HOMEWORK 4:

Q.1)

Analysis of Variance Table

```
Response: lpsa
          Df Sum Sq Mean Sq F value  Pr(>F)    
lcavol     1   37.3    37.3   70.30 2.9e-10 ***
lweight    1    4.6     4.6    8.69  0.0054 ** 
age        1    0.0     0.0    0.02  0.8780    
lbph       1    0.6     0.6    1.14  0.2914    
svi        1    3.4     3.4    6.31  0.0162 *  
lcp        1    0.1     0.1    0.09  0.7600    
gleason    1    1.6     1.6    2.94  0.0945 .  
pgg45      1    0.5     0.5    0.87  0.3578    
train      1    0.2     0.2    0.43  0.5181    
Residuals 39   20.7     0.5                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We fitted linear regression model for prostate data to calculate lpsa. After that we get AIC value for the model which is -34.1. For 5- fold cross validation, test error we get is 0.57 and for 10-fold cross validation we get error of 0.56 which is slightly less than 5-fold CV.

Table above shows significance of each column, calculated by cross validation.

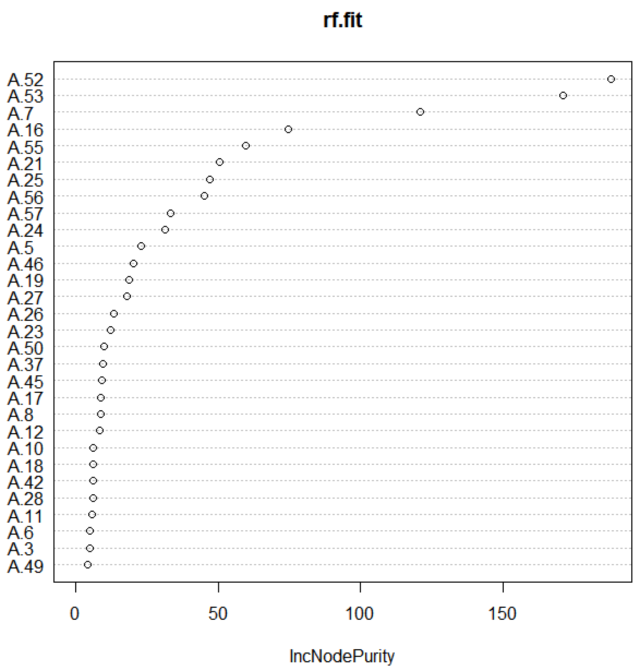
To calculate minimum bic and cp, we are using best subset selection. Result table below shows best subsets for n variables.

According to cp, best subset we get is for 4 varibles i.e. cp is least for 4 variables. Whereas, bic is least for 6 variables.

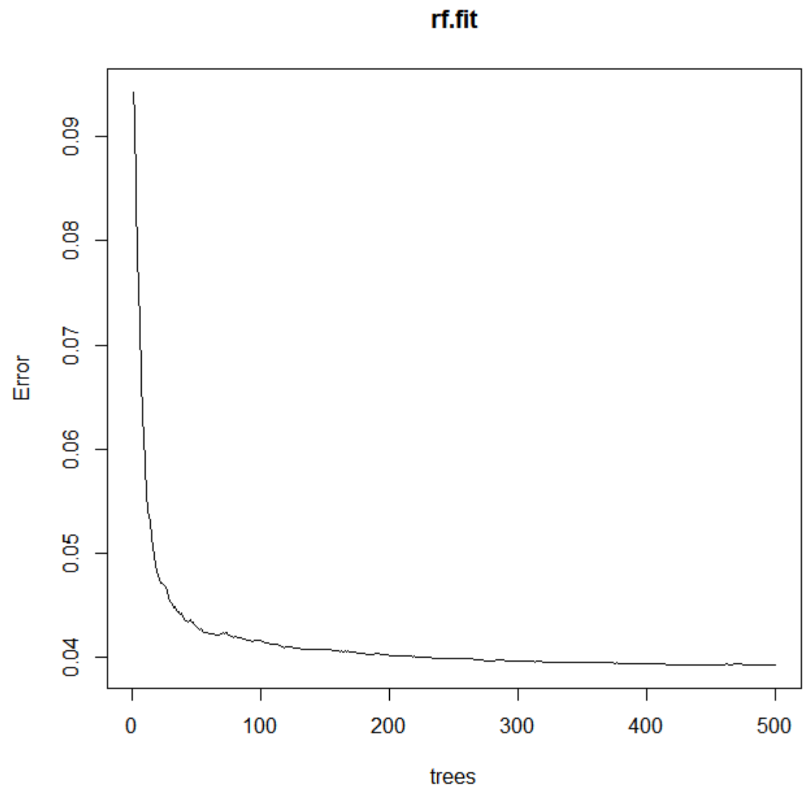
Least value of 4.46. And least value of bic is -56.5 for 6 variables.

