Assignment 2 – sxb180048 – Sai Pratheek Banda

ENERGY DATASET

Classification Problem:

The aim of the problem is to predict the classification of the energy consumption into high and low. The interesting part of this classification is that it will help us understand the patterns for energy consumption thus understanding why energy consumption might be low or high. **Class 0 (Low)**: 10357 Observations| **Class 1 (High)**: 8048 Observations

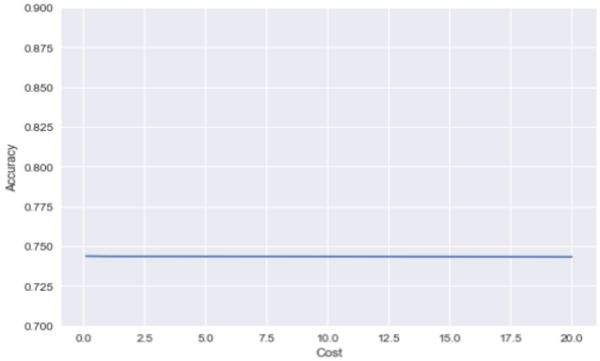
EDA:

1) Features 'T_out, RH_7, T9, RH_4, T3, rv2, RH_9, T1, T7 ' were dropped due to high correlation.
2) Date, Visibility features are insignificant due to its inability to explain the target variable.
3) Lights Feature was removed due to high null values.
4) The outliers were removed based on the Inter Quartile Range.

Support Vector Machine (Linear):

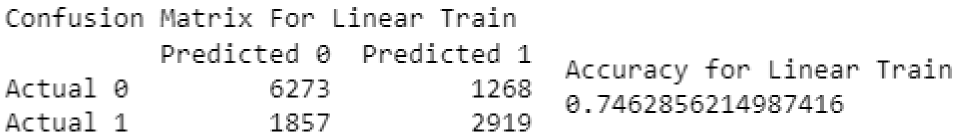
Parameters:

To linearly separate the data the **Regularization parameter** 'C' which, for a high value chooses a smaller-margin hyperplane resulting in better data point classification.
Implementing the **Cross Validation with 3-folds** I found that there is not significant change in the Accuracy Metric.
Thus, **default value of C = 1** was chosen to run the model.
From which we can conclude that the data was linearly separated with the lowest significant margin.



With the default regularization Parameter SVM is fit and predicted on train data without Cross Validation to compare the Accuracy metric and have a base line to check with the test data.

As shown below in the metric. We can observe an Accuracy of 74.6% on the train data.



```
Confusion Matrix For Linear Train
          Predicted 0  Predicted 1
Actual 0         6273         1268
Actual 1         1857         2919
```

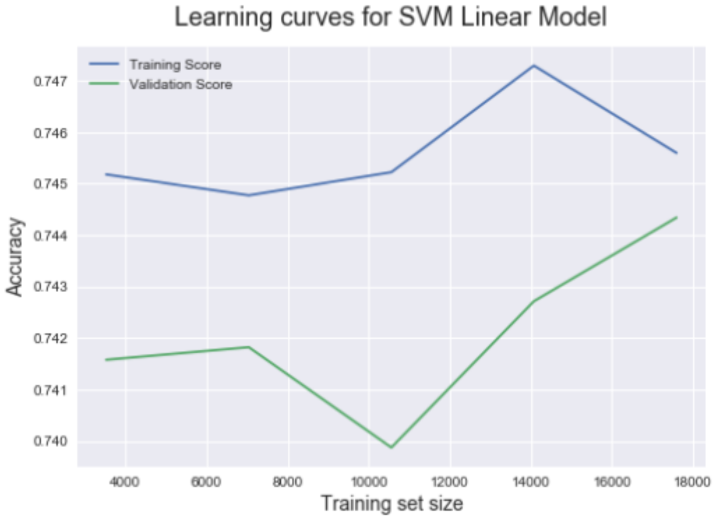
```
Accuracy for Linear Train
0.7462856214987416
```

Learning Curve:

Learning Curve observes the change in metrics as with an increase in the size of data given for training data.

