Report Homework 4

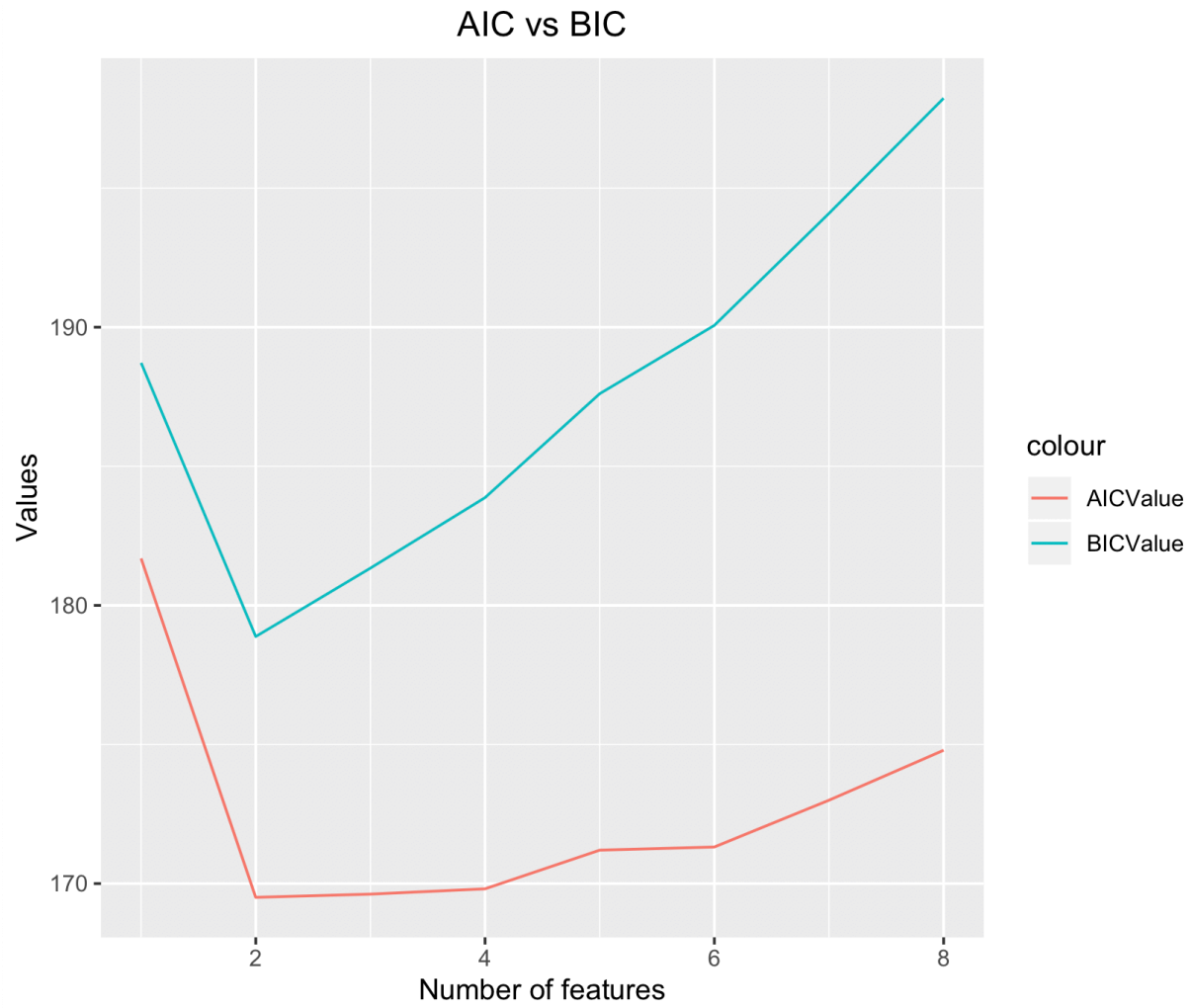
Problem 1:

In this problem, the prostate dataset is given to perform a best linear regression analysis. And, compute AIC, BIC and five fold and ten fold cross validation and bootstrap .632 estimates of prediction error.

After loading data, did a sanity check over data to have NA values in dataset or not and divided data into training and testing set.