```
         lcavol lweight age lbph svi lcp gleason pgg45 train
1  ( 1 ) "*"    " "     " " " "  " " " " " "     " "   " "  
2  ( 1 ) "*"    "*"     " " " "  " " " " " "     " "   " "  
3  ( 1 ) "*"    "*"     "*" " "  " " " " " "     " "   " "  
4  ( 1 ) "*"    "*"     "*" " "  "*" " " " "     " "   " "  
5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " "     " "   " "  
6  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     " "   " "  
7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " "     "*"   " "  
8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"   " "  
9  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*"     "*"   "*"  
```

Q.4)



From the figure we can see the graph of variable importance. A.52 is the most important variable. So without A.52 correctness of random forest model is the least as compared to other inputs. So the graph above shows sensitivity for all the inputs

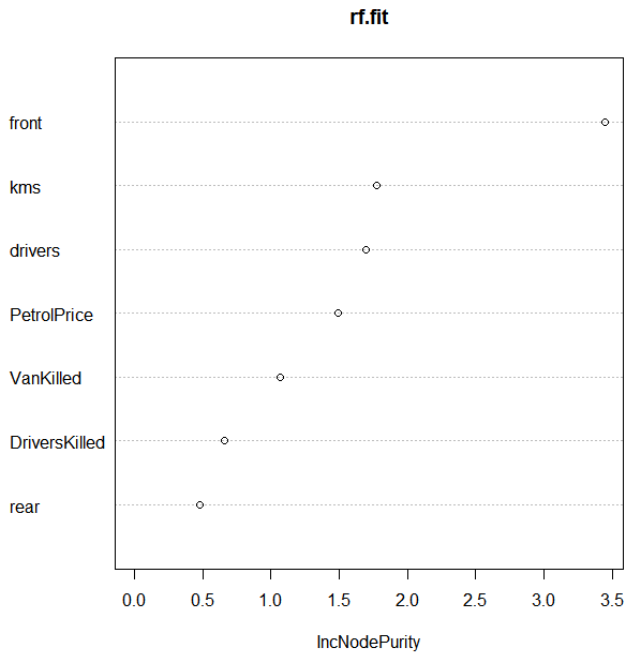


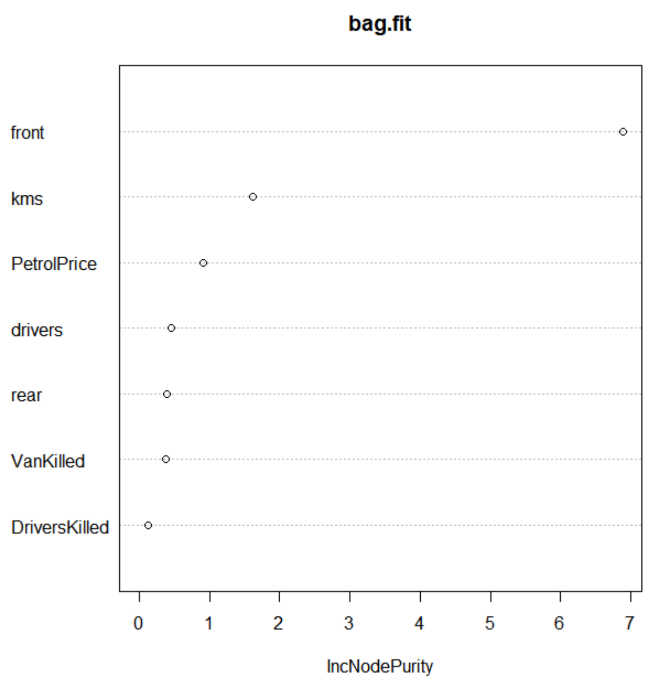
The figure above shows OOB error for various number of trees used in random forest. We can see that OOB error exponentially decreases and is constant from about 400 trees.

We get very small test error i.e. 0.005215123859 for this model.

Q.3)Surprisingly we are getting least error for logistic regression. For ensemble methods also the error is very less. We get approximately 3% error in boosting and 5% error for bagging. For random forest error i s quite large i.e. 6.5% and maximum among all the classification methods.

The committee machine data gives quite less accuracy which is in the fraction of percentage as compare d to other data for which accuracy is just around 95-96% as observed above.





Variable importance for various methods is shown in the figure above.

Q.7)

For the linear kernel svm, optimal cost is 0.01. We get training error of just 0.131 for this cost.

```
      cost error dispersion
1   0.0100 0.131     0.0264
2   0.0178 0.132     0.0263
3   0.0316 0.132     0.0261
4   0.0562 0.132     0.0263
5   0.1000 0.132     0.0261
6   0.1778 0.132     0.0260
7   0.3162 0.132     0.0260
8   0.5623 0.132     0.0259
9   1.0000 0.132     0.0260
10  1.7783 0.132     0.0261
11  3.1623 0.132     0.0260
12  5.6234 0.132     0.0260
13 10.0000 0.132     0.0259
```

For radial kernel svm, we get optimal cost at 0.5623 and the error we get for that is 0.131. Table below show error for various values of costs.

```
- Detailed performance results:
      cost error dispersion
1   0.0100 0.308     0.0554
2   0.0178 0.278     0.0525
3   0.0316 0.233     0.0470
4   0.0562 0.177     0.0366
5   0.1000 0.141     0.0245
6   0.1778 0.133     0.0231
7   0.3162 0.131     0.0227
8   0.5623 0.131     0.0226
9   1.0000 0.135     0.0233
10  1.7783 0.138     0.0243
11  3.1623 0.138     0.0241
12  5.6234 0.139     0.0244
13 10.0000 0.144     0.0271
```

For polynomial kernel with degree 2, least error is 0.154 for the optimal cost of 1.78.

```
- Detailed performance results:
      cost error dispersion
1   0.0100 0.315     0.0567
2   0.0178 0.295     0.0577
3   0.0316 0.273     0.0571
4   0.0562 0.254     0.0539
5   0.1000 0.233     0.0530
6   0.1778 0.213     0.0489
7   0.3162 0.186     0.0413
8   0.5623 0.166     0.0322
9   1.0000 0.156     0.0304
10  1.7783 0.154     0.0310
11  3.1623 0.154     0.0295
12  5.6234 0.158     0.0317
13 10.0000 0.160     0.0358
```