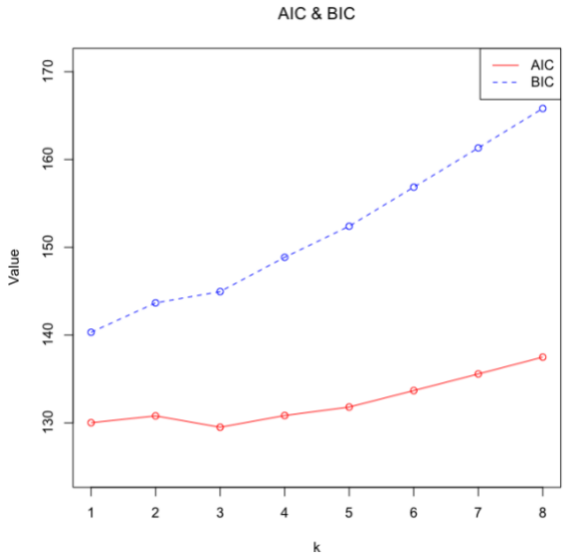
Abdoulaye Niamou
Data mining Homework4

1.

AIC, BIC and Linear Regression

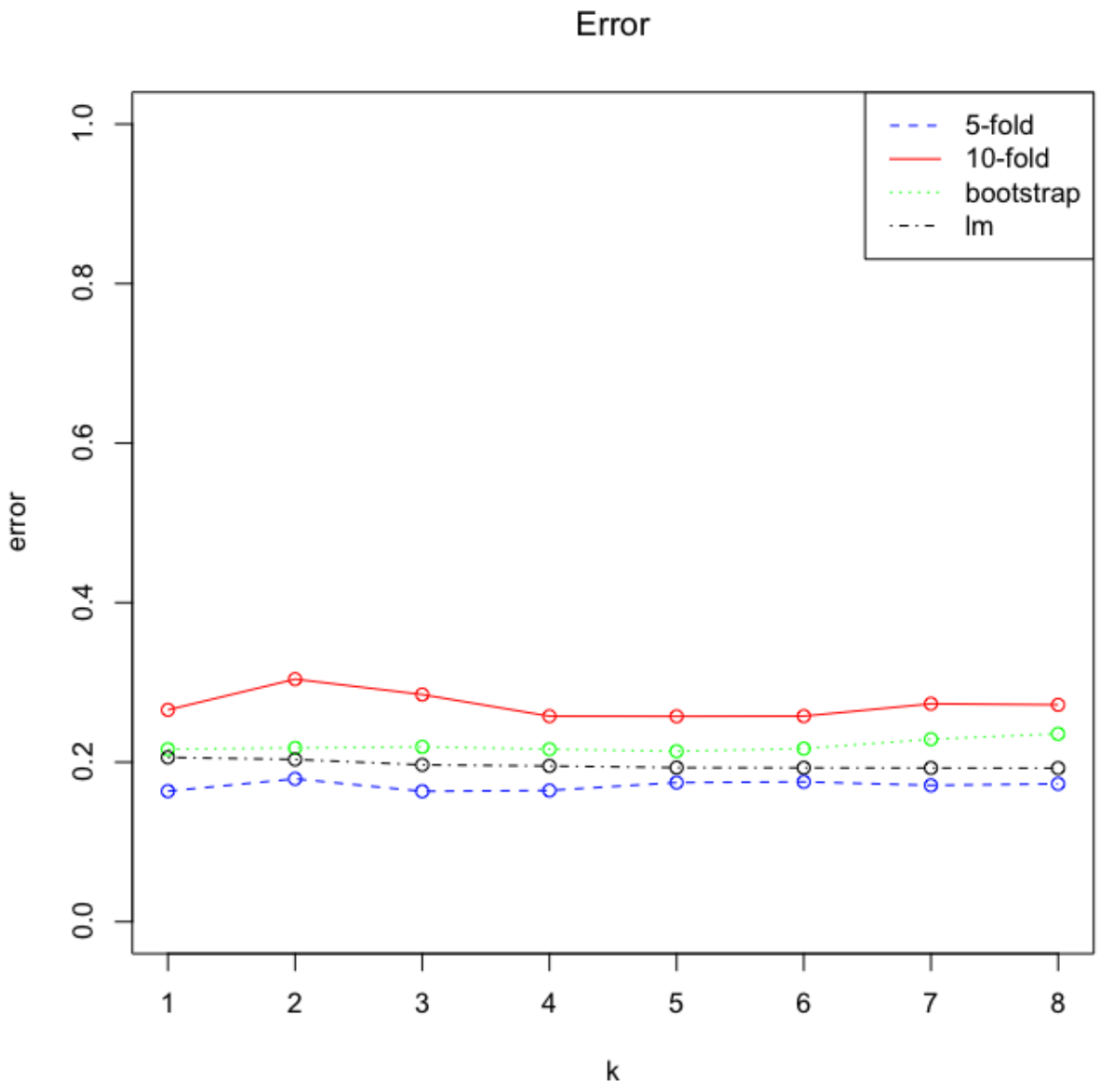
```
Selection Algorithm: exhaustive
         lcavol lweight age lbph svi lcp gleason pgg45 lpsa
1  ( 1 ) " "    " "     "*" " "  " " " " " "     " "   " "
2  ( 1 ) " "    " "     " " " "  " " " " "*"     "*"   " "
3  ( 1 ) " "    " "     "*" " "  " " " " "*"     "*"   " "
4  ( 1 ) " "    " "     "*" " "  " " "*" "*"     "*"   " "
5  ( 1 ) " "    " "     "*" "*"  " " "*" "*"     "*"   " "
6  ( 1 ) " "    "*"     "*" "*"  " " "*" "*"     "*"   " "
7  ( 1 ) " "    "*"     "*" "*"  "*" "*" "*"     "*"   " "
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"   " "
```




```
> aic_list
[1] 130.0186 130.7968 129.5066 130.8388 131.8057 133.6803 135.5732 137.4892
> bic_list
[1] 140.3174 143.6704 144.9549 148.8618 152.4034 156.8527 161.3203 165.8110
```

```
> errors_lm
[1] 0.2059847 0.2034065 0.1966227 0.1952737 0.1932049 0.1929553 0.1927423
[8] 0.1925755
Selection Algorithm: exhaustive
         lcavol lweight age lbph svi lcp gleason pgg45 lpsa
1  ( 1 ) " "    " "     "*" " "  " " " " " " " "    " "  " "
