STATISTICAL DATA MINING – I
HOMEWORK IV

NAME: SIDDIQ SYED

UB Person # 50291566

Class # 50

1) (20 points) (Exercise 7.9) For the prostate data of Chapter 3, carry out a bestsubset linear regression analysis, as in Table 3.3 (third column from the left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap .632 estimates of prediction error.

Exploratory analysis of the data

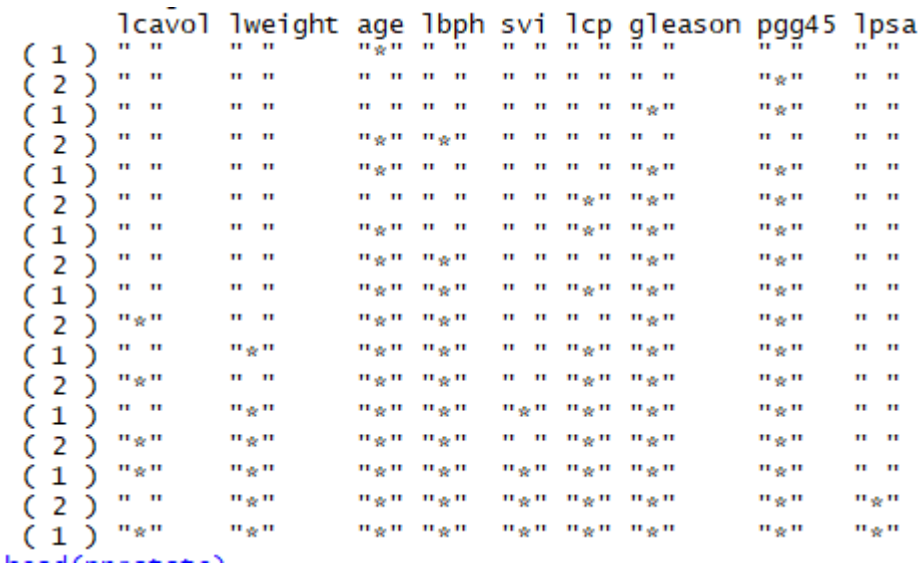
Loading the data in to R studio for which the column train has TRUE and FALSE values which says that the model that can be used for prediction should be of classification.

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa	train
1	-0.580	2.77	50	-1.39	0	-1.39	6	0	-0.431	TRUE
2	-0.994	3.32	58	-1.39	0	-1.39	6	0	-0.163	TRUE
3	-0.511	2.69	74	-1.39	0	-1.39	7	20	-0.163	TRUE
4	-1.204	3.28	58	-1.39	0	-1.39	6	0	-0.163	TRUE
5	0.751	3.43	62	-1.39	0	-1.39	6	0	0.372	TRUE
6	-1.050	3.23	50	-1.39	0	-1.39	6	0	0.765	TRUE

Figure 1: Head data for prostate dataset

The train column needs to be converted into numeric binary form for easy application of model building.

Building exhaustive subset model to check what are the most important parameters that are required to train the best model for the given dataset.



From the above image it is to be seen that the column age and pgg45 are the best single parameter models for the dataset. And like wise for 2,3.. can be observed in the above image.

Observing the training and test errors for the dataset with different combinations for the above obtained important parameters.

It is observed below that for age,gleason and pgg45 parameters used for model building with linear regression below are the errors obtained.

