Q1. We load the prostate dataset from ElemStatLearn package for our analysis. We can see that there is a column train which can be used to decide our train and test datasets. We carry out best subset selection for linear regression and store the output matrix which contains the order of the variables used for best subset selection.

a) We compute the AIC of the different models for a different combination of variables. The following are the values of AIC. We can see that the model with the least AIC is a 7 variable model

1	2	3	4	5	6	7	8
24.766	12.108	9.803	7.679	8.209	7.194	7.021	9.00

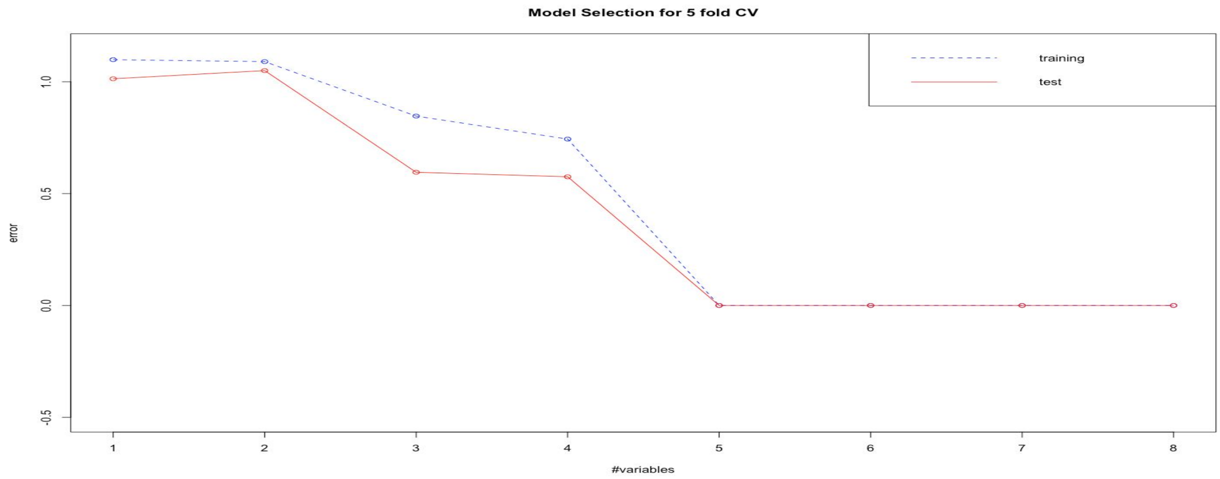
The test mean squared error for the 7 variable model is: 0.516

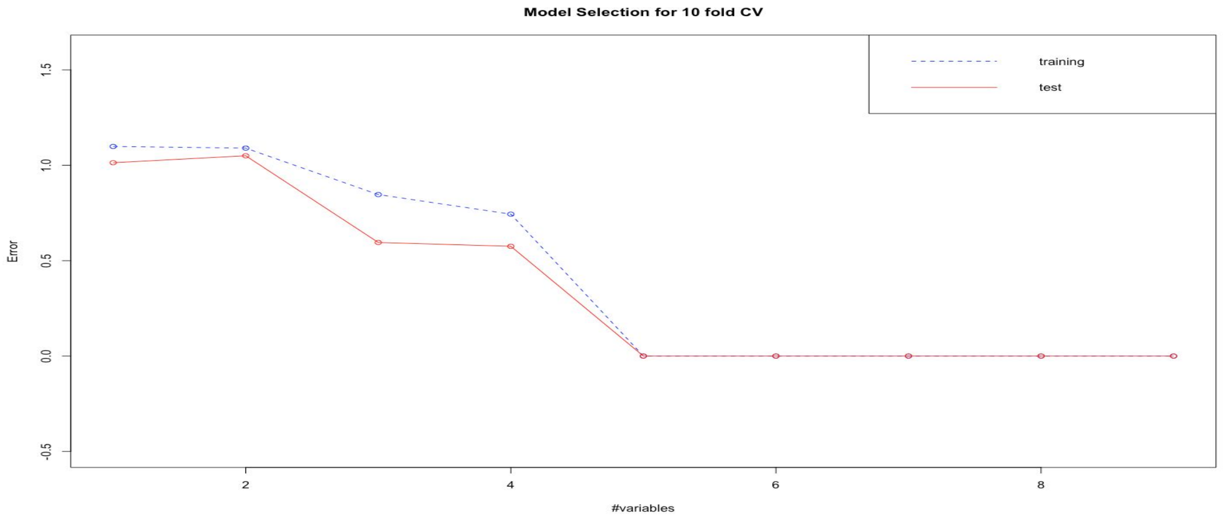
b) We compute the BIC of the different models for a different combination of variables. The following are the values of BIC. We can see that the model with the least BIC is a 2 variable model