2  ( 1 ) " "    " "     " " " "  " " " " " " "*"    "*"  " "
3  ( 1 ) " "    " "     "*" " "  " " " " " " "*"    "*"  " "
4  ( 1 ) " "    " "     "*" " "  " " "*" "*" "*"    "*"  " "
5  ( 1 ) " "    " "     "*" " "  "*" "*" "*" "*"    "*"  " "
6  ( 1 ) " "    " "     "*" " "  "*" "*" "*" "*"    "*"  "*"
7  ( 1 ) " "    " "     "*" "*"  "*" "*" "*" "*"    "*"  "*"
8  ( 1 ) "*"    " "     "*" "*"  "*" "*" "*" "*"    "*"  "*"
> test_errors_5fold
[1] 0.1636624 0.1791347 0.1634921 0.1645180 0.1744424 0.1753804 0.1709710
[8] 0.1728674
Selection Algorithm: exhaustive
         lcavol lweight age lbph svi lcp gleason pgg45 lpsa
1  ( 1 ) " "    " "     "*" " "  " " " " " " " "    " "  " "
2  ( 1 ) " "    " "     " " " "  " " "*" " " " "    "*"  " "
3  ( 1 ) " "    " "     " " " "  " " "*" "*" "*"    "*"  " "
4  ( 1 ) " "    " "     "*" " "  " " "*" "*" "*"    "*"  " "
5  ( 1 ) " "    " "     "*" "*"  " " "*" "*" "*"    "*"  " "
6  ( 1 ) " "    " "     "*" "*"  " " "*" "*" "*"    "*"  "*"
7  ( 1 ) "*"    " "     "*" "*"  " " "*" "*" "*"    "*"  "*"
8  ( 1 ) "*"    " "     "*" "*"  "*" "*" "*" "*"    "*"  "*"
> test_errors_10fold
[1] 0.2655908 0.3040962 0.2847995 0.2577092 0.2575082 0.2577841 0.2732500
[8] 0.2719529
```