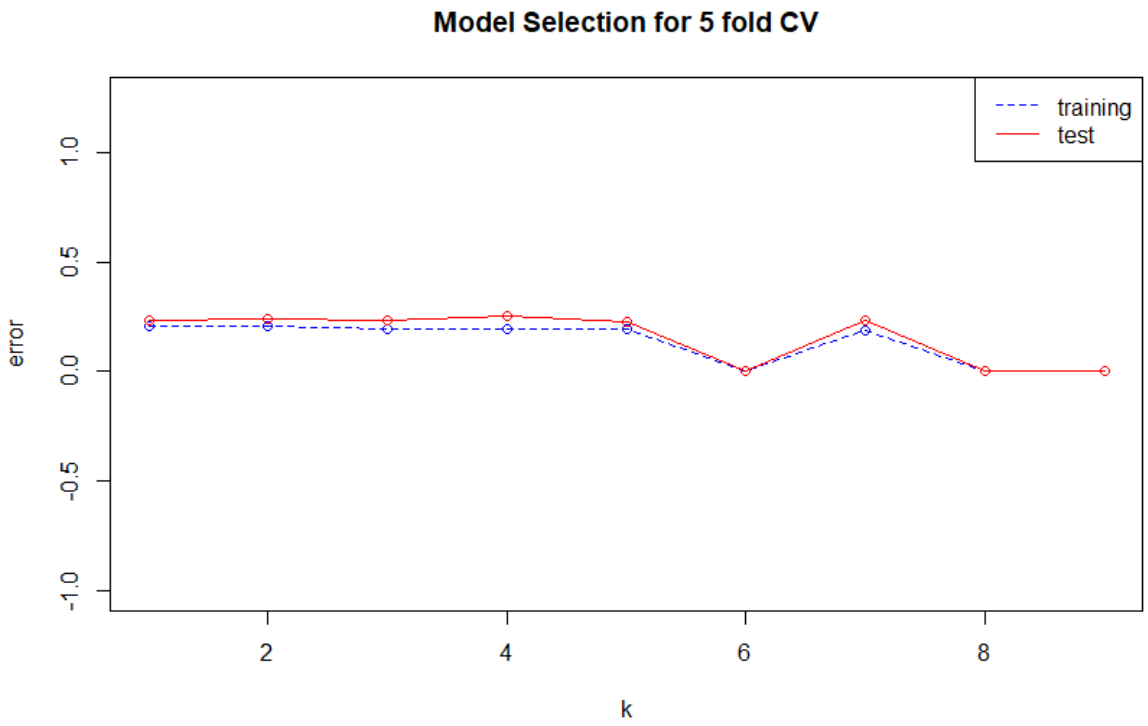
```
> test.error_lm
[1] 0.164
> train.error_lm
[1] 0.21
```

And for 4 best subset model below are the errors obtained.

```
> test.error_lm
[1] 0.159
> train.error_lm
[1] 0.209
```

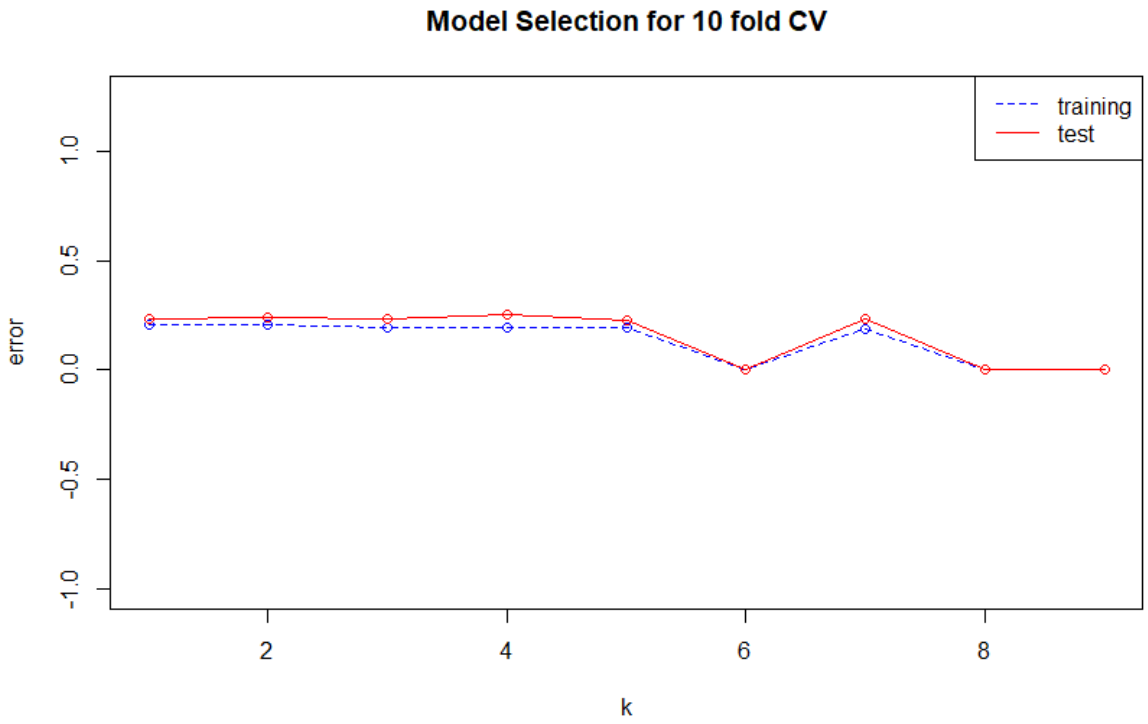
Which is evident that the error rate has been decreased.