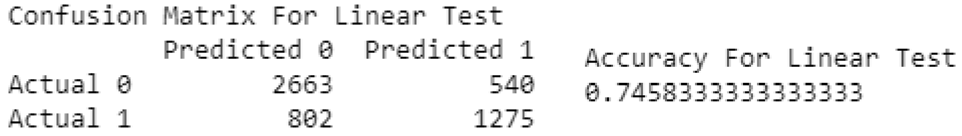
For the above model we plot a training curve which results in increase in accuracy up to 80% of train data but decreases with 100% of train data, but for the validation data metric increases as it understands the data better with increase of influx training data.

As learning curve implements cross validation, we can conclude that the decrease in accuracy in training is the model trying not to overfit thus generalizing well.



Results on Test Data:

The model tests well on the test data with an accuracy metric of 74.58% which implies that the data can be linearly seperated but a more accurate result is possible.

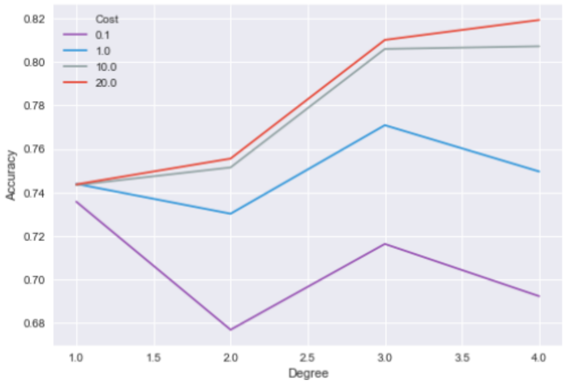


```
Confusion Matrix For Linear Test
          Predicted 0  Predicted 1
Actual 0         2663          540    Accuracy For Linear Test
Actual 1          802         1275    0.7458333333333333
```

Support Vector Machine (Polynomial):

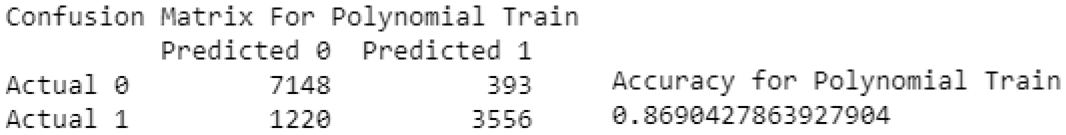
Parameters:

Polynomial Seperation of the data points is experimented with multiple Regularization Parameter and multiple Degree of the polynomial to find the the combination of the best parameter.
Cross Validation is implemented to find the best results which were found with C =20 & Degree of the Polynomial 4



With the best parameters for polynomial separation, SVM is fit and predicted on train data without Cross Validation to compare the Accuracy metric and have a base line to check with the test data.

We can observe an Accuracy of 86.9% on the train data, which is a significant improvement from the linearly separated model. Hinting that the data may be explained better with a polynomial rather than a linear equation

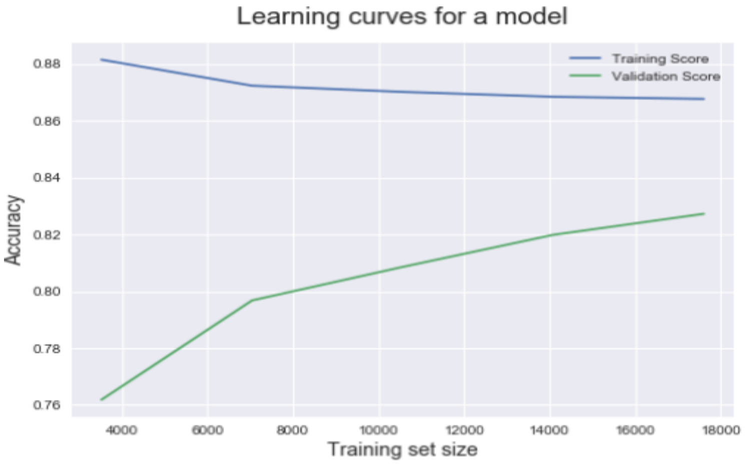


```
Confusion Matrix For Polynomial Train
          Predicted 0  Predicted 1
Actual 0         7148          393    Accuracy for Polynomial Train
Actual 1         1220         3556    0.8690427863927904
```

Learning Curve:

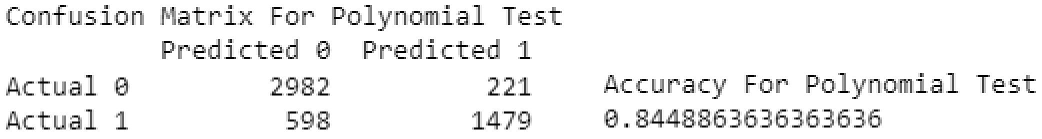
For the model we plot a training curve which results in a gradual decrease in accuracy with the train data but a simultaneous increase in the validation data with increase of influx training data.