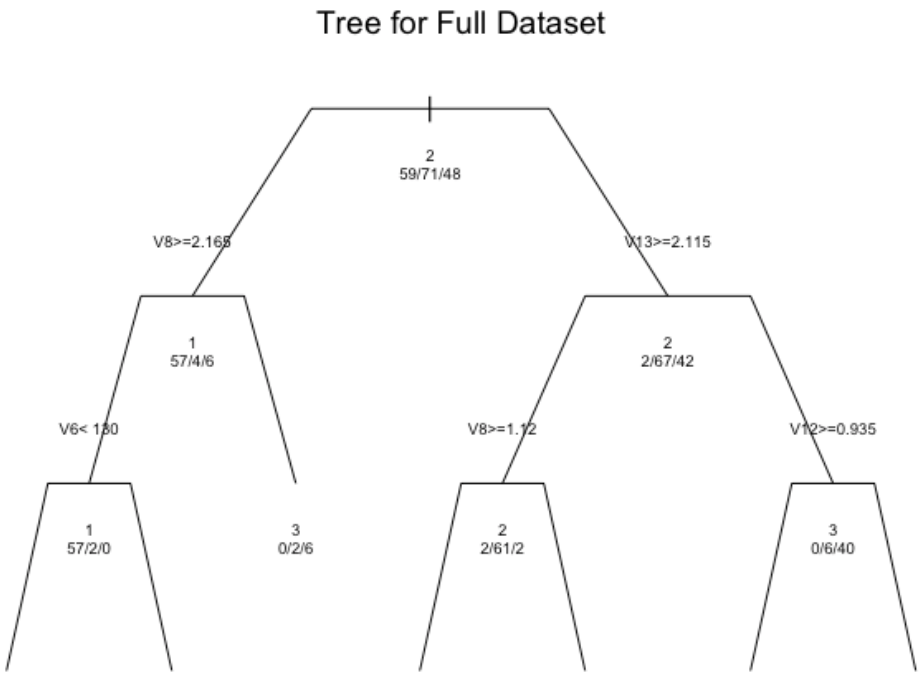
```
> errors_boot
[1] 0.2162353 0.2178410 0.2192147 0.2162492 0.2137146 0.2170170 0.2288264
[8] 0.2355435
```



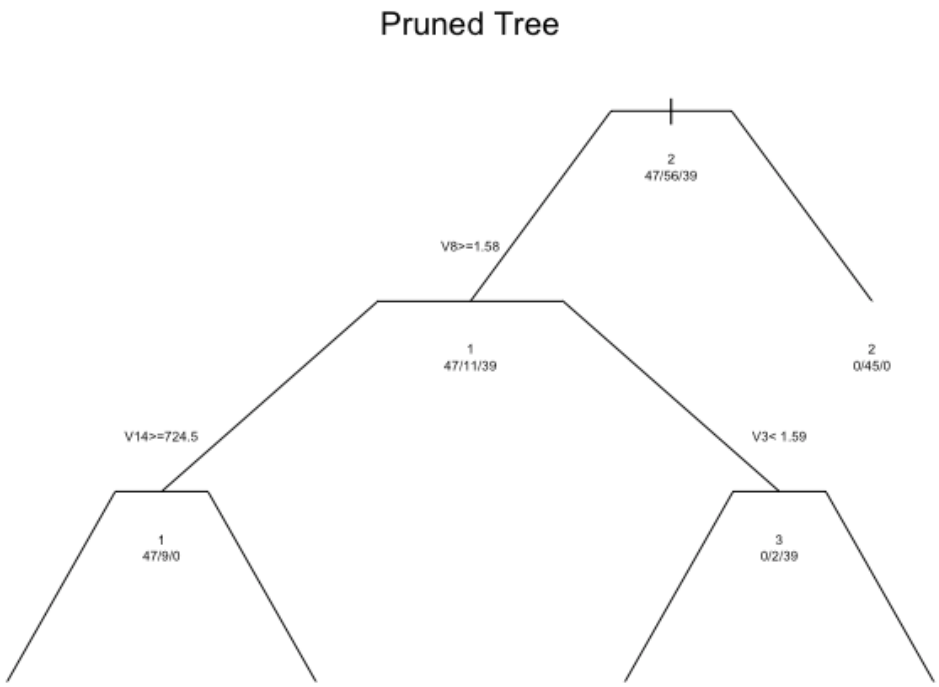
we can see that 5-fold across validations and linear model have lower errors and therefore perform better than the other two methods. For the 5-fold across validations, k = 3 performs best.