Applying 5 fold cross-validation for the model building of each best subset linear regression model we can see the below graph between error and number of parameters.



From the above plot it is evident that the error is very least for 6 8 and 9 parameters.

Let's see for the 10 fold CV of the same.



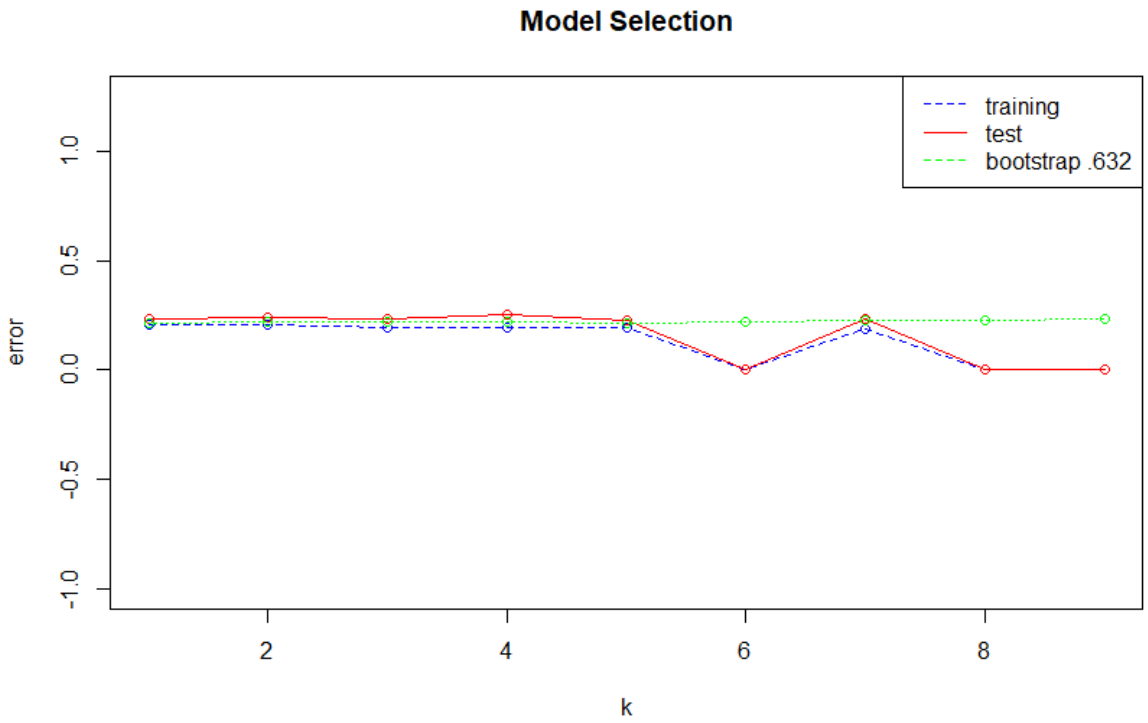
From the two graphs it is evident that the error rate is same for both 5 and 10 folds.

Looking at the CP and BIC values for the exhaustive subset model for increasing number of parameters.

```
> my_summary$cp
[1] -0.150 -0.238  0.567  1.629  2.835  4.431  6.157  8.073 10.000
> my_summary$bic
[1]  6.44  8.54 11.58 14.87 18.32 22.21 26.24 30.48 34.74
>
> which.min(my_summary$cp) #Cp says 2 variables is best
[1] 2
> which.min(my_summary$bic)
[1] 1
```

Above values clearly says that the 2 and 1 parameter models are the best. As per the cross validation errors 6, 8 and 9 were the lowest error rate models.

From bootstrap .632 prediction error which is plotted below.



