EAS 508 Homework – 3

Name: Venkata Satya Surya Sai Vineet Atyam

UB Person number: 50419767

UBIT Name: vatyam

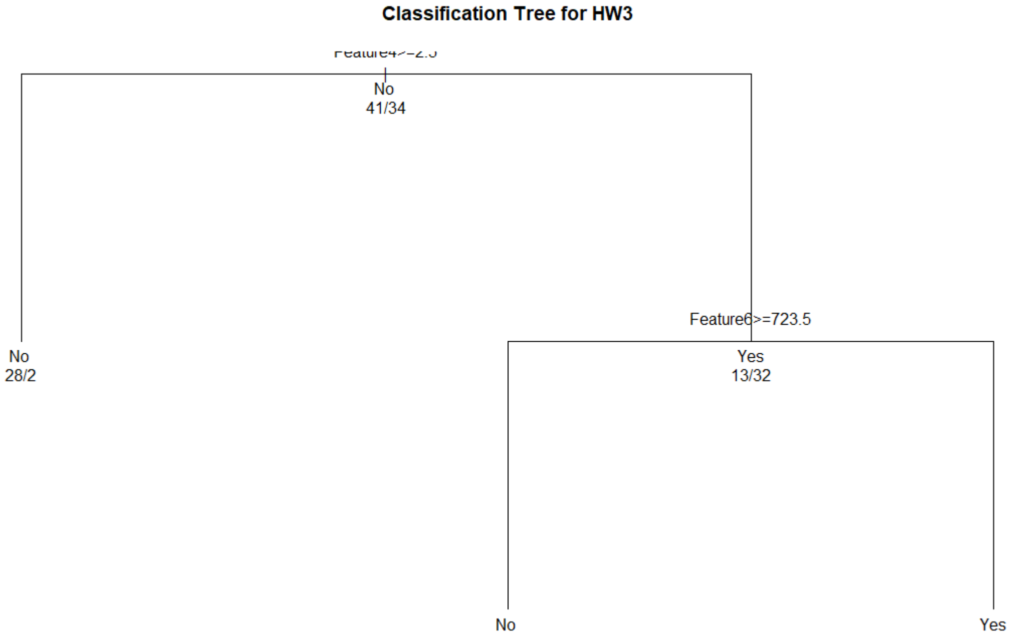
MODEL	TRAIN ACCURACY	TEST ACCURACY
DECISION TREE	0.92	0.96
BAGGING	0.90	1
RANDOM FOREST	0.92	1
BOOSTING	0.882	1

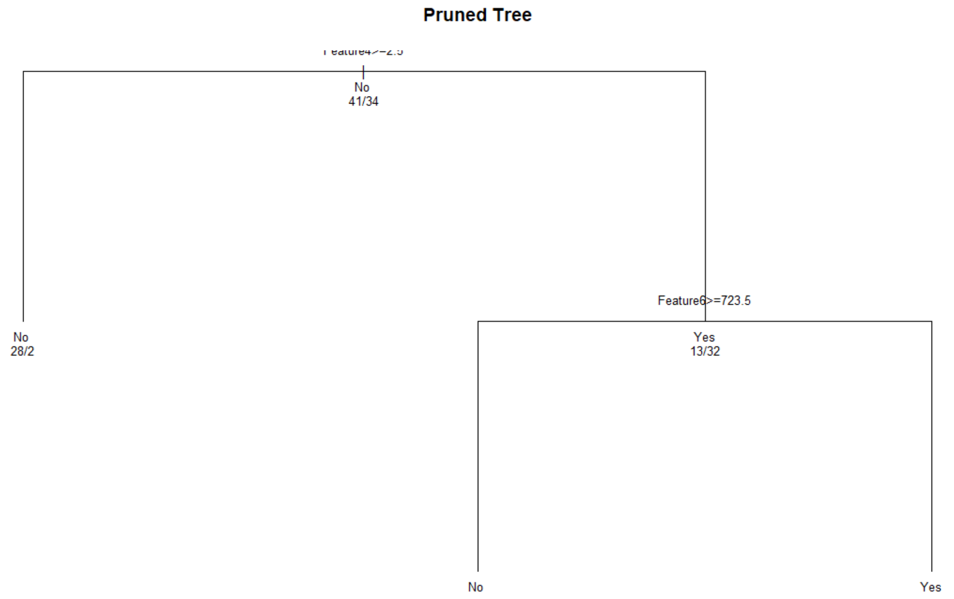
1.) The Model classification has been done on the basis of Property1 where the data is split into ‘Yes’ and ‘No’ based on the mean of the data, hence the given data has about 48 data points above the mean and 52 data points below the mean.