As learning curve implements cross validation, we can conclude that the model is learning better with increase in data and is generalizing well with an 86.7 % on train and 82.7% on validation data.



Test Data:

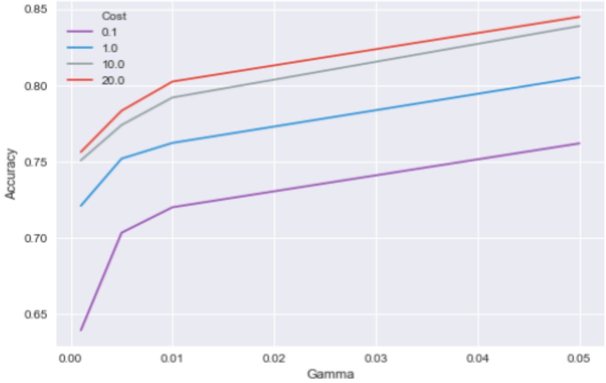
The model tests well on the test data with an accuracy metric of **84.48%** which implies that the data can be polinomialy seperated signifcantly better than linear.



```
Confusion Matrix For Polynomial Test
          Predicted 0  Predicted 1
Actual 0         2982          221    Accuracy For Polynomial Test
Actual 1          598         1479    0.8448863636363636
```

Support Vector Machine (Radial):

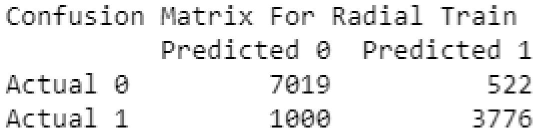
Parameters:



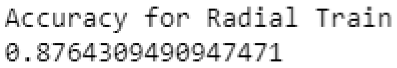
Non- Linear Seperation of the data points is experimented with the combination of multiple Regularization Parameter and multiple Gamma values which with increase in its value data points closer to the seperation line are used and vice versa.
Cross Validation is implemented to find the best best parameters on the metric of accuracy which were found with **C =20 & Gamma = 0.05**

With the best parameters for non-linear separation, SVM is fit and predicted on train data without Cross Validation to compare the Accuracy metric and have a base line to check with the test data.

We can observe an **Accuracy of 87.6% on the train data**, which is an improvement from the polynomial separated model. Hinting that the data may be **explained better with non-linear rather than a linear or polynomial**

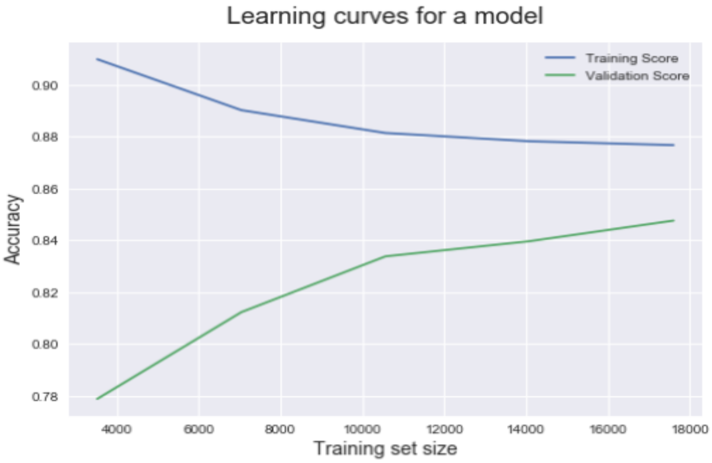


```
Confusion Matrix For Radial Train
          Predicted 0  Predicted 1
Actual 0         7019          522
Actual 1         1000         3776
```



```
Accuracy for Radial Train
0.8764309490947471
```

Learning Curve:



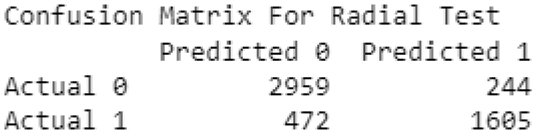
With increase of influx of data there is a gradual decrease in accuracy with the train data but a simultaneous and significant increase in the validation data.