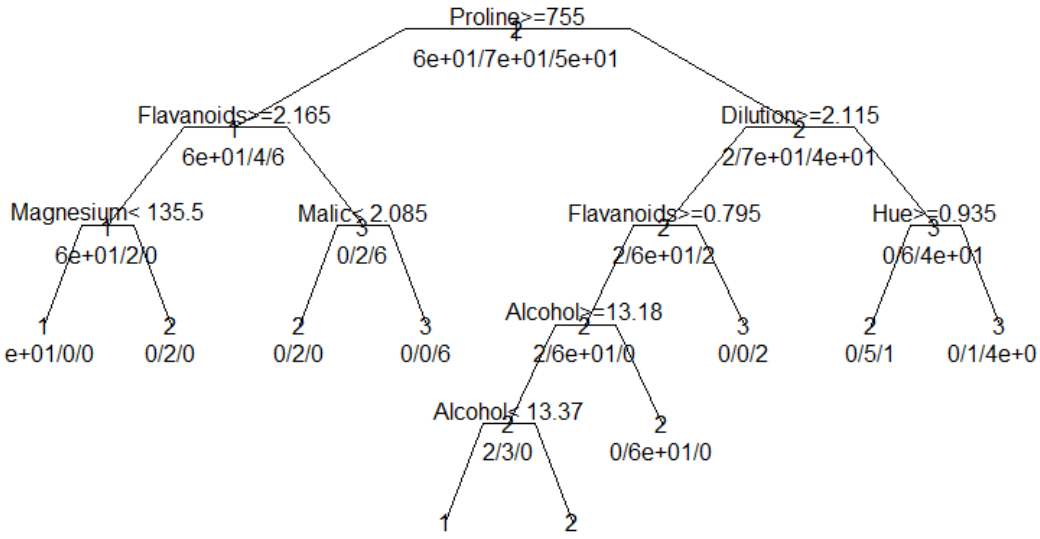
The green line which has almost the constant rate of prediction error which for this dataset is not able to perform effectively with bootstrap .632 prediction errors.

2) (10 points) A access the wine data from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/wine). These data are the results of a chemical analysis of 178 wines grown over the decade 1970-1979 in the same region of Italy, but derived from three different cultivars (Barolo, Grignolino, Barbera). The Babera wines were predominately from a period that was much later than that of the Barolo and Grignolino wines. The analysis determined the quantities MalicAcid, Ash, AlcAsh, Mg, Phenols, Proa, Color, Hue, OD, and Proline. There are 50 Barolo wines, 71 Grignolino wines, and 48 Barbera wines. Construct the appropriate-size classification tree for this dataset. How many training and testing samples fall into each node? Describe the resulting tree and your approach.

The wine dataset has been loaded and looking at the summary below.

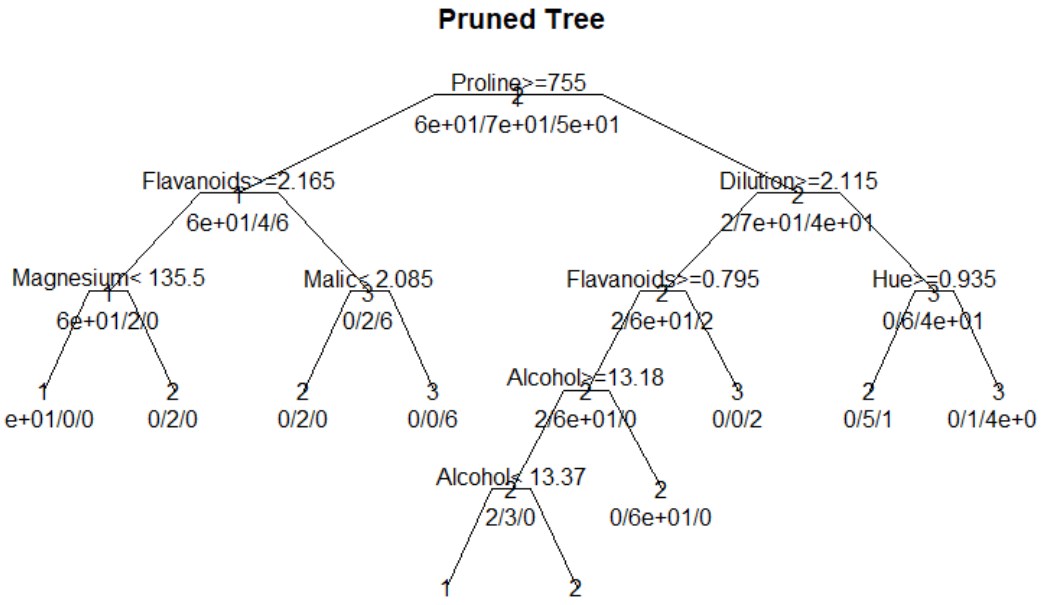
```
> summary(wine_data)
 Type      Alcohol         Malic            Ash          Alcalinity     Magnesium        Phenols      Flavanoids
 1:59   Min.   :11.0   Min.   :0.74   Min.   :1.36   Min.   :10.6   Min.   : 70.0   Min.   :0.98   Min.   :0.34
 2:71   1st Qu.:12.4   1st Qu.:1.60   1st Qu.:2.21   1st Qu.:17.2   1st Qu.: 88.0   1st Qu.:1.74   1st Qu.:1.21
 3:48   Median :13.1   Median :1.86   Median :2.36   Median :19.5   Median : 98.0   Median :2.36   Median :2.13
        Mean   :13.0   Mean   :2.34   Mean   :2.37   Mean   :19.5   Mean   : 99.7   Mean   :2.30   Mean   :2.03
        3rd Qu.:13.7   3rd Qu.:3.08   3rd Qu.:2.56   3rd Qu.:21.5   3rd Qu.:107.0   3rd Qu.:2.80   3rd Qu.:2.88
        Max.   :14.8   Max.   :5.80   Max.   :3.23   Max.   :30.0   Max.   :162.0   Max.   :3.88   Max.   :5.08
 Nonflavanoids   Proanthocyanins      Color             Hue            Dilution        Proline
 Min.   :0.130   Min.   :0.41    Min.   : 1.28   Min.   :0.480   Min.   :1.27   Min.   : 278
 1st Qu.:0.270   1st Qu.:1.25    1st Qu.: 3.22   1st Qu.:0.782   1st Qu.:1.94   1st Qu.: 500
 Median :0.340   Median :1.55    Median : 4.69   Median :0.965   Median :2.78   Median : 674
 Mean   :0.362   Mean   :1.59    Mean   : 5.06   Mean   :0.957   Mean   :2.61   Mean   : 747
 3rd Qu.:0.438   3rd Qu.:1.95    3rd Qu.: 6.20   3rd Qu.:1.120   3rd Qu.:3.17   3rd Qu.: 985
 Max.   :0.660   Max.   :3.58    Max.   :13.00   Max.   :1.710   Max.   :4.00   Max.   :1680
```

