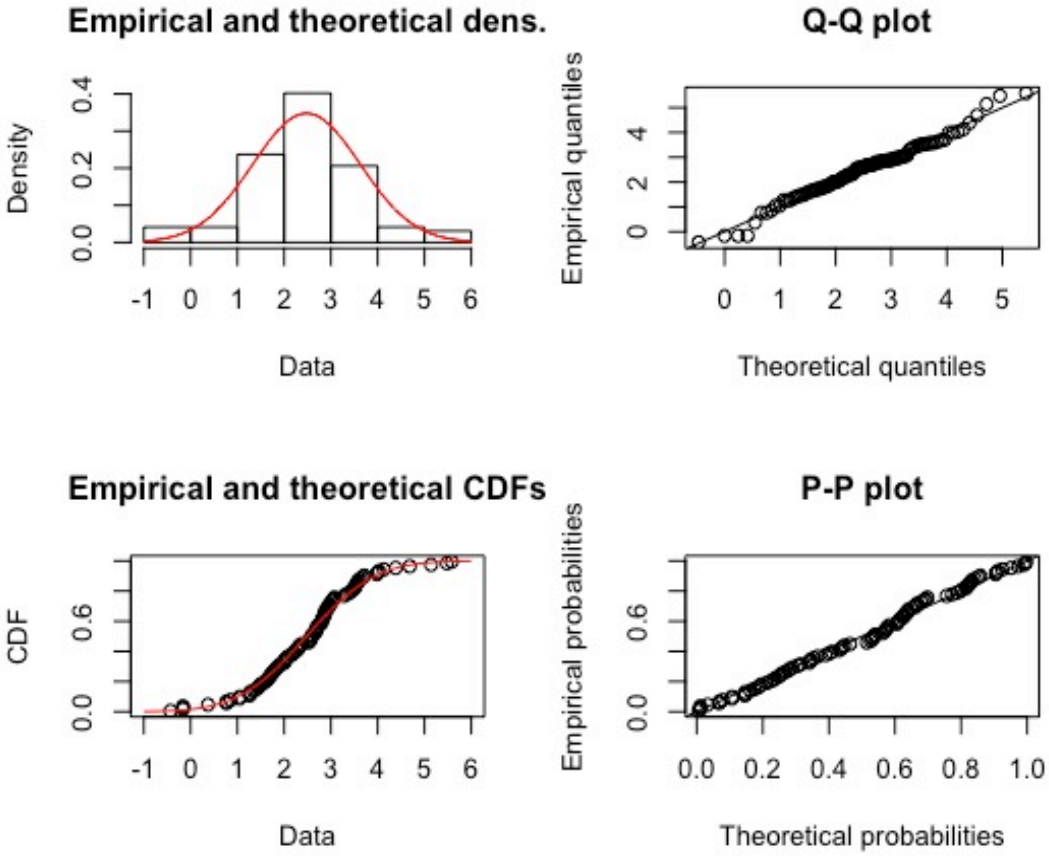
Homework 4

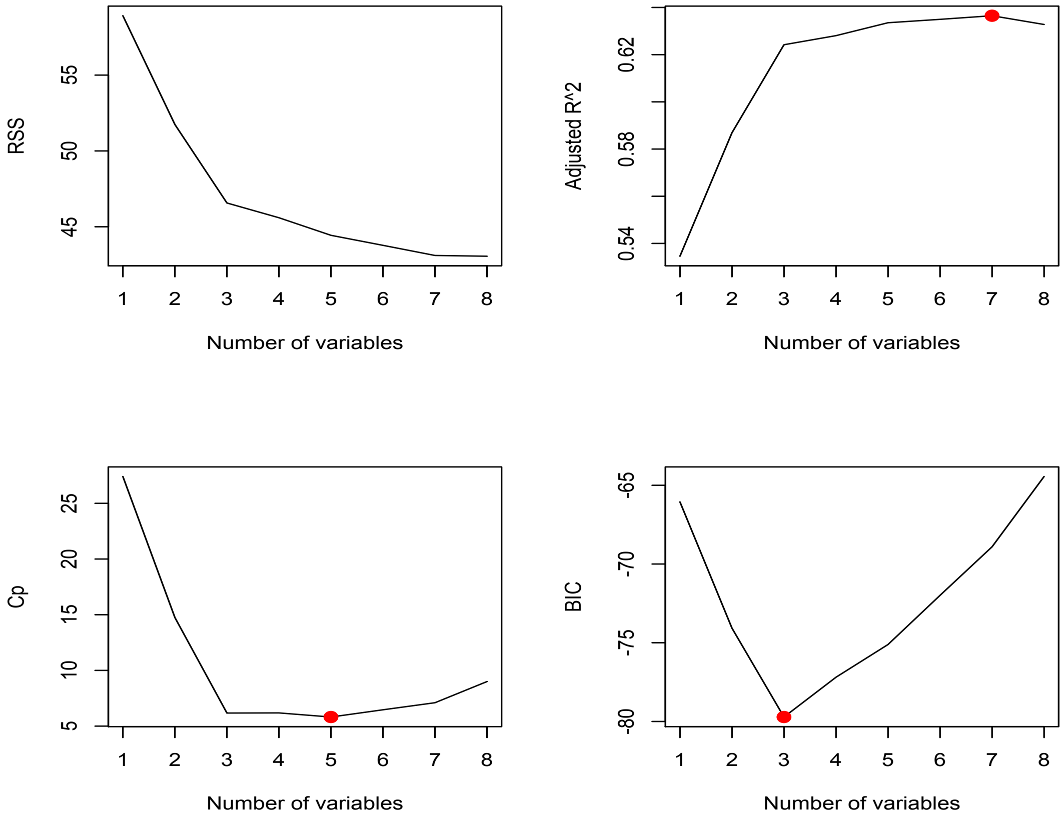
1. For the prostate data of Chapter 3, carry out a best- subset linear regression analysis, as in ESL Table 3.3 (third column from the left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap .632 estimates of prediction error.

Solution:

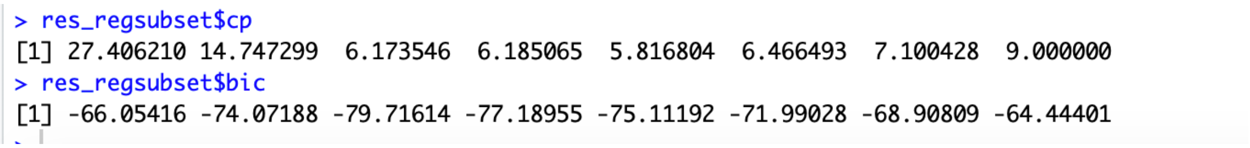
- Downloaded the 'Prostate' dataset and eliminated the last column as it is a flag indicating 'train' data.
- Checked for distribution of the data. The graph below clearly shows that the distribution is uniform normal.



- As the data is normally distributed i.e. Gaussian, by looking at above plots we can conclude that Mallows Cp will represent AIC.
- Went ahead to plot the AIC, BIC, Cp.



- From the graphs above, we can see that BIC is minimum for p=3 while Cp is minimum at p=5. However, BIC increases p=3 onwards.
- At p=3, Cp = 6.173546 and at p=5 Cp = 5.816804.
  At p=3, BIC = -79.71614



```
> res_regsubset$cp
[1] 27.406210 14.747299  6.173546  6.185065  5.816804  6.466493  7.100428  9.000000
> res_regsubset$bic
[1] -66.05416 -74.07188 -79.71614 -77.18955 -75.11192 -71.99028 -68.90809 -64.44401
>
```