As learning curve implements cross validation, we can conclude that the model is learning better than any other Kernels of SVM.

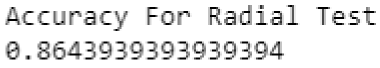
With increase in data the model converges and generalizing well with an **87.8 %** on train and **84.8%** on validation data.

Test Data:

This model is the most accurate among all the kernels of the Support Vector Machines with an accuracy of 86.4% , highest f-1 score on test data.

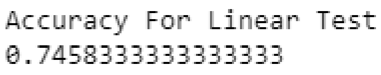
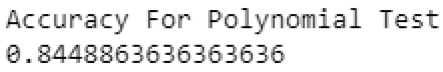
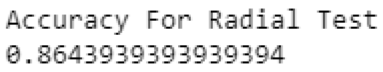


```
Confusion Matrix For Radial Test
          Predicted 0  Predicted 1
Actual 0         2959          244
Actual 1          472         1605
```



```
Accuracy For Radial Test
0.8643939393939394
```

Comparison of Support Vector Machine Kernels:

 Accuracy For Linear Test 0.7458333333333333	 Accuracy For Polynomial Test 0.8448863636363636	 Accuracy For Radial Test 0.8643939393939394
F-1 Score: 66%	F-1 Score: 78%	F-1 Score: 82%

Comparing the three kernels we can conclude that the radial kernel was the most accurate of them all with the highest 'Precision', 'F-1' and 'Accuracy' Metrics. You can imply that the data isn't linearly separable nor can the data be explained with the greatest accuracy with a polynomial separation. Non – Linear separation is the best Support Vector Machine model which generalizes the best among the lot.

Decision Tree:

Parameters:

The feature selection is not significant in case of Decision Trees as Entropy is calculated for each feature and thus selects the best ones to split on

The 2 sets of Parameters which were experimented on are:

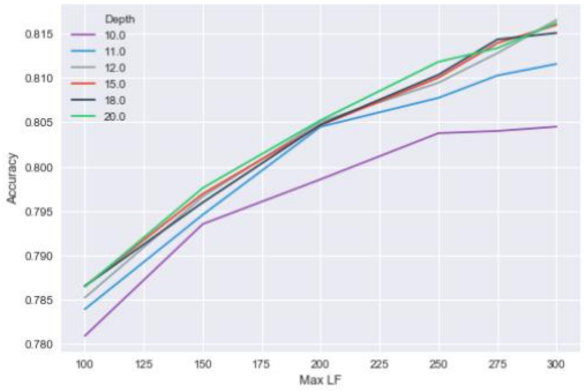
- Depth and Maximum number of Leaf Nodes
- Depth and Percentage of Maximum Features

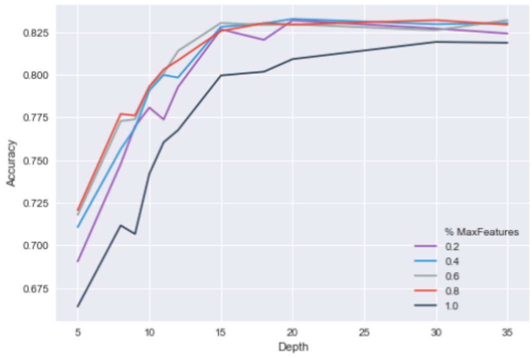
Cross Validation is implemented to find the best parameters.

In the first set of parameters the Depth = 15 and Maximum number of Leaf nodes = 275

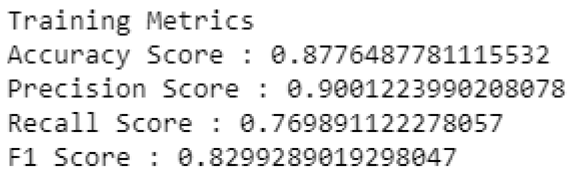
In the Second Set of parameters the Depth =15 and Percentage of Maximum Features = 0.6

A combination of these parameter is used to find the best decision tree.