It is indicated that 1 is Barolo wine 2 is Grignolino wines and 3 is Barbera wines. Now building a tree for the given dataset



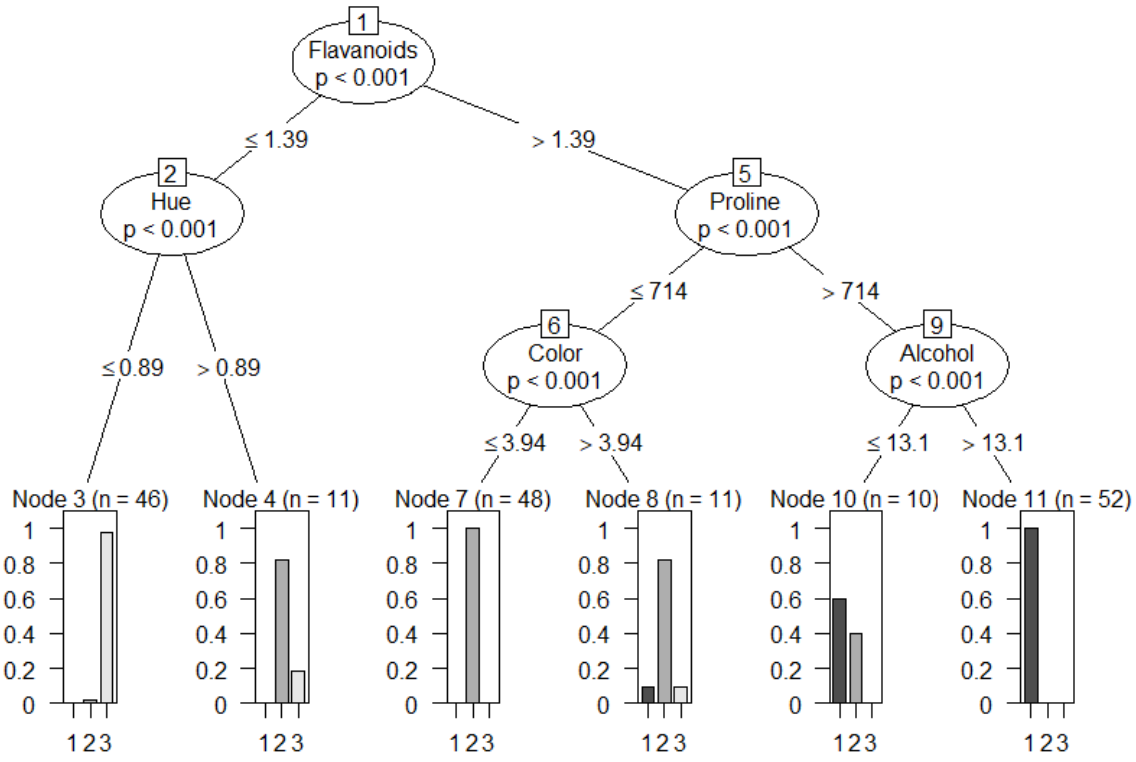
From which we can observe that the parameter Proline is prominently dependent and then the parameters Flavanoids and Dilutions the classification is dependent.