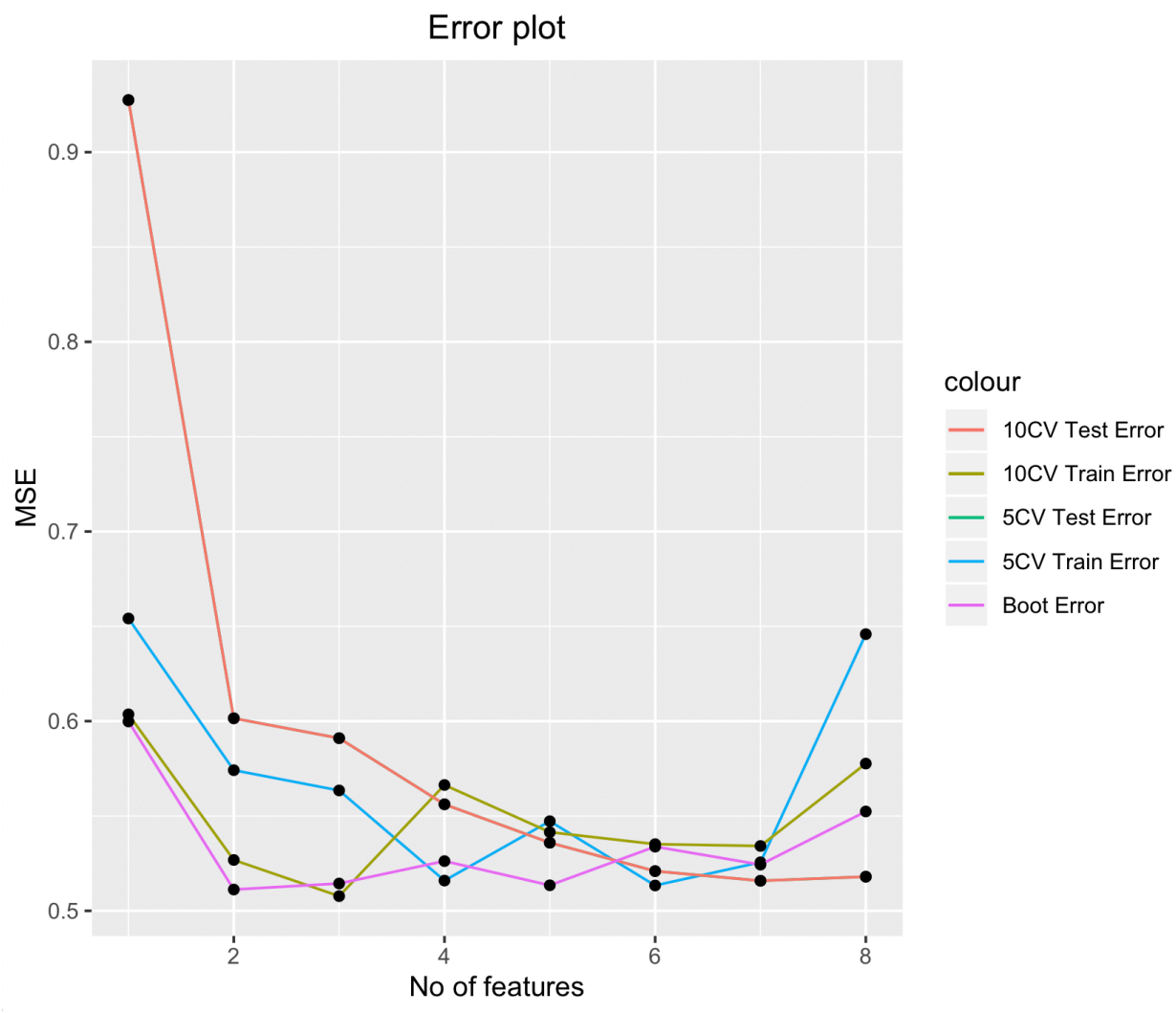
Then, performed an exhaustive subset selection method using the method, regsubsets in the leaps package. The best features are lcp and lpsa according to the Cp value.

Then, plotted graph for AIC and BIC values across different subset selection. Below is the graph:



From AIC and BIC values, they too agree with Cp value that 2 feature subset gives the best results.

