a steady increase in both of the curves respectively until depth 5 and then the accuracy of the testing curve decreased while the accuracy of the training curve increased. The model has an accuracy of 76.93% with confusion matrix [[32 10] [13 36]] when the depth of the tree is 5.

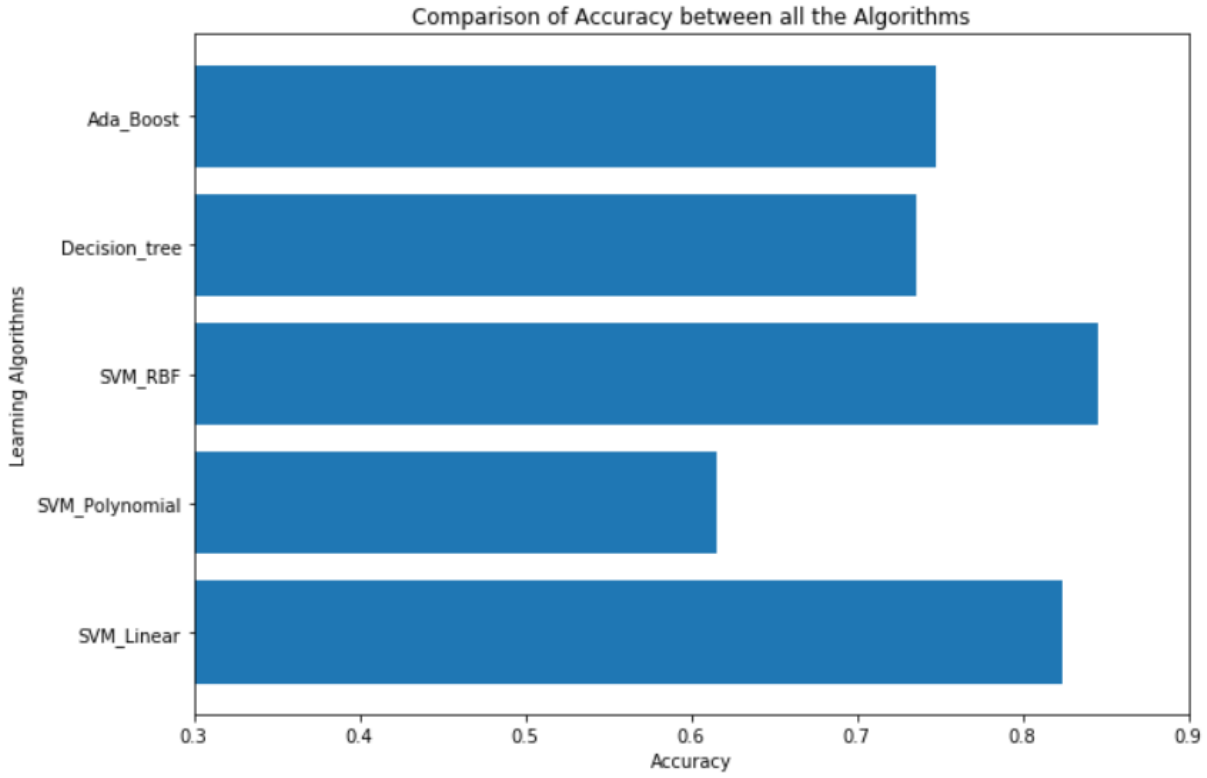
## AdaBoost Algorithm

I used AdaBoost algorithm to implement ensemble supervised learning on the dataset. AdaBoost has been used on Decision Tree in this model, to decrease the error of the model and make the model perform better. For this I have used decision tree with different values of depth [5,10,15] and it is observed that the tree which is pruned 5 has the best accuracy and gave promising results for this model. The accuracy which was predicted of the tree that is pruned with a depth of 5 is 75.82%. From the plot, you may see that the training accuracy of the model has 100% accuracy, this can happen because of the boosting weights.




## Overall Performance Comparision




The figure above shows the accuracy of all the models and algorithms that have been use in this assignment and the ones which gave the best results on the test data. As you can interpret from the figure that RBF SVM Kernel has performed best out of all the other models and algorithm which is followed by Linear SVM Kernel.

## Conclusion

From the above observations, we can conclude that for Dataset 1, the best accuracy is obtained by the model Decision Tree and similarly for the second dataset 2, the best accuracy is obtained from the model RBF SVM Kernel. This assignment has given me a broader view on how to handle the data and how to use the data in a better method using certain models. Overall, it was great to find out the results and work on this assignment. Both the datasets were not only diverse in their data but were challenging as well.