Upon further pruning the dataset to find the appropriate size classification tree we can see that the dataset forms the tree below.



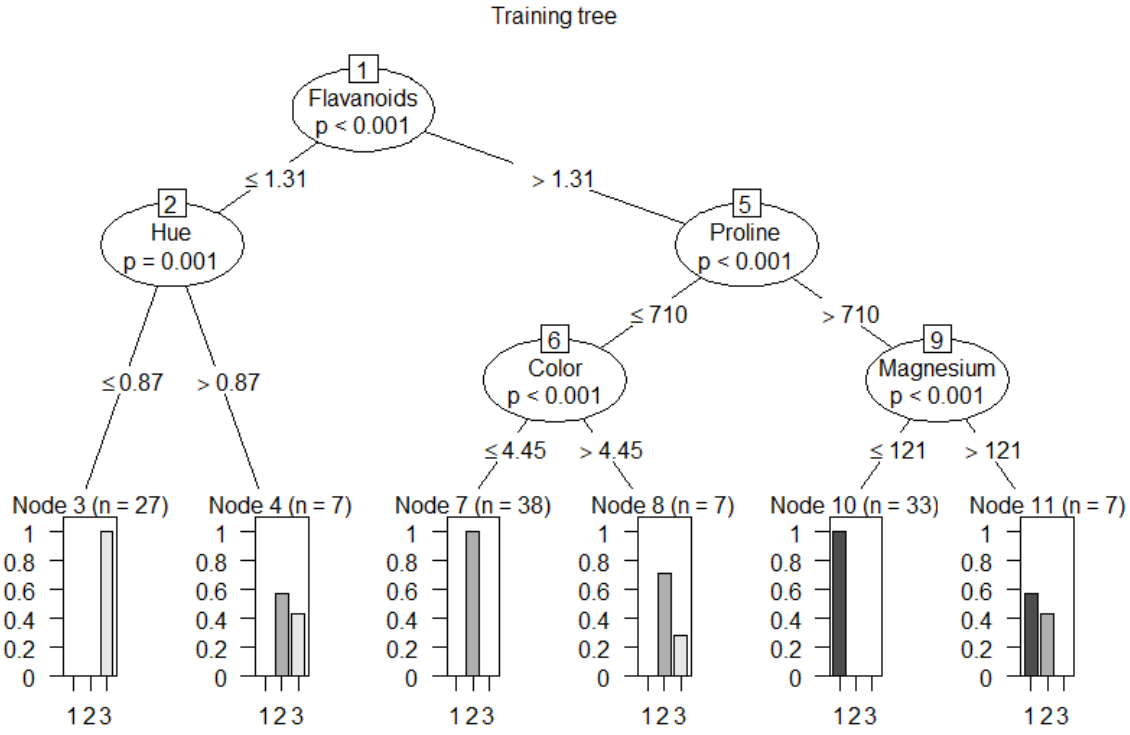
Even upon pruning the tree remains the same for the given dataset, which says that the appropriate size of the same is as obtained in the above tree.