- Went on to perform 5-fold and 10-fold cross validation, and bootstrap.
- For 5-fold, prediction errors were:
  0.6415488, 0.4871423, 0.5216380, 0.4215537, 0.4524955, 0.4839025, 0.4711124, 0.4758440
  The minimum error 0.4215537 was observed at p = 4.
- For 10-fold, prediction errors were:
  1.216592, 1.325291, 1.405090, 1.436872, 1.393015, 1.438767, 1.410463, 1.421770
  The minimum error 1.216592 was observed at p = 4.
- For bootstrap, the minimum error was 0.5122783 observed at p=5.
  0.6407578, 0.5629932, 0.5312754, 0.5223366, 0.5122783, 0.5271584, 0.5317484, 0.5377916

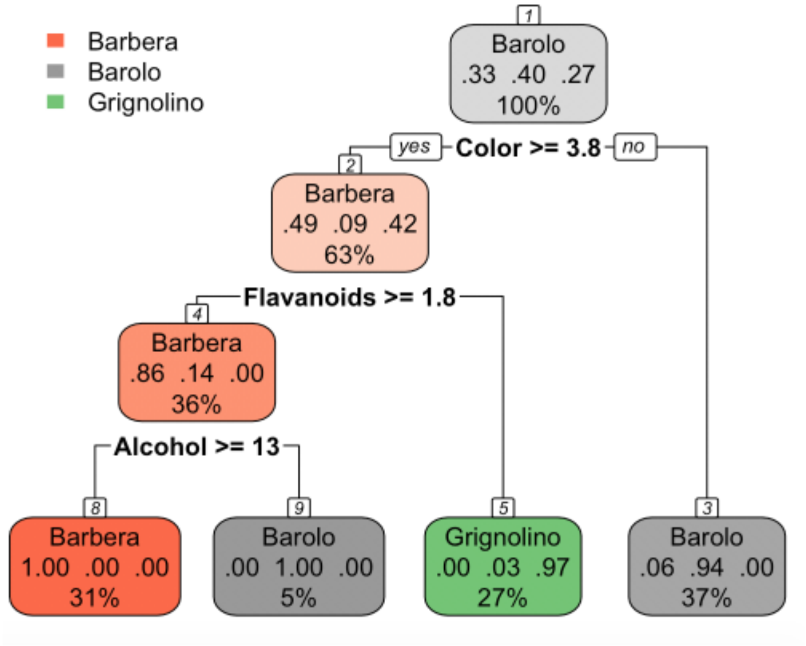
2. Access the wine data from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/wine). These data are the results of a chemical analysis of 178 wines grown over the decade 1970-1979 in the same region of Italy, but derived from three different cultivars (Barolo, Grignolino, 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.

Solution:

- The wine dataset contains 178 obs. and 14 variables.
- Used the function rpart() to fit the model on the training dataset and got the following Variable importance matrix:

V8	V14	V11	V13	V12	V7	V3	V2	V5	V10	V4	V6
17	13	13	11	10	9	7	7	6	2	2	2

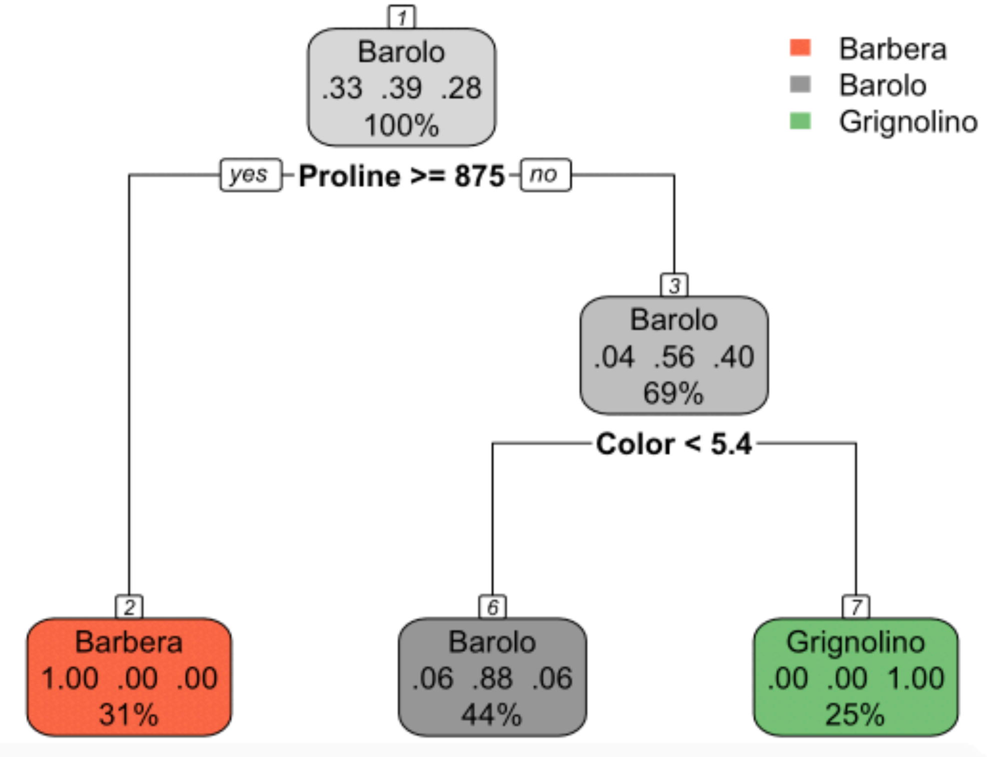
- The following tree was generated from the training set.



- The following were the observations:

	Node1	Node2	Node3	Node4	Node5	Node8	Node9
#Obs	142	90	52	44	39	44	7
Barolo	47	44	3	3	0	44	0
Barbera	57	8	49	7	1	0	7
Grignolino	38	38	0	0	38	0	0

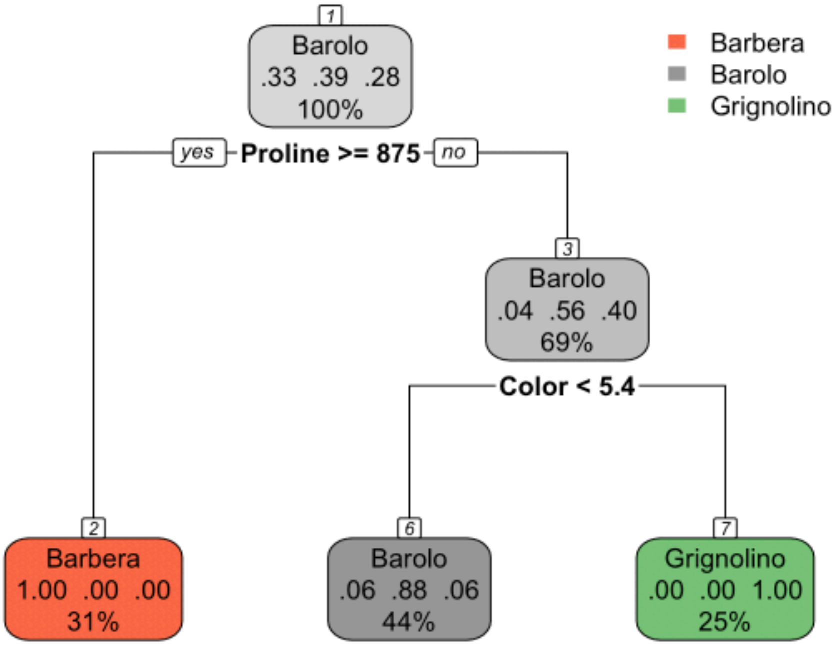
- The following tree was generated from the test set.