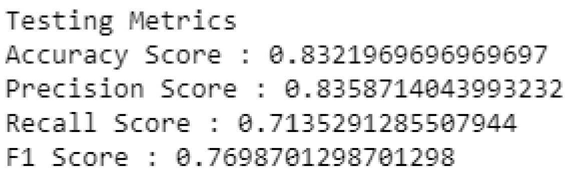
With the best parameters of a decision tree is fit and predicted on train data without Cross Validation to compare the Accuracy metric and have a base line to check with the test data. As shown below in the metric, we can observe an Accuracy of 87.7% on the train.



```
Training Metrics
Accuracy Score : 0.8776487781115532
Precision Score : 0.9001223990208078
Recall Score : 0.769891122278057
F1 Score : 0.8299289019298047
```

Test Data:

The decision tree did not perform as well as the radial kernel did but has an Accuracy of 83.2%

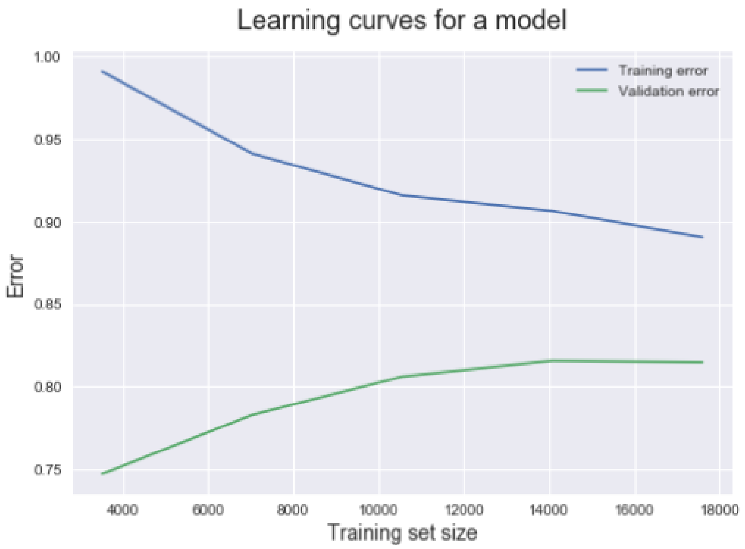


```
Testing Metrics
Accuracy Score : 0.8321969696969697
Precision Score : 0.8358714043993232
Recall Score : 0.7135291285507944
F1 Score : 0.7698701298701298
```

Learning Curve:

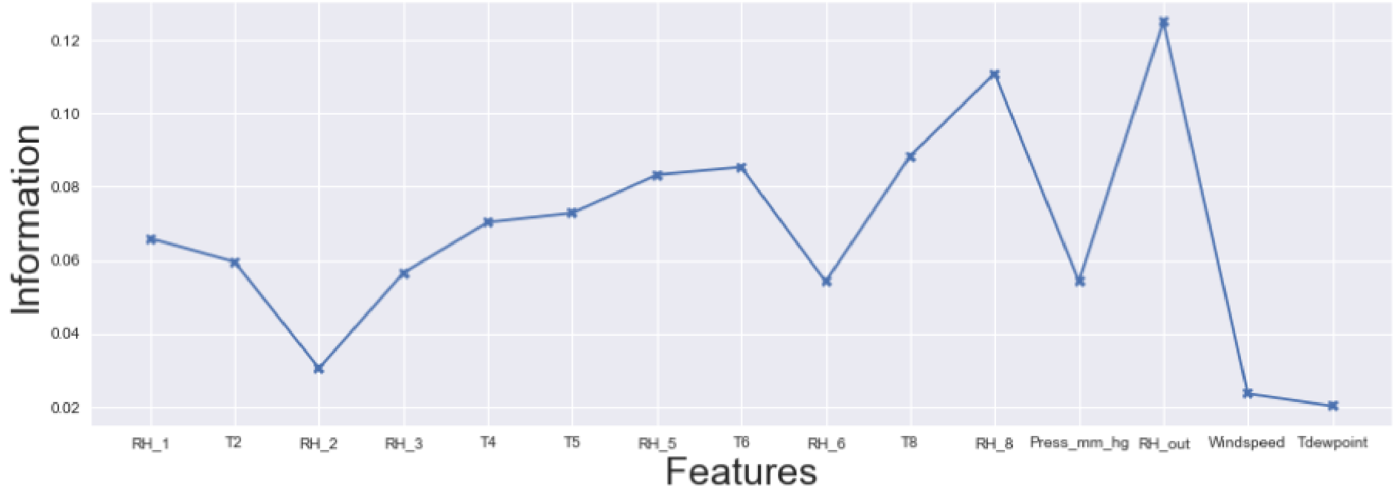
As the data influx increases the train and test scores converge, which implies that the model is generalizing well.

