EAS596, Homework_4

Abhishek Kumar, Class#1

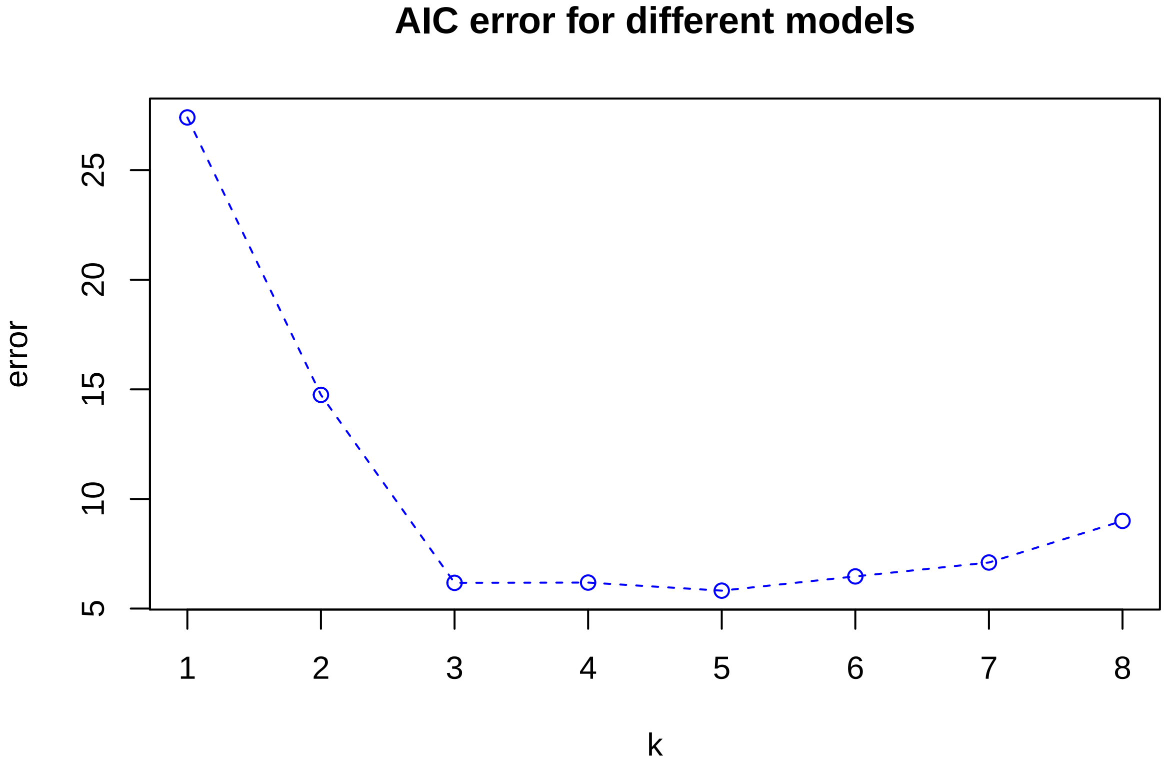
26/11/2018

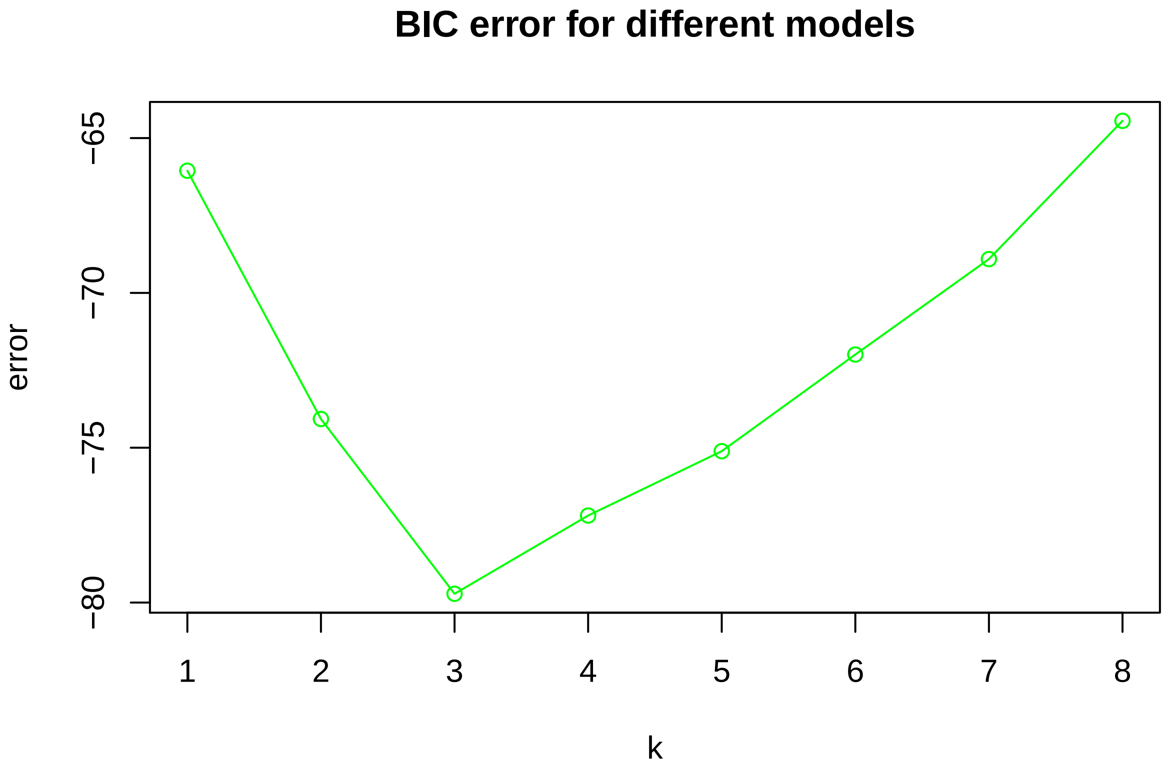
SOLUTION 1

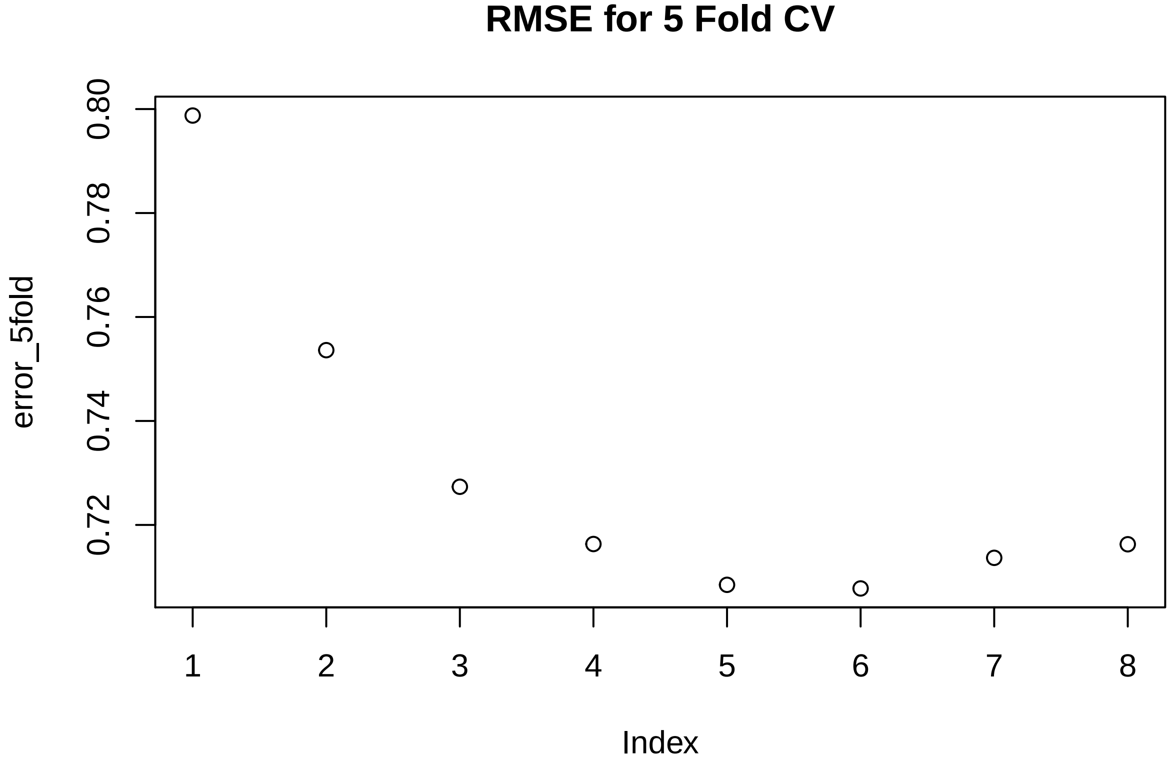
```
## 'data.frame':    97 obs. of  10 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
##  $ train  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