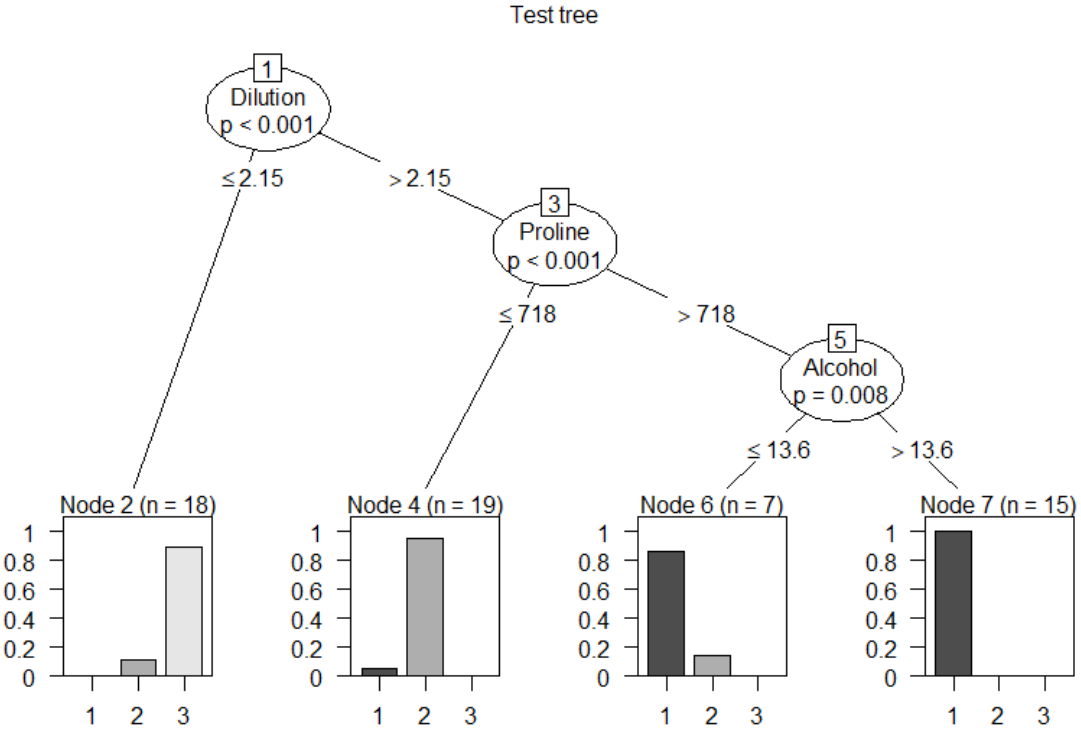
Now let us see the try to visualize the conditional interference trees for the given dataset.



From the above tree visualization we can actually see the number of nodes that this dataset is required to classify the wines type. It is about 11 nodes that are required to classify the same. Also the parameters that are influencing the classification has also been changed with the conditional approach.

Let's look at the training and testing samples trees.





Above tree diagrams clearly says that the samples will have various parameters as influential one's as the samples of training and testing changes. Hence for the same the prediction and error calculation rates will be highly fluctuating. Hence for the same prediction of one test value is feasible and then a set of testing values will always end up unpredictable.

3) (10 points) Apply bagging, boosting, and random forests to a data set of your choice (not one used in the committee machines labs). Fit the models on a training set, and evaluate them on a test set. How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?

For the dataset chosen below is the summary of the same.

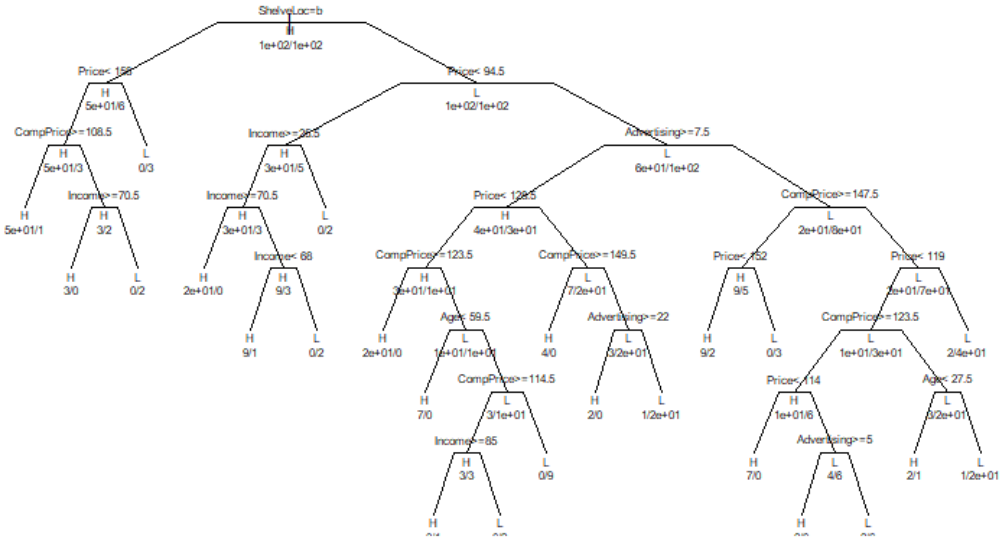
```
     Sales          CompPrice       Income        Advertising      Population      Price      ShelveLoc        Age       
 Min.   : 0.00   Min.   : 77   Min.   : 21.0   Min.   : 0.00   Min.   : 10   Min.   : 24   Bad    : 96   Min.   :25.0  
 1st Qu.: 5.39   1st Qu.:115   1st Qu.: 42.8   1st Qu.: 0.00   1st Qu.:139   1st Qu.:100   Good   : 85   1st Qu.:39.8  
 Median : 7.49   Median :125   Median : 69.0   Median : 5.00   Median :272   Median :117   Medium:219   Median :54.5  
 Mean   : 7.50   Mean   :125   Mean   : 68.7   Mean   : 6.63   Mean   :265   Mean   :116                Mean   :53.3  
 3rd Qu.: 9.32   3rd Qu.:135   3rd Qu.: 91.0   3rd Qu.:12.00   3rd Qu.:398   3rd Qu.:131                3rd Qu.:66.0  
 Max.   :16.27   Max.   :175   Max.   :120.0   Max.   :29.00   Max.   :509   Max.   :191                Max.   :80.0  
   Education    Urban       US     
 Min.   :10.0   No :118   No :142  
 1st Qu.:12.0   Yes:282   Yes:258  
 Median :14.0                      
 Mean   :13.9                      
 3rd Qu.:16.0                      
 Max.   :18.0                      
```