The final Training and Validation accuracy metrics are 89% and 81% respectively.



Information Gain:

The following graph shows the most information gain with respective to features and RH_out being the most important.



Boosting:

Parameters:

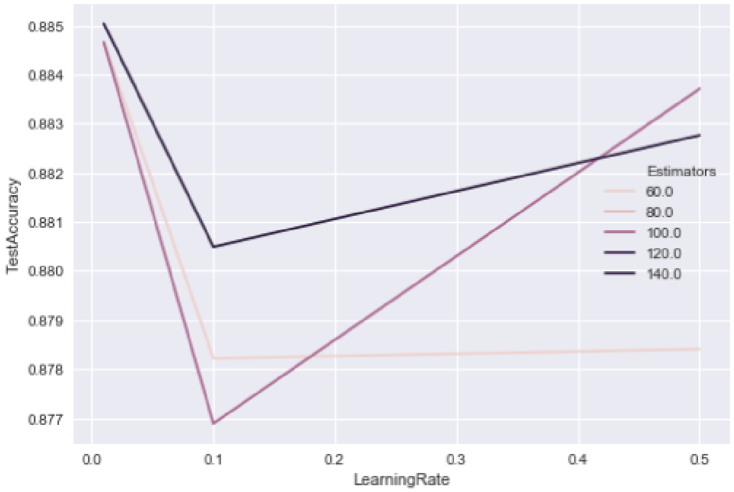
Boosting is done to the above decision tree for an ensemble method.

The parameters experimented with are the learning rate and estimators.

Estimators are the weak learns we need to boost and improve the accuracy of the required classification.

The best parameters are Estimators: 100

And Learning Rate: 0.5



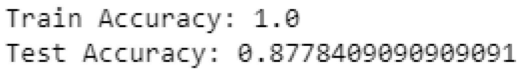
Learning Curve, Train and Test:

With increase in the influx of training data the train score is 100% and the validation increases but shows a very low accuracy score.

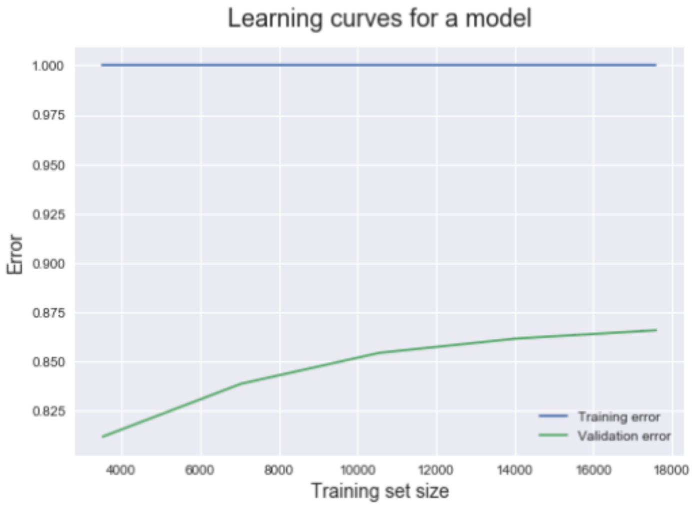
Thus, there may a point of overfitting the data.

The below measures also show the same without cross validation.

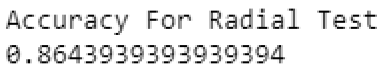
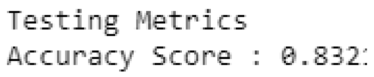
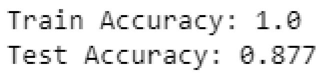
Yes, booting does improve the records but it might be prone to overfitting the model.



```
Train Accuracy: 1.0
Test Accuracy: 0.8778409090909091
```



Comparison of The Different Models:

SVM	Decision Tree	Ensemble Method(Boosting)
A tedious preprocessing which includes scaling the features. Classification of the data is done with decent accuracy without overfitting to the noise of the data They generalize well over the data	The Decision tree is the fastest of all the algorithms when performed and does an easy job in classification and has an accuracy better than some kernels of SVM.	Boosts the accuracy of the decision tree to achieve a train accuracy of 100% which might results in overfitting the train data but does well on the Test data.
 Accuracy For Radial Test 0.8643939393939394	 Testing Metrics Accuracy Score : 0.832	 Train Accuracy: 1.0 Test Accuracy: 0.877