```
## Minimum Classification Error while using Cp is at k =  5
```

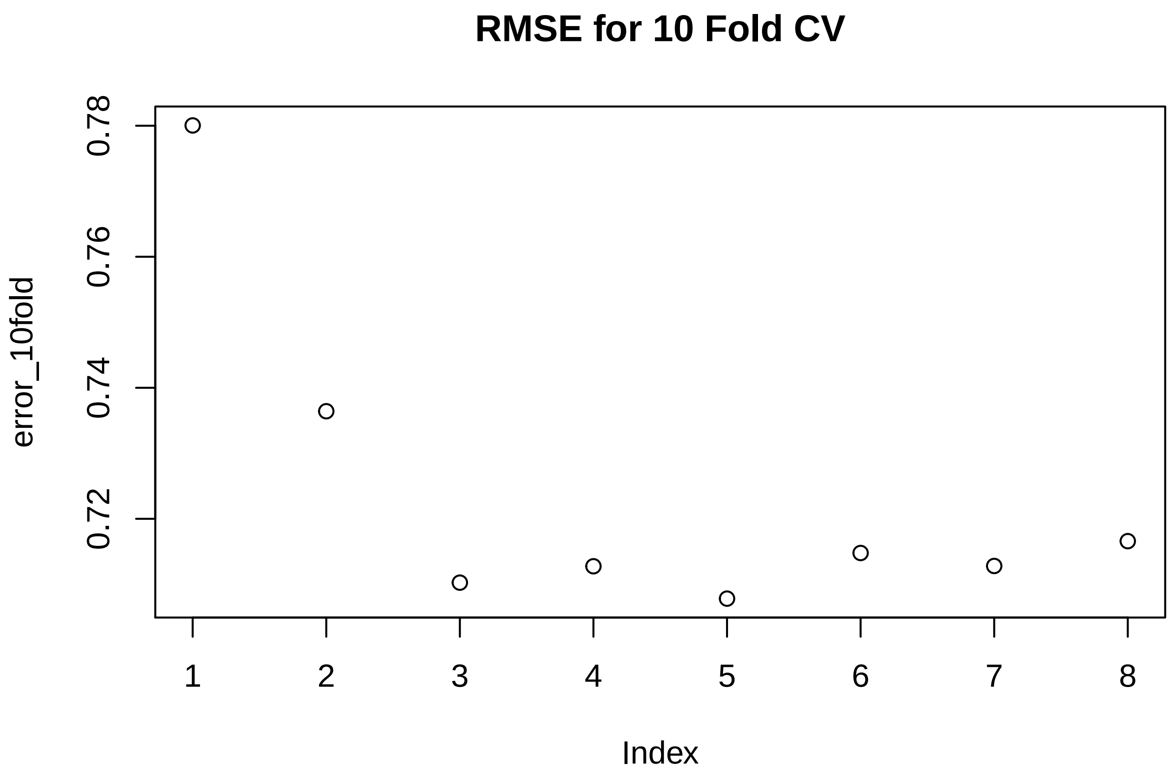
```
## Minimum Classification Error while using BIC is at k =  3
```



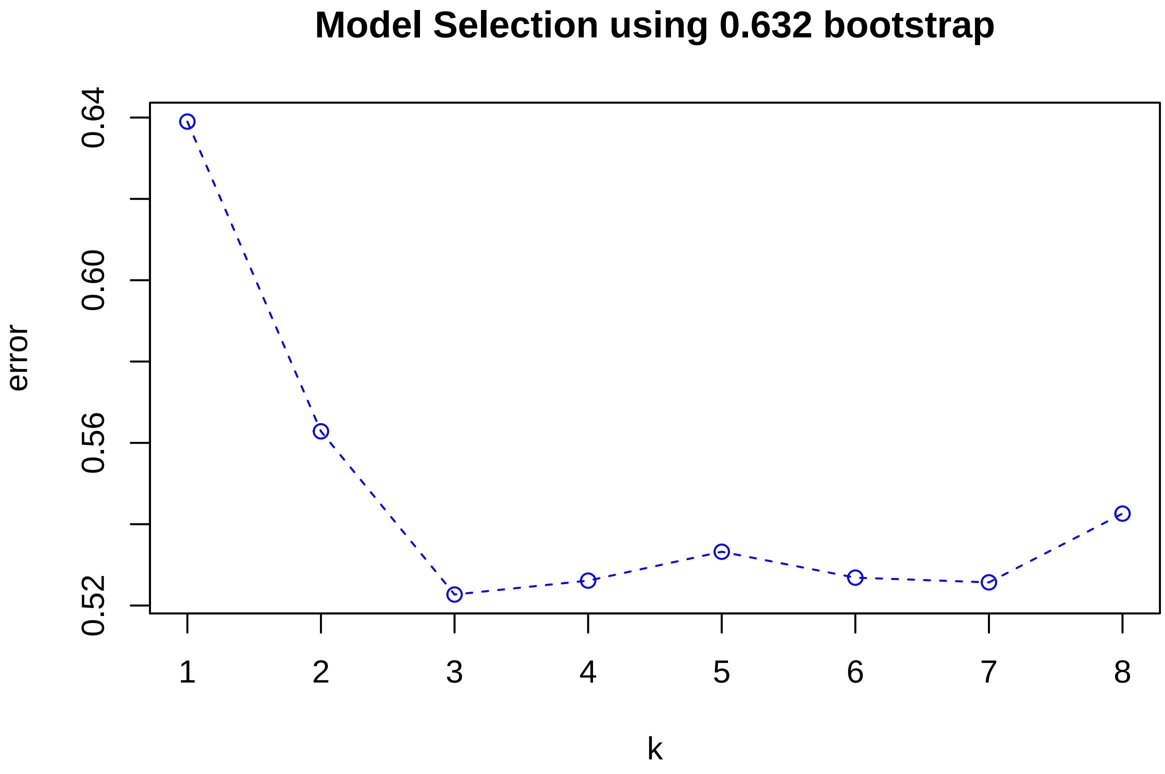