Then, performed 5CV, 10CV and bootstrap .632. Here is the plot for errors:



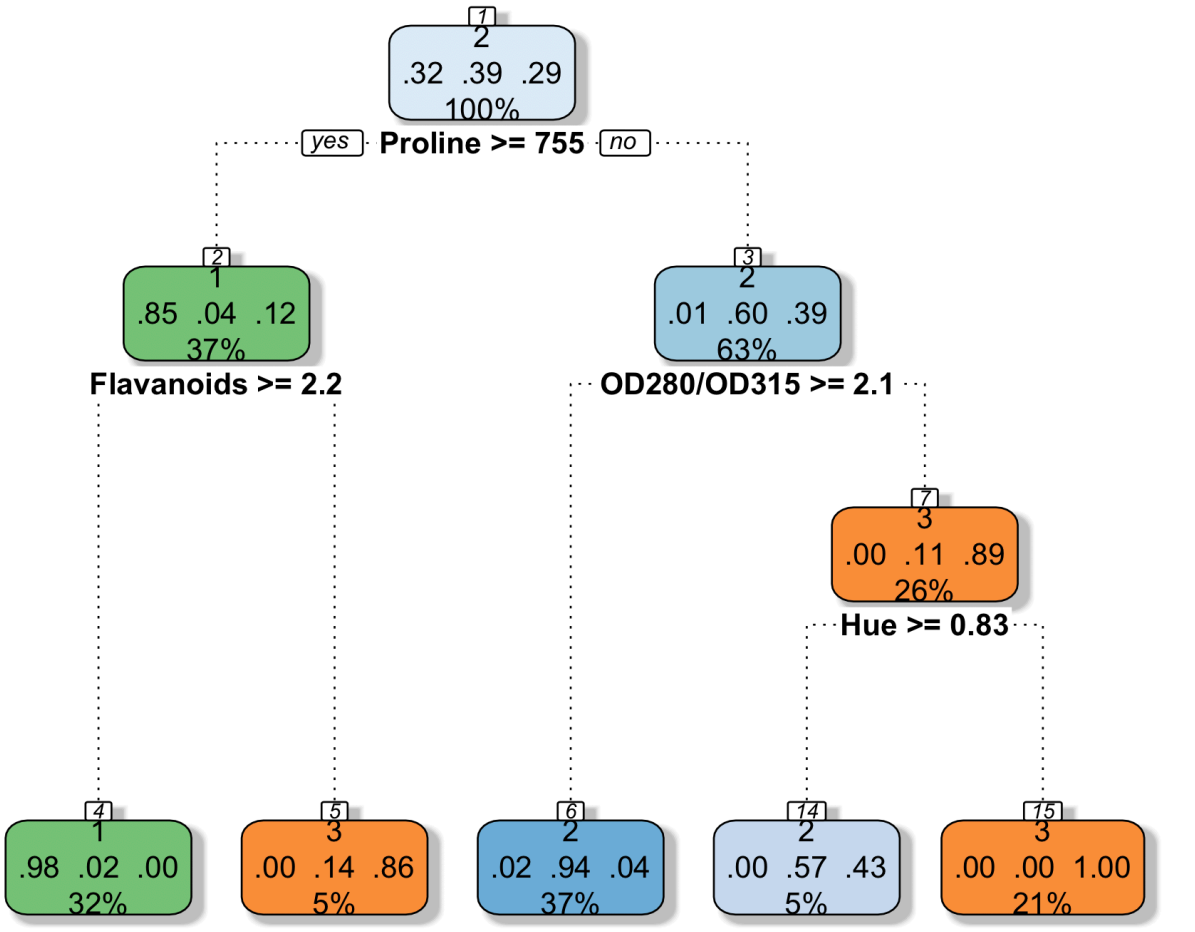
From this we can say that in almost all of them we get that using 2 feature subsets gives the best result and best for this dataset. The MSE is least for Bootstrap with less features. And, then it is less for 10 fold cross validation for more features.

Problem 2:

In problem, we are given a wine dataset. We had to construct a classification tree.

After loading data, did a sanity check that it has NA values or not. Then, The target feature is wine feature which is numeric feature in dataset which is to be categorial value. Then, converted it into categorial value.

Then, performed a decision tree on the dataset and plotting the decisions gives the tree.



From this tree we can say that data is not that much connected or it is separable since the purity at leaf nodes is high despite smaller trees.

And, this tree is made by only four features - Proline, Flavanoids, OD280/OD315 and Hue despite being 13 features used for this.