DECISION TREE

Feature 4 and Feature 6 were used to construct the decision tree with a high accuracy and a more robust model.

There is no difference with the pruning as only 2 features were used.

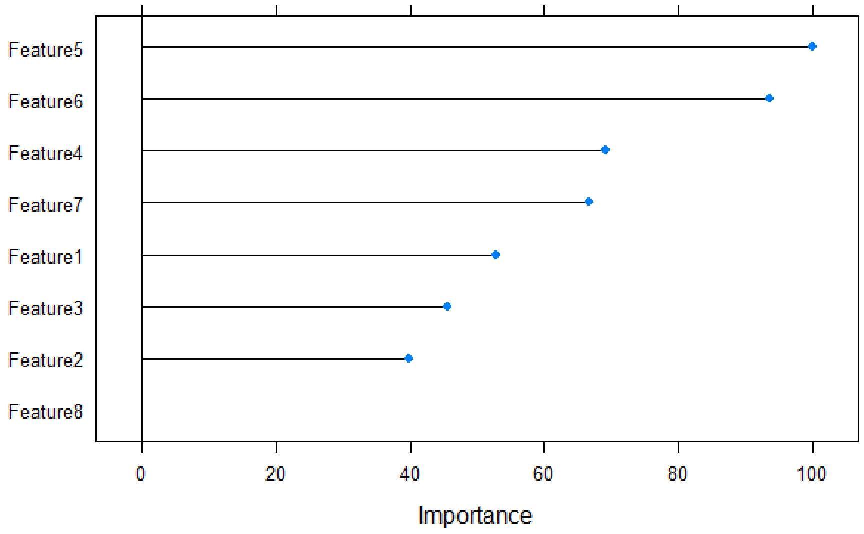




BAGGING

Feature Importance:

Feature5 100.00 , Feature6 93.76 , Feature4 69.08 , Feature7 66.86 , Feature1 52.91 , Feature3 45.49 , Feature2 39.83 , Feature8 0.00



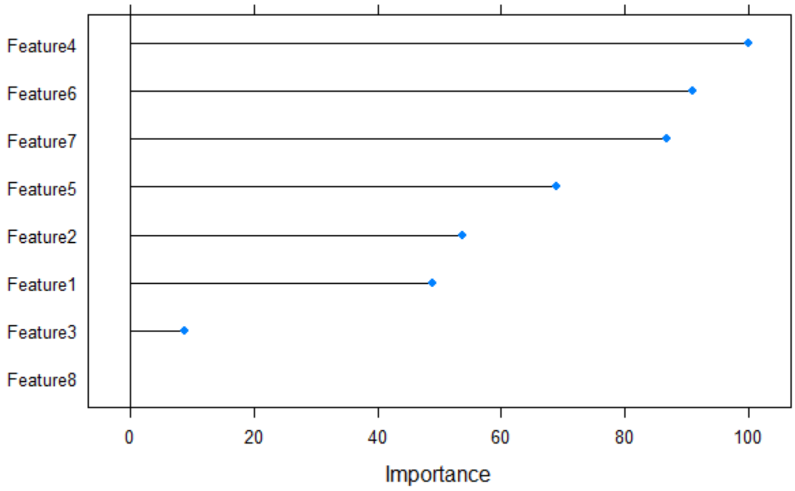
In the bagging model, nbagg can be changed but since the model is very highly dependent on few features, there is no change in the accuracy with the increase in the nbagg value.