- Following were the observations:

	Node1	Node2	Node3	Node6	Node7
#Obs	36	11	25	16	9
Barolo	12	11	1	1	0
Barbera	14	0	14	14	0
Grignolino	10	0	10	1	9

- On Pruning, following results were observed:



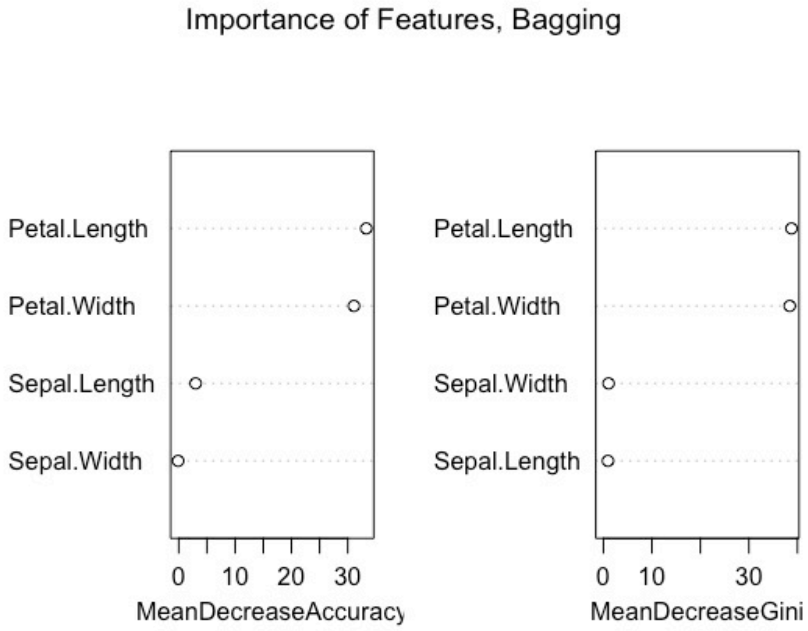
	Node1	Node2	Node3	Node6	Node7
#Obs	36	11	25	16	9
Barolo	12	11	1	1	0
Barbera	14	0	14	14	0
Grignolino	10	0	10	1	9

- **Observation –** Here the full tree is very good so pruning does not have effect on it. Pruning means to eliminate leaves that are not increasing the accuracy significantly thereby, to prevent overfitting.

3. 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?

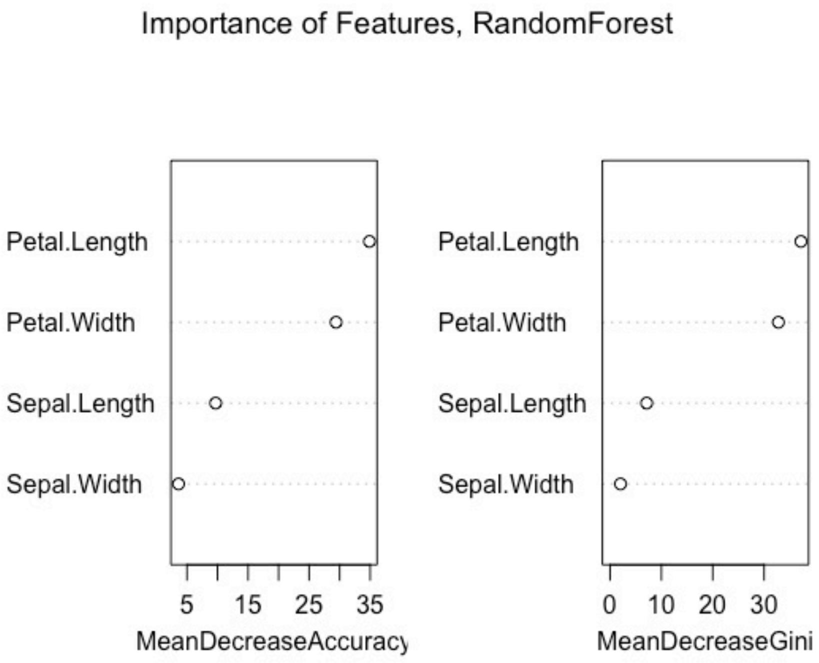
Solution:

- Selected the Iris dataset. It has 150 observations and 5 variables.
- Performed Ensemble Methods on the dataset.
- **Bagging**:

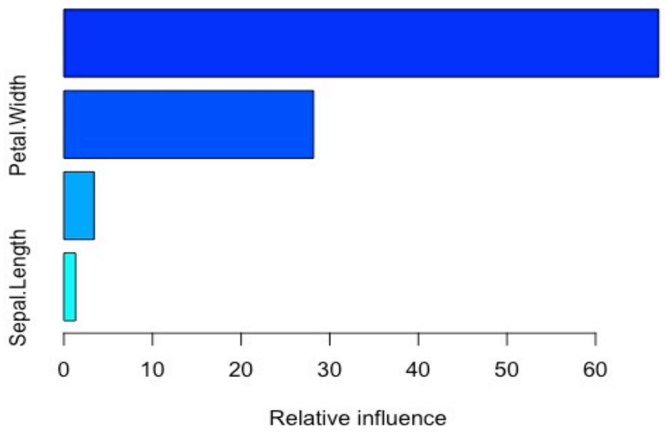


  - From the graph above it can be seen that Petal.Length and Petal.Width are the most important features.
  - On fitting the model on test data, the accuracy was 96.67%.
  - The RMSE was 0.1825742.

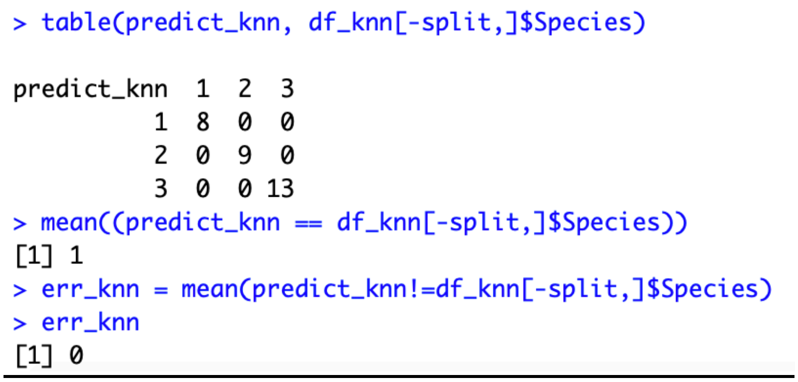
- **Random Forest:**



  - From the graph above it can be seen that Petal.Length and Petal.Width are the most important features.
  - On fitting the model on test data, the accuracy was 96.67%.
  - The RMSE was 0.1825742.
  - This result was the same as bagging.
- **Boosting**



    - From the graph above it can be seen that Petal.Length and Petal.Width are the most important features.
    - The RMSE was 8.944468.
- **KNN**



```
> table(predict_knn, df_knn[-split,]$Species)

predict_knn  1  2  3
          1  8  0  0
          2  0  9  0
          3  0  0 13
> mean((predict_knn == df_knn[-split,]$Species))
[1] 1
> err_knn = mean(predict_knn!=df_knn[-split,]$Species)
> err_knn
[1] 0
```

- According to me, the screenshot above is the most important result from this experiment.
- It is clear that there is no misclassification error.
- KNN predicts this test set accurately.

Note: Cannot use Logistic Regression because dataset has more than 2 classes in response variable.

Advantages of committee machines:

1. Committee methods take a simple unweighted average of the predictions from each model, essentially giving equal probability to each model.
2. Bagging improves prediction accuracy.
3. Random forest can handle high dimensional spaces as well as large number of training examples.

Disadvantages of committee machines:

1. When we bag a large number of trees, it is no longer possible to represent the resulting statistical learning procedure using a single tree, and it is no longer clear which variables are most important to the procedure. Thus, bagging improves prediction accuracy at the expense of interpretability.
2. Boosting is Time and Computationally expensive.