The distributions of training set in the leaf nodes are the following:

Class 1 Node 4 - 45
Class 3 Node 5 - 7
Class 2 Node 6 - 53

Class 3 Node 7 - 0
Predicting class 1 and 2 is relatively easy since the purity of leaf nodes is high. Predicting class 3 is so difficult because the purity of their leaf nodes is very less.

The distributions of testing set in the leaf nodes are the following:

Node 4 - 14
Node 5 - 1
Node 6 - 16
Node 7 - 5

Following are the accuracy rate for training and testing dataset.

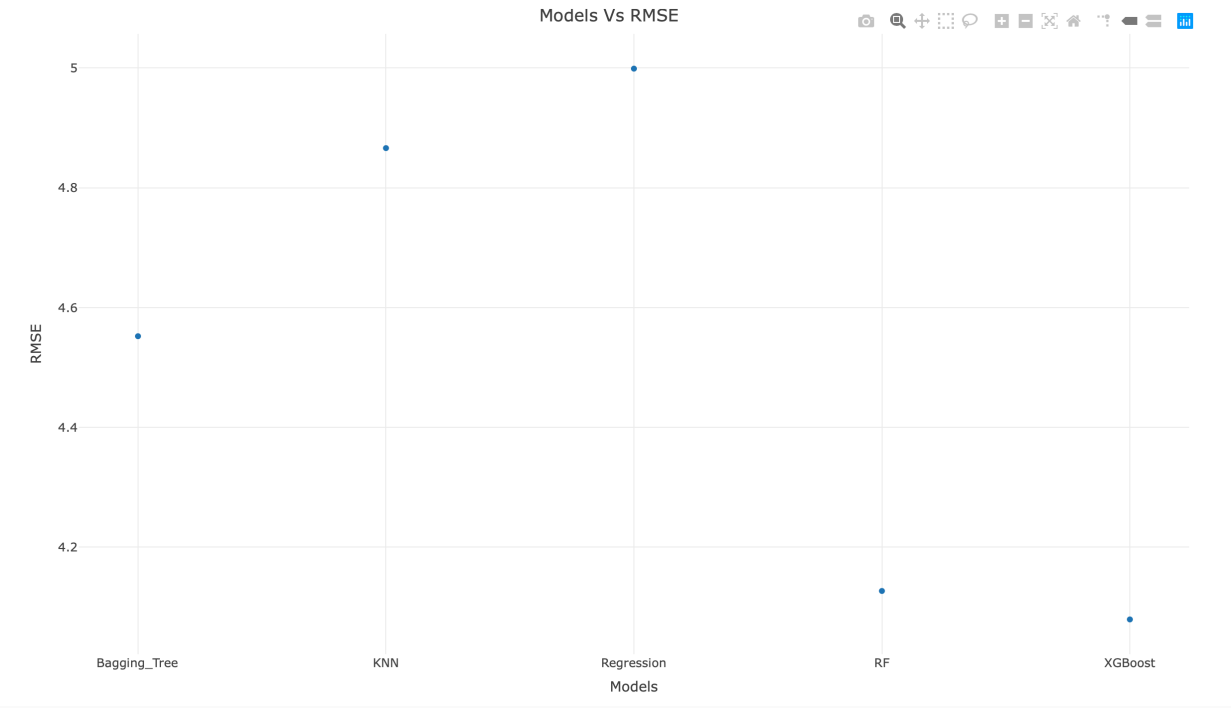
Training Accuracy : 94.36%.
Testing Accuracy: 86.11%.

Problem 3:

In this problem, we have to choose a dataset. Used a Boston dataset to apply bagging, boosting and Random forest.