1	2	3	4	5	6	7	8
-43.35	-51.29	-51.15	-51.09	-48.42	-47.49	-45.75	-41.57

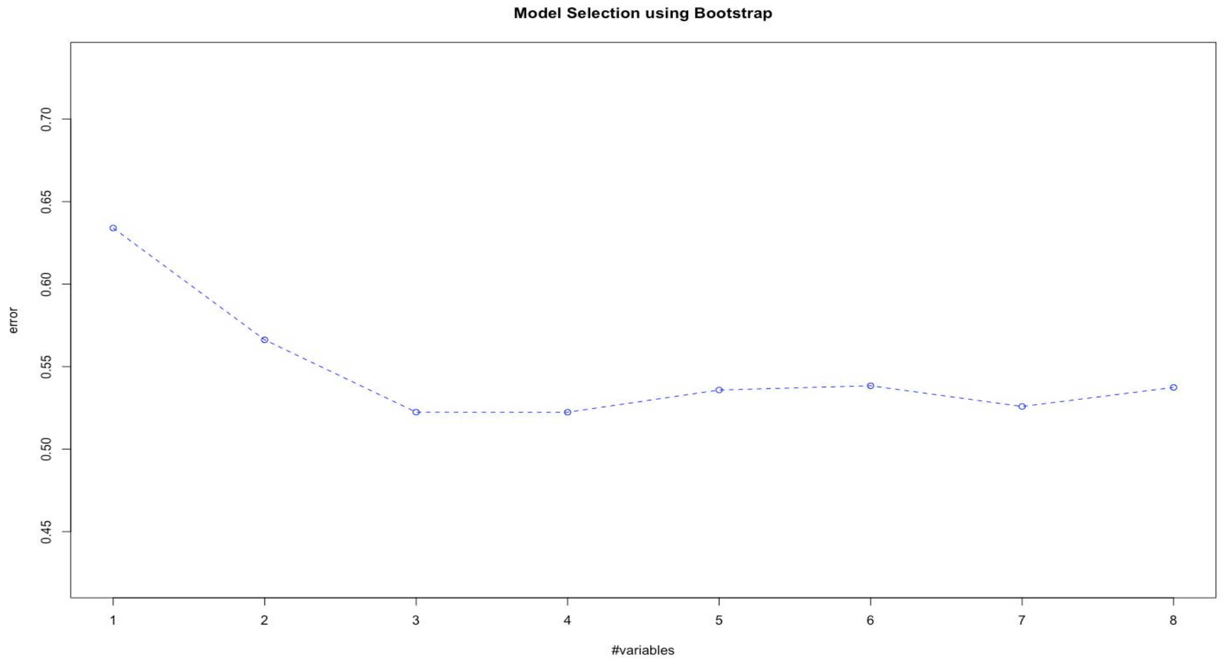
The test mean squared error for the 2 variable models is: 0.492

c) We select the variables in the order using best subset selection for linear regression. We run 5 fold cross-validation and 10 fold cross-validation for each combination of variables and see that error is close to zero after 5 variables. Hence the best model is one with 5 variables as we can see from the plot below.