```
## The minimum error with 5 fold cross-validation is at k =  6
```



```
## The minimum error with 10 fold cross-validation is at k =  5
```

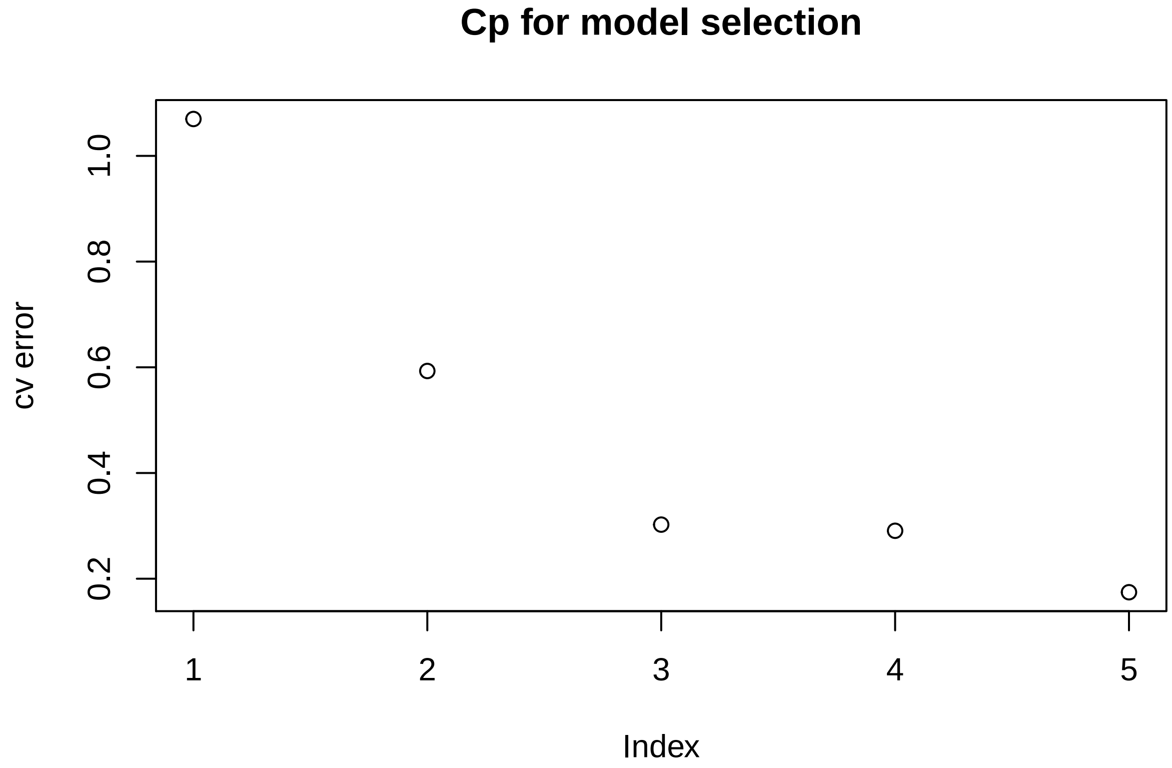


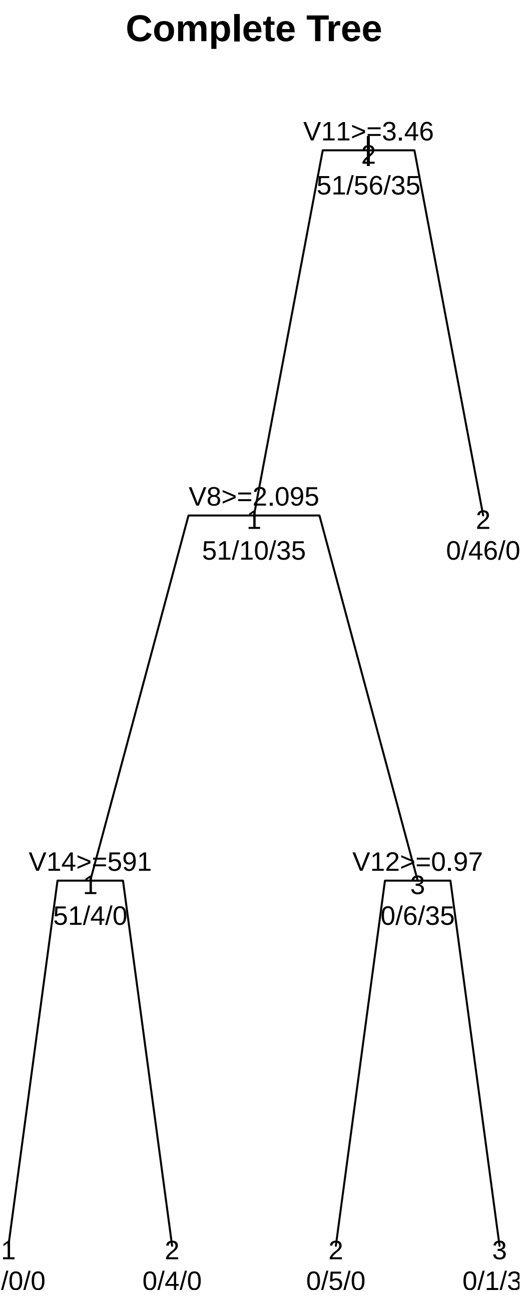
```
##          lcavol lweight age lbph svi lcp gleason pgg45
## 1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "
## 2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "
## 3  ( 1 ) "*"    "*"     " " " "  "*" " " " "     " "
## 4  ( 1 ) "*"    "*"     " " "*"  "*" " " " "     " "
## 5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "
## 6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     "*"
## 7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"
## 8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"
```

We see that almost all the methods agree that the best model is approximately at k(number of features in the model) = 3,5,6. We should select k=3 as it will not overfit the data as we are including less number of variables in the model.

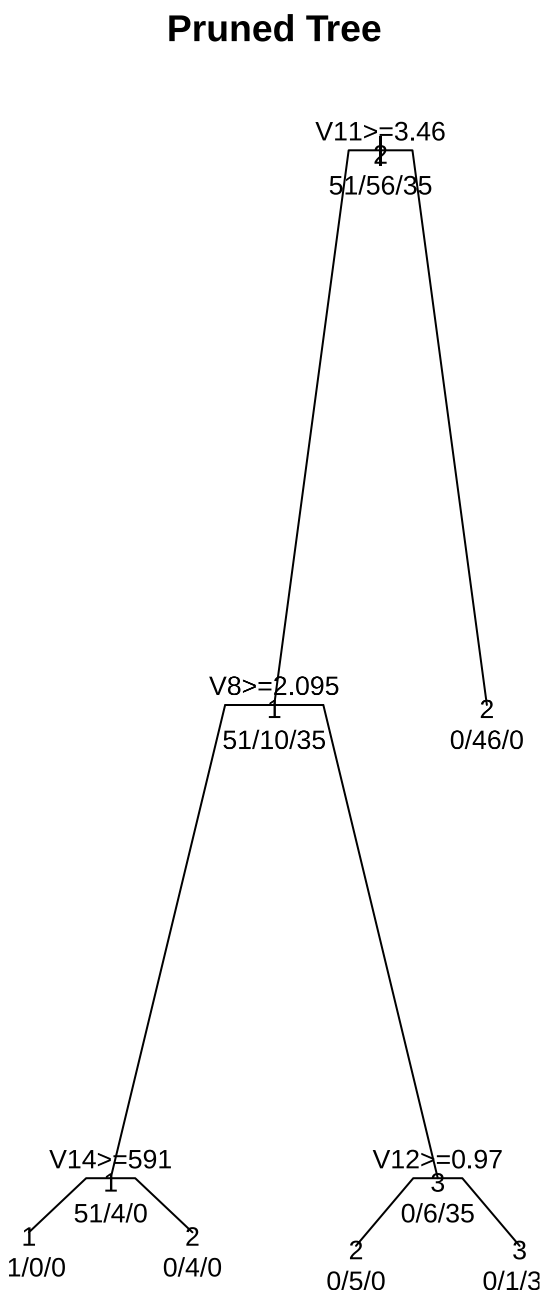
SOLUTION 2