From which Sales data has to be predicted.

Let's convert the column to classification by ranging the values above 7 as high and below 7 as low. Splitting the data in training and test sets.

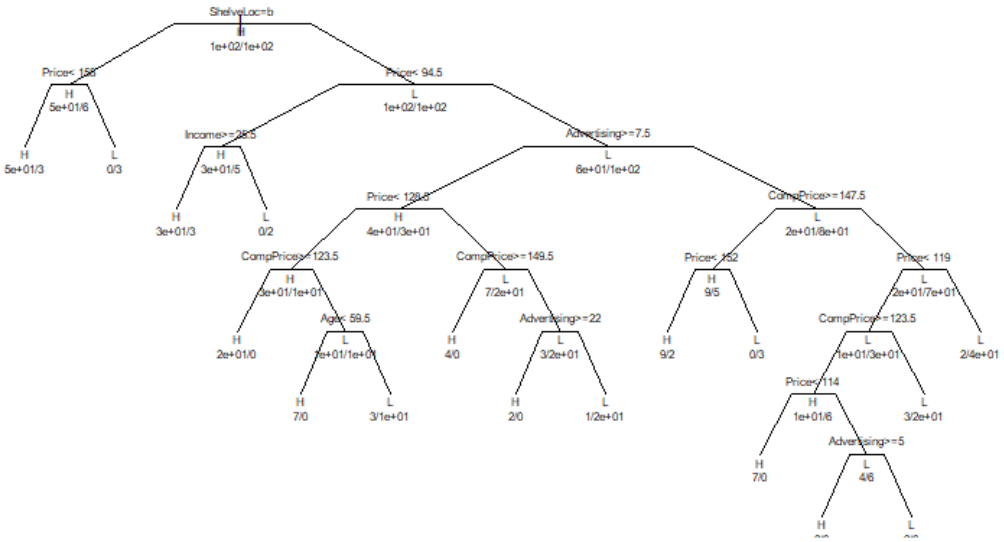
Below is the tree for the dataset taken. '

Main Tree



Upon pruning the above model for the single tree we obtain the below Pruned tree.

Pruned Tree



Computing the test error for the above tree model.

```
> miss_class_tree
[1] 0.256
```

Similarly looking at the test error for the

Random Forests -

```
> miss_class_rf
[1] 0.195
```

Bagging -

```
> miss_class_bagging
[1] 0.218
```

Boosting for 0.1 shrinkage -

```
> miss_class_bagging_1
[1] 0.158
```

It is clear that the for the given dataset Boosting performs very well with low error rate and the random forests and then Bagging.

Let's check to compare with logistic regression and KNN methods.

The error obtained using logistic regression is

```
> log_test_err
[1] 0.0902
```

And for KNN -

```
> knn_test_error
[1] 0.218
```

For us to surprise logistic regression performs better with this dataset. But when discussed on paper the Boosting rates should be low. This is completely dependent on the type of dataset. For larger set of data Boosting is expected to perform very well. But for the dataset that we have performed logistic regression works fine. For every data set depending on its type simplistic model can have advantages of having more accurate and redundant results. Where as Boosting and RF have their advantages in their suitable types of datasets.