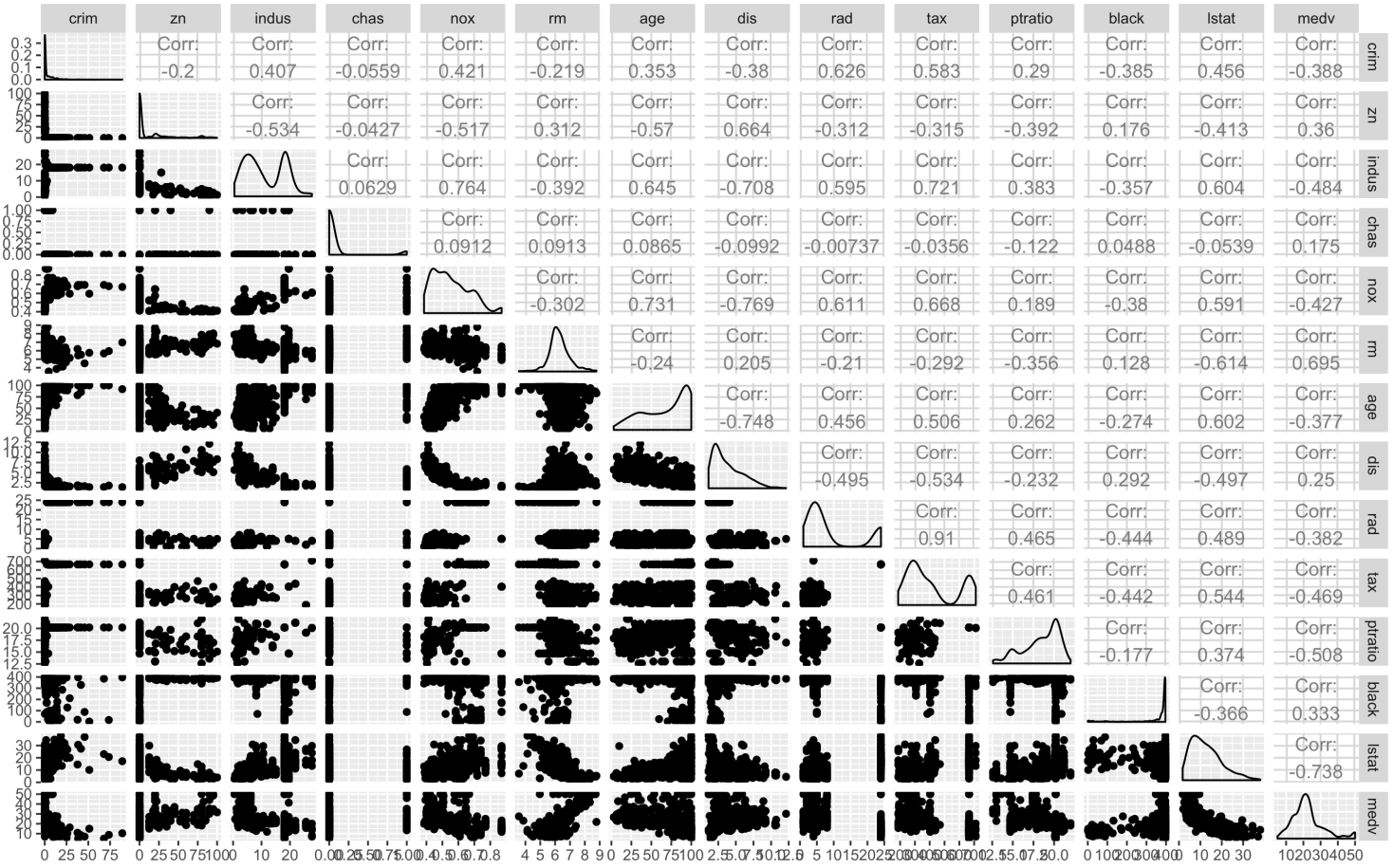
After loading data, did a sanity check and divided data into training and testing sets. Then, performed bagging, boosting, Random forest, linear regression and KNN model and checked the results.

Here is the result:



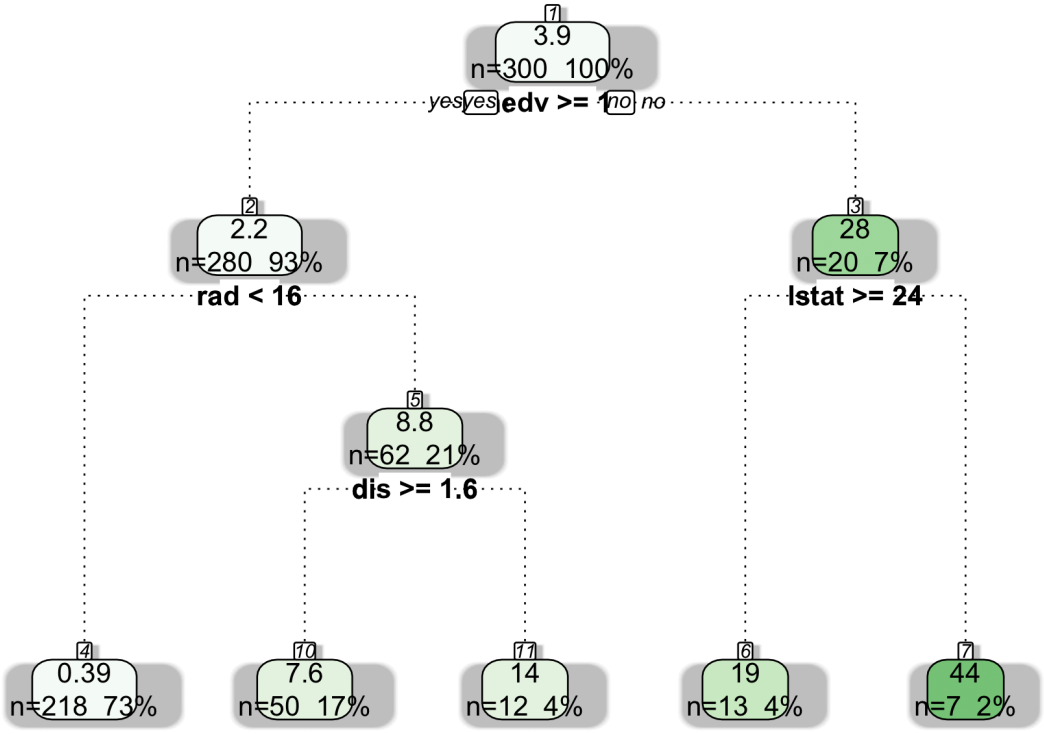
As we can see from the figure that the RMSE is lowest for Random Forest and Boosting. This is because the tree model are performing good for this dataset could be that the proper blob like separation along x axis and y axis.

Then, plotted a pair plot:



From this pairplot, we can say that there is good separation structure by which taking mean is easier which confirms that the tree model are performing better.

Then, plotted the decision tree. Here is the result:



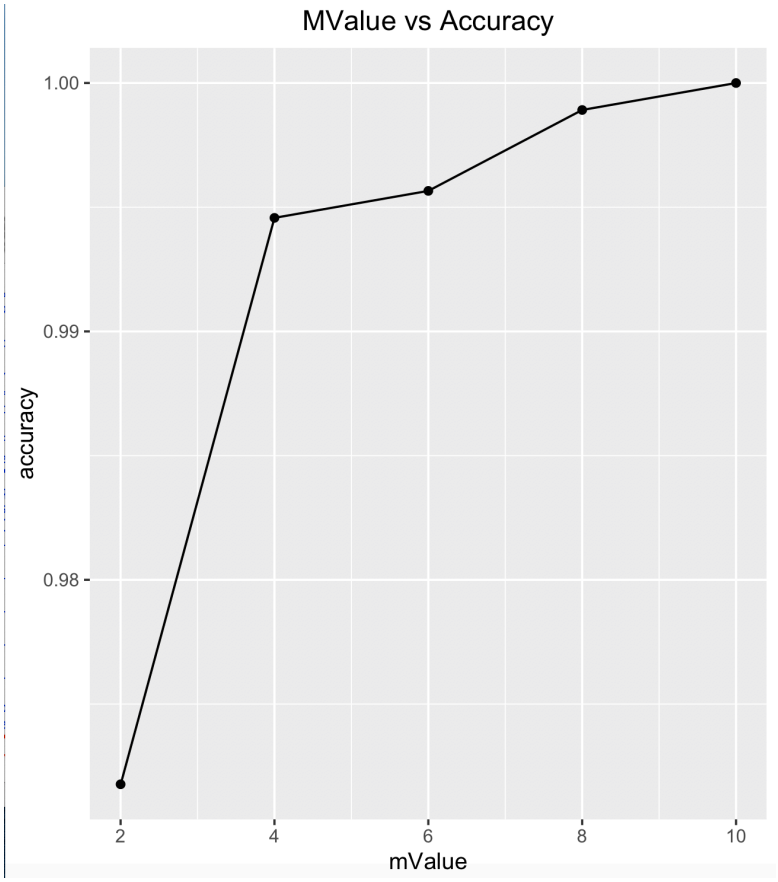
Rattle 2019-Nov-26 16:15:18 akshayadlakha

Problem 4:

In this problem, we are given a SPAM dataset. We had to perform a series of random forest classifiers to explore the sensitivity of m(randomly chosen values).