```
## Mean Classification Test Error:  0.08333333
```





```
## Mean Classification Test Error :  0.08333333
```



```
## Mean Classification Test Error :  0.08333333
```

Calculating number of training and test data falling on each node:

```
## No. of training data falling on node 1 is  51
```

```
## No. of test data falling on node 1 is  11
```

```
## No. of training data falling on node 2 is  4
```

```
## No. of test data falling on node 1 is  0
```

```
## No. of training data falling on node 3 is  5
```

```
## No. of test data falling on node 3 is  16
## No. of training data falling on node 4 is  36
## No. of test data falling on node 4 is  0
## No. of training data falling on node 5 is  46
## No. of test data falling on node 5 is  9
```

The confusion matrix foe the test set is :

```
##    tree_pred
##      1  2  3
##   1  8  0  0
##   2  3 12  0
##   3  0  0 13
```

The tree was built using minimum split of 5, number of cross validations = 10 and the value of complexity parameter to be = 0 after iterating through all permutations of these parameters. And then we pruned the tree with complexity parameter, cp as the minimum error using cross-validation on the complete tree.

Let's look at the pruned tree and see how many data points fall into each node(leaves). We have 5 leaves node and as we can see from the tree plot that 46, 51, 4, 5 and 36 data points falls into each leaf nodes respectively.

SOLUTION 3

We will use Boston dataset and try to classify houses that lies in high crime zone or not. For this we will assume high crime zones where crime rate is more than the median crime rate and low crime zone where crime rate is below median crime rate.