2.

We have 57 cases of Class1 wine, 2 cases of Class2 wine and 0 case of Class3 wine were classified under the first node. 2 cases of Class2 wine and 6 cases of Class3 wine were classified under the second node. 2 cases of Class1 wine, 61 cases of Class2 wine



and 2 case2 of Class3 wine were classified under the third node. 0 case of Class1 wine, 6 cases of Class2 wine and 40 cases of Class3 wine were classified under the forth node.



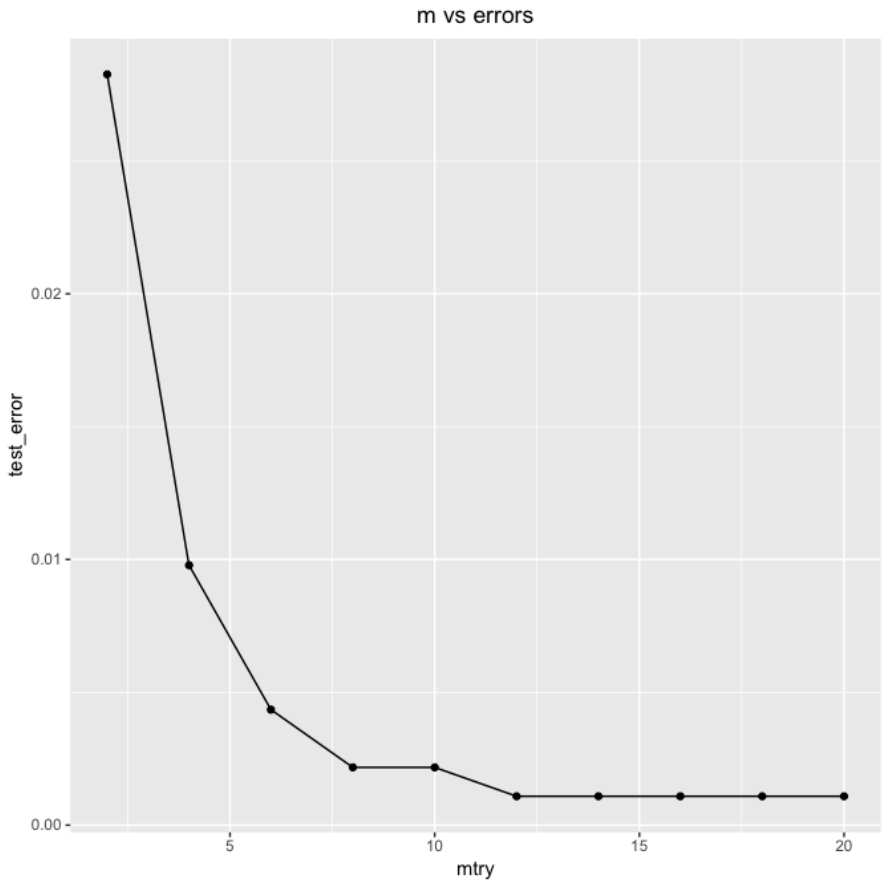
From the pruned tree we see that 47 case4 of Class1 wine, 9 cases of Class2 wine and 0 case of Class3 wine were classified under the first node; 0 case of Class1 wine, 2 cases of Class2 wine and 39 cases of Class3 wine were classified under the second node; 0 case of Class1 wine, 45 cases of Class2 wine and 0 case of Class3 wine were classified under the third node.