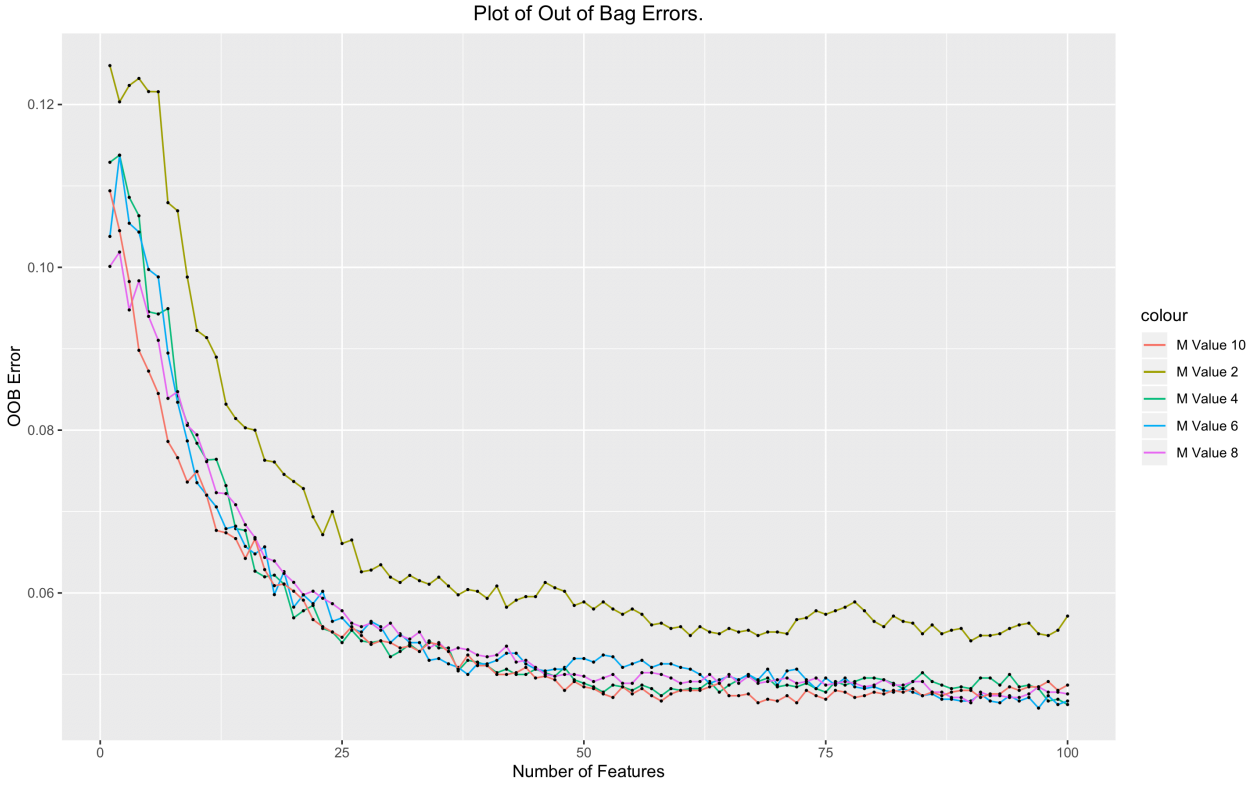
After loading data, did a sanity check whether it has NA values or not. Then, performed a series of random classifiers. Iterated through different values of m and plotting it gives the following graph.

As the tree is built



per expectation, the accuracy is increasing as number of features in a increases.

Random forest model is with 100 trees. Out of Bag errors are plotted for different values of m. Here the result:



The error is decreasing as the number of features increasing. And, we can see that the error of M Value 2 is more out of all M values. The curve for M Value 2 is above all of the curves.

Problem 5:

Random Forest Classifiers, like its name implies, it consists of a large number of individual decision trees that works as the whole. The subset of data is randomly selected. In Random Forest, Each individual tree gives a class prediction and the class with the most votes becomes our model's prediction.