```
## Logistic Regression Train Error :  0.0817942
```

```
## 
##  Logistic Regression Test error :  0.1102362
```

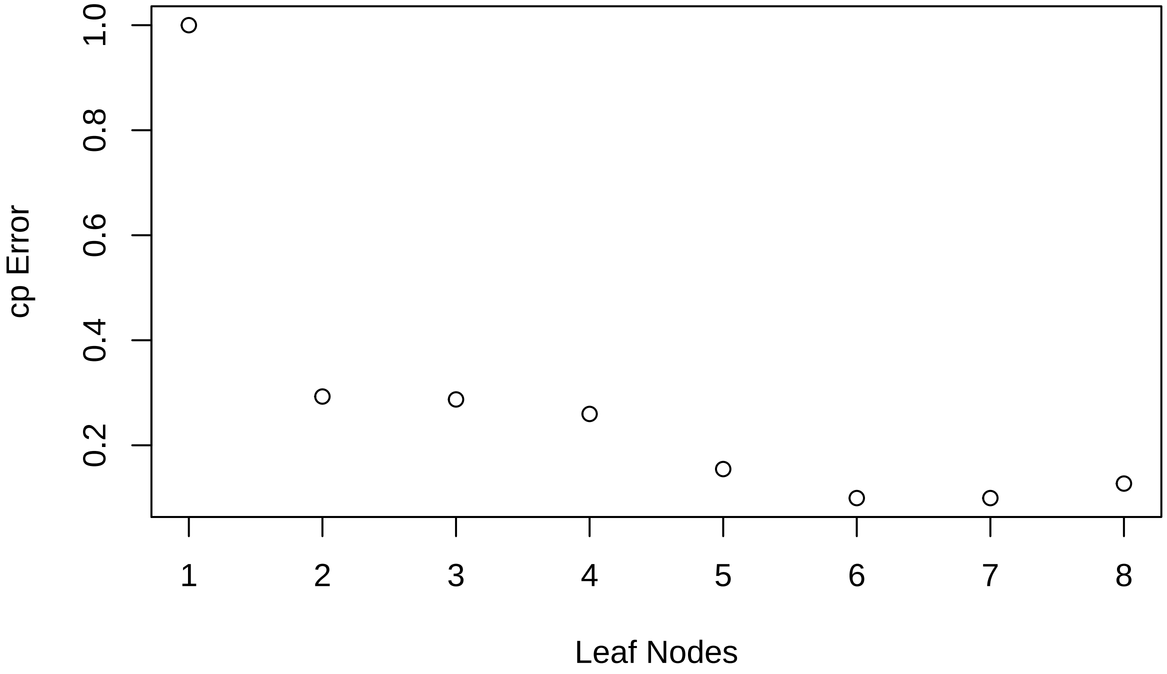
```
## Loading required package: class
```

```
## KNN Test Error for k=1:10, : 0.1181102 0.08661417 0.1102362 0.07874016 0.09448819 0.07874016 0.094488
```

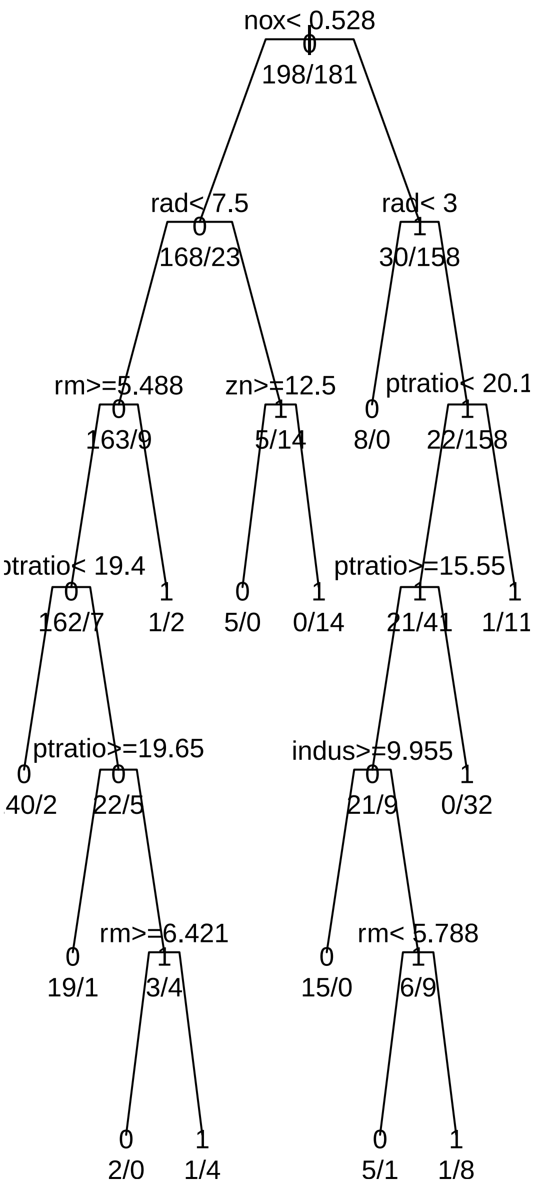
```
## For KNN, the test error is minimum at, K= 4  and the test Error is :  0.07874016
```