From those results we can see that the pruned tree performs well.

```
  tree_pred
    1  2  3
1  12  0  0
2   0 14  1
3   0  2  7
```

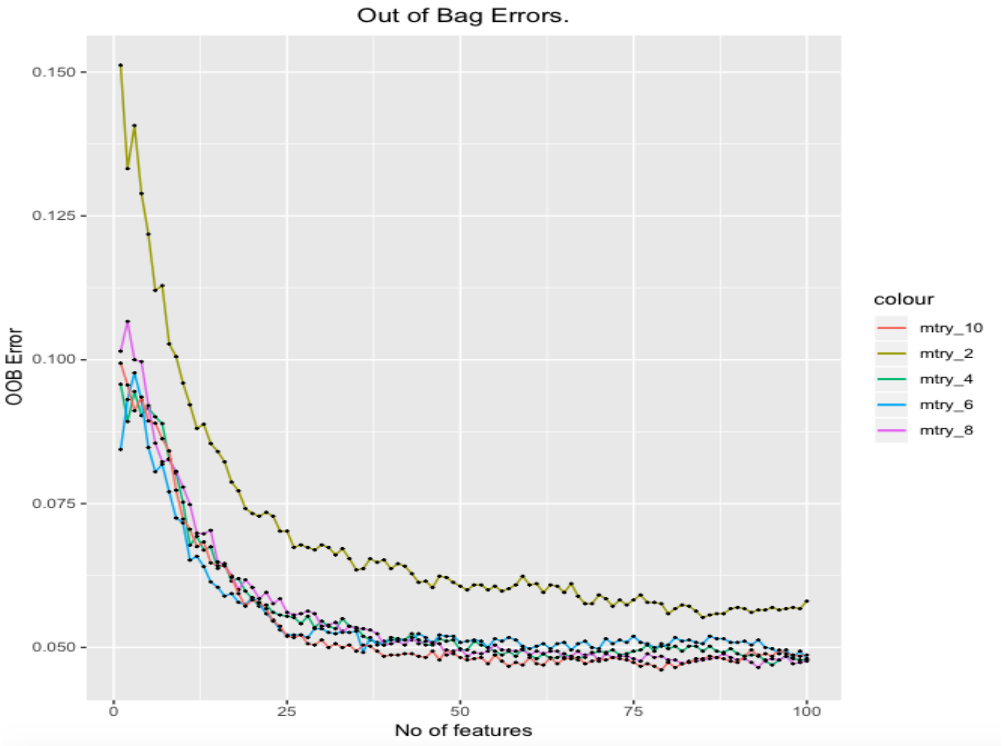
```
> mean(tree_pred != wine_test_data$V1)
[1] 0.08333333
```

4.




```
> cbind(m, TestErrors)
       m     TestErrors
 [1,]  2 0.028260869565
 [2,]  4 0.009782608696
 [3,]  6 0.004347826087
 [4,]  8 0.002173913043
 [5,] 10 0.002173913043
 [6,] 12 0.001086956522
 [7,] 14 0.001086956522
 [8,] 16 0.001086956522
 [9,] 18 0.001086956522
[10,] 20 0.001086956522
```

The following figure shows the relationship between m vs OBB errors. The OBB errors have the same trending with the test errors. The OBB errors will decrease with the increase of the number of features (m). And the error of mtry_2 (where m = 2) is consistently more out of all the curves which approve this pattern again.



5.

A random forest randomly selects observations and specific variables to build multiple decision trees, Then averages the results. After the of trees are built, each tree chooses the class, and the class chosen the most by simple majority is the or predicted class. It is random because we do not know the predicted class.