Accuracy	Kappa
0.9053571	0.8082319

RANDOM FOREST

Random Forest Variable importance

Feature4 100.000 , Feature6 90.965 , Feature7 86.796 , Feature5 68.951 , Feature2 53.721 , Feature1 48.799 , Feature3 8.637 , Feature8 0.000



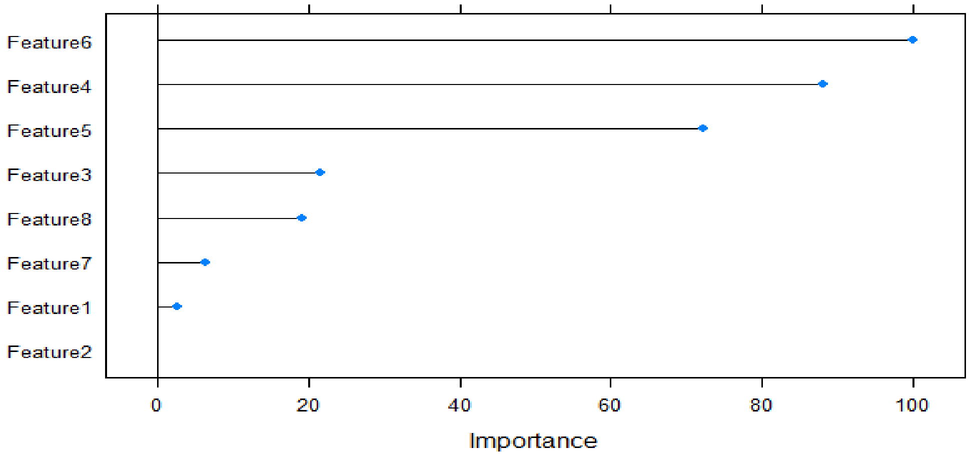
The features have different importance as compared to the bagging model and has better accuracy than the bagging model. The tuned model prevents over fitting better and hence works better for test dataset predictions.

The final value used for the model was mtry = 2.

BOOSTING

gbm variable importance

Feature6 100.000 , Feature4 88.119 , Feature5 72.299 , Feature3 21.582 , Feature8 19.150 , Feature7 6.376 , Feature1 2.510 , Feature2 0.000



The final values used for the model were n.trees = 100, interaction.depth = 1, shrinkage = 0.1

2.) **Random Forest Regression**

The Random Forest Regression using the default mtry value turned out to be 0.72, and when the value was changed to mtry = 6, there was a considerable change in rmse value to 0.81., which means the model has been tuned for better accuracy and robustness.

RMSE : 0.81021 (mtry = 6)

RMSE : 0.7201 (mtry = 2 default)

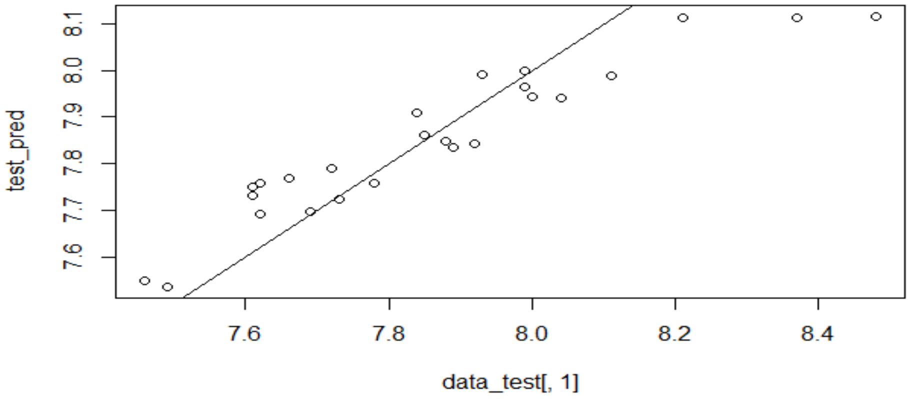
Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 6

Mean of squared residuals: 0.005417074

% Var explained: 80.04



There is a change in accuracy when comparing classification and regression because in classification, anything in a given tolerance is classified under the same category, whereas in case of regression, the values are to be found specifically and hence results in more errors present.