```
## Classifiaction error for a Single Tree : 0.03937008
```

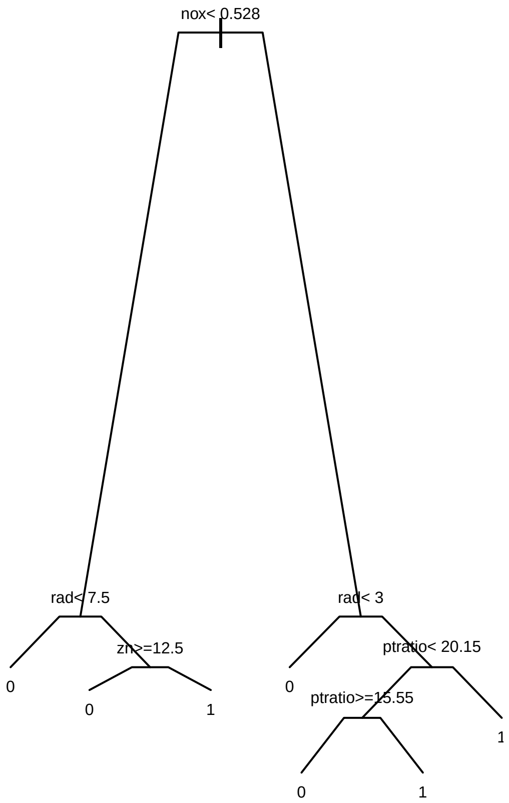
```
## Classification Error for a Single Tree with pruning:  0.05511811
```


cp Error
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
Leaf Nodes

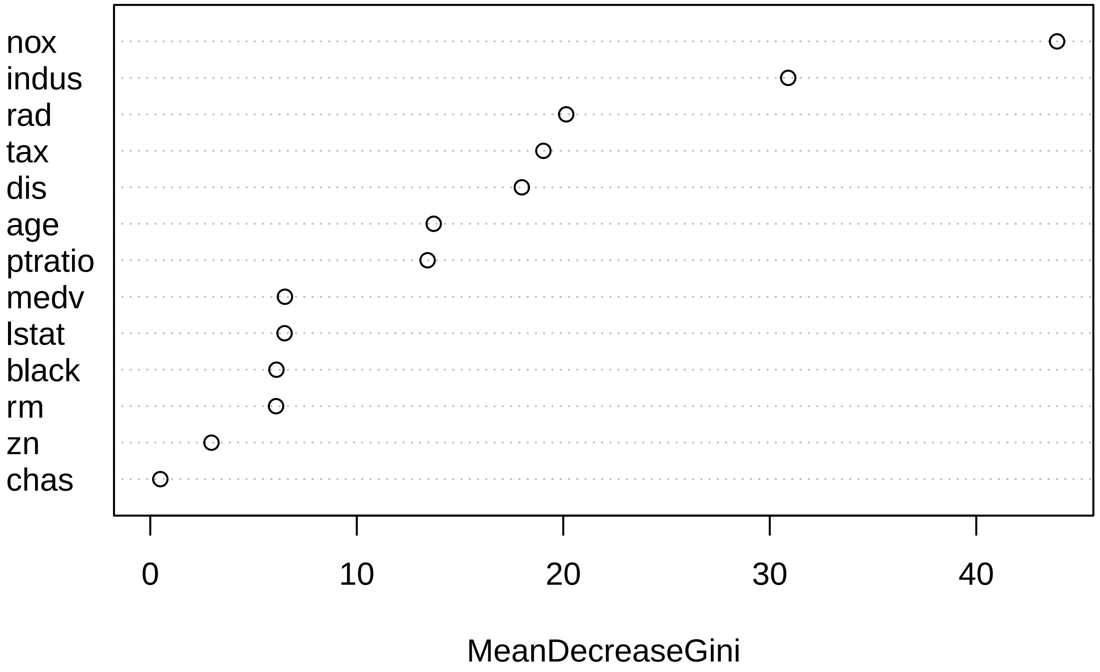
Complete Tree



Pruned Tree

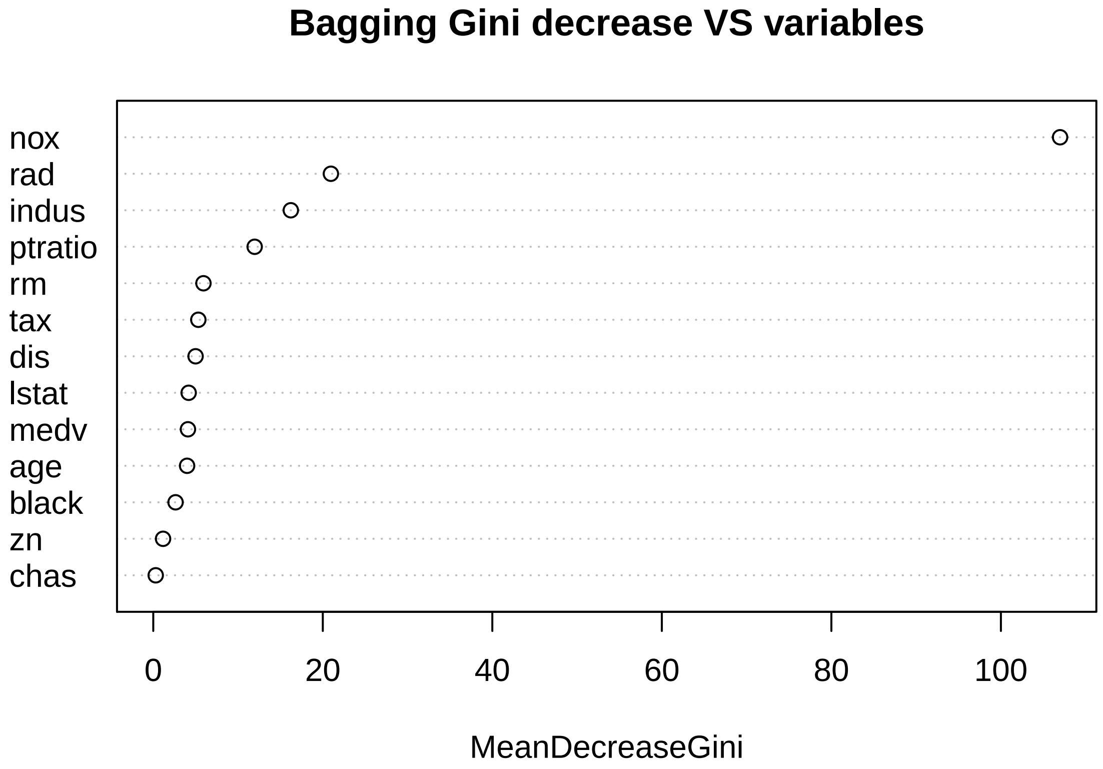


Random Forest Gini decrease VS variables



```
##         MeanDecreaseGini
## zn             2.9634876
## indus         30.8929258
## chas           0.4865255
## nox           43.9133194
## rm             6.0905434
## age           13.7250842
## dis           17.9925855
## rad           20.1425671
## tax           19.0370443
## ptratio       13.4289716
## black          6.1108055
## lstat          6.5037884
## medv           6.5206467
```

```
## Classification error for test error using Random Forest :  0.02362205
```

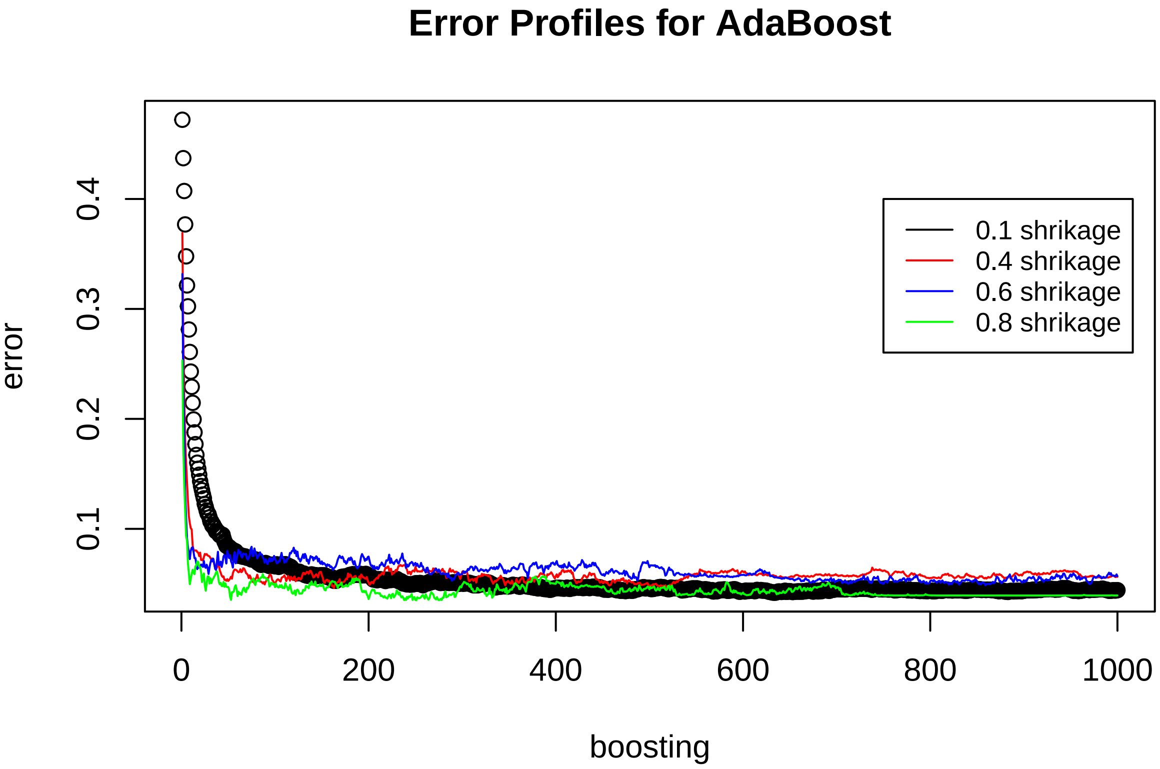


```
##         MeanDecreaseGini
## zn             1.1537401
## indus         16.2273007
## chas           0.2892555
## nox          106.9821833
## rm             5.9202510
## age            3.9902632
## dis            4.9918628
## rad           20.9560436
## tax            5.3146915
## ptratio       11.9618871
## black          2.6293943
## lstat          4.1748335
## medv           4.0898553
```

```
## Classification error for test error using Bagging : 0.01574803
```

```
## Misclassification error for AdaBoost with shrinkage = 0.1 0.05189309
```

```
## Misclassification error for AdaBoost with shrinkage = 0.6 0.04338883
```



We have implemented several methods for the Boston dataset such as Random Forest, Bagging, Boosting(AdaBoost) and other simplistic methods such as Logistic Regression and KNN. The errors on test sets were: Logistic Regression = 0.1102, KNN = 0.0787, Random Forest = 0.0236, Bagging = 0.1574, AdaBoost = 0.04338. From the Assignment 1 correlation plot for the same dataset we saw that the variables were linearly inseperable and hence we see a poor classification error by logistic regression while the committee machines methods performs quite better than logistic regression. We see that Bagging performs the best with a test error of 0.01574. Also, KNN does not perform as good as the committee machines methods because of high dimensional nature of our data.

The only disadvantage of committee machines over the simple classification methods for our data set is that it is computationally expensive.