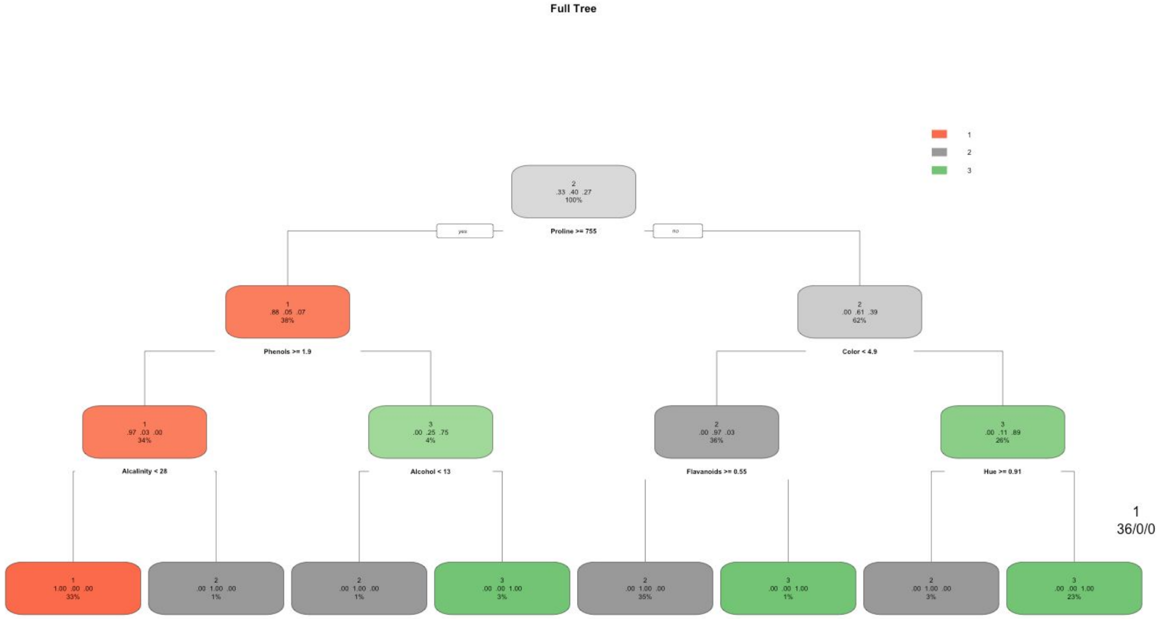


d) We run subset selection via bootstrap and see that error is lowest for a model with 3 variables as shown below

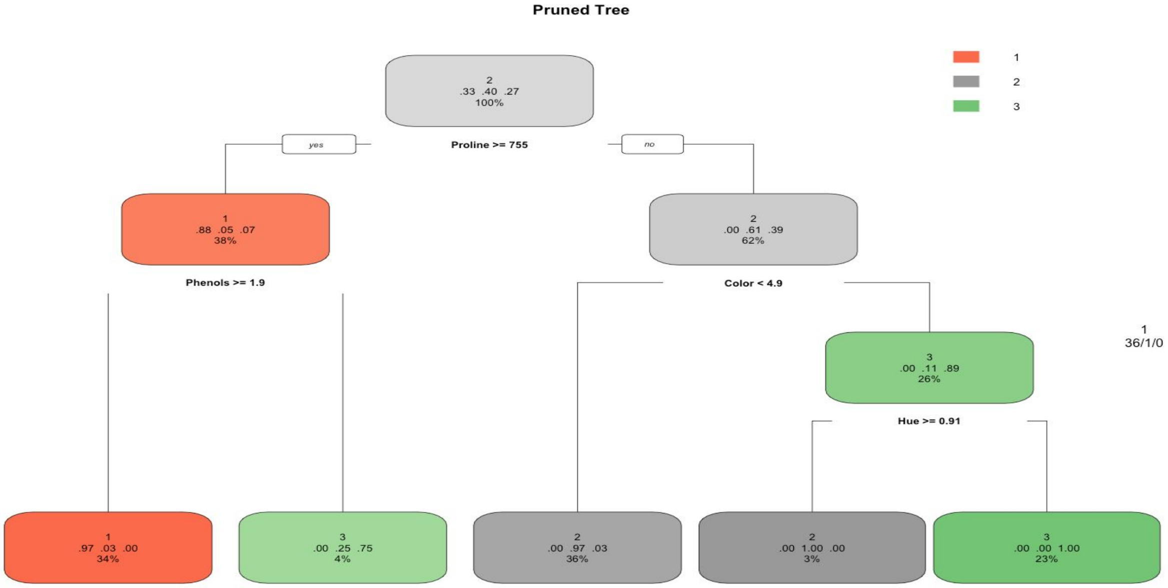


Q2. The full tree for the wine dataset is as shown below:

Class 1 is Barolo, 2 is Grignolino and 3 is Barbera

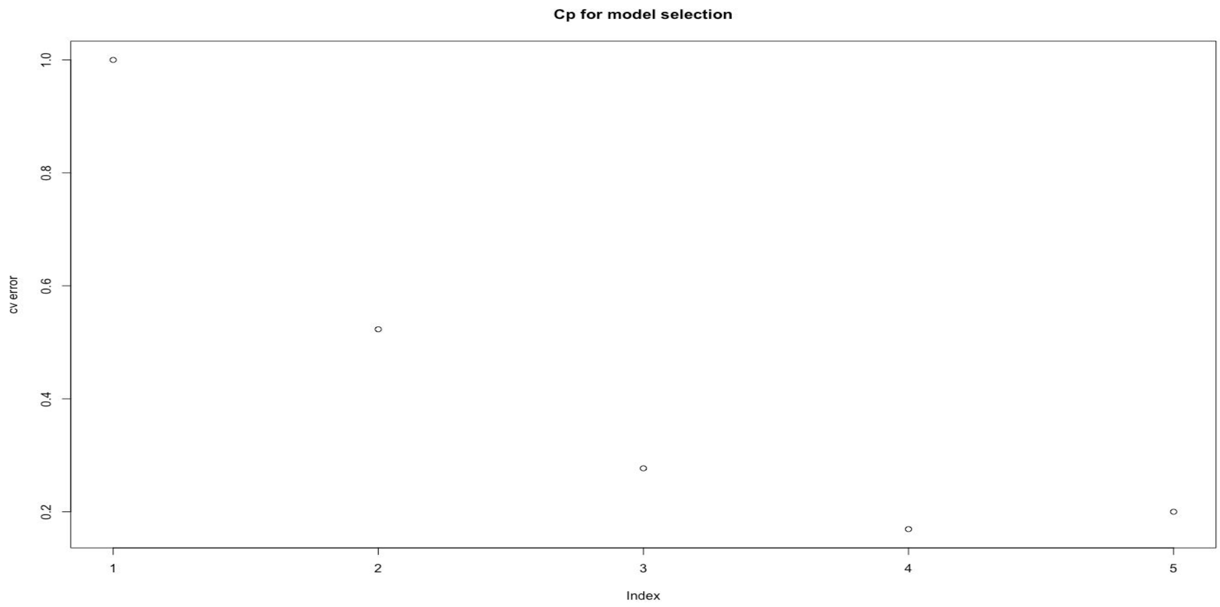


The pruned tree for the wine dataset is as shown below:



There are 5 nodes in the pruned tree at the end. The number of testing samples in each of the nodes are as follows:
Node 1: 36 Barolo and 1 Grignolino
Node 2: 1 Grignolino and 3 Barbera
Node 3: 38 Grignolino and 1 Barbera
Node 4: 3 Grignolino (A pure node)
Node 5: 25 Barbera (A pure node)

The resulting tree has been achieved after pruning it using a value of complexity parameter (cp). We use cross-validation to obtain the appropriate cp value which will determine the threshold for improvement of the fit for extra attributes. The following graph shows the cv error. We choose cp with the lowest cv error.


Cp for model selection

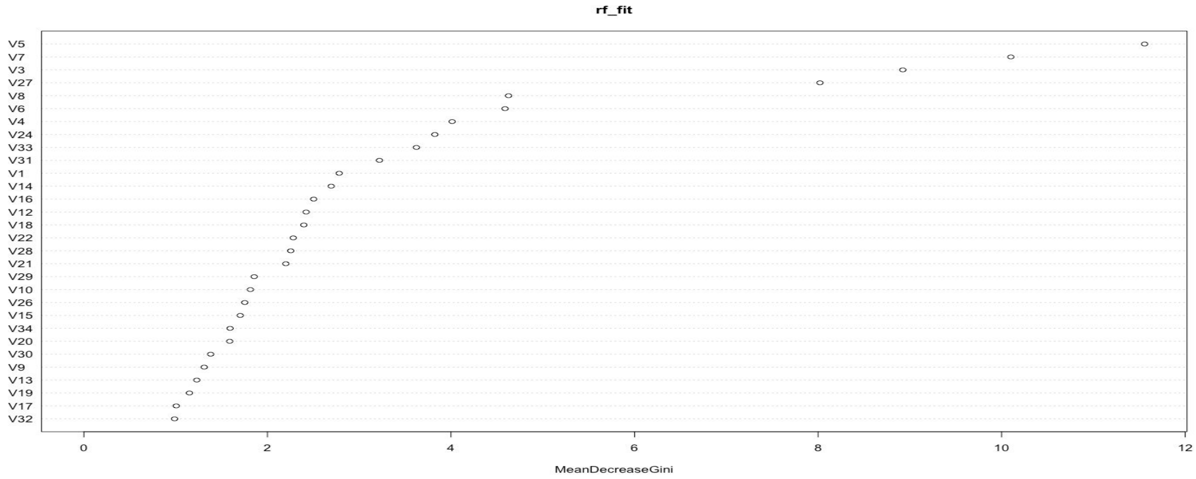
The cp value corresponding to the lowest cv-error is 0.01538462 which we will use to prune the full tree. The training error and test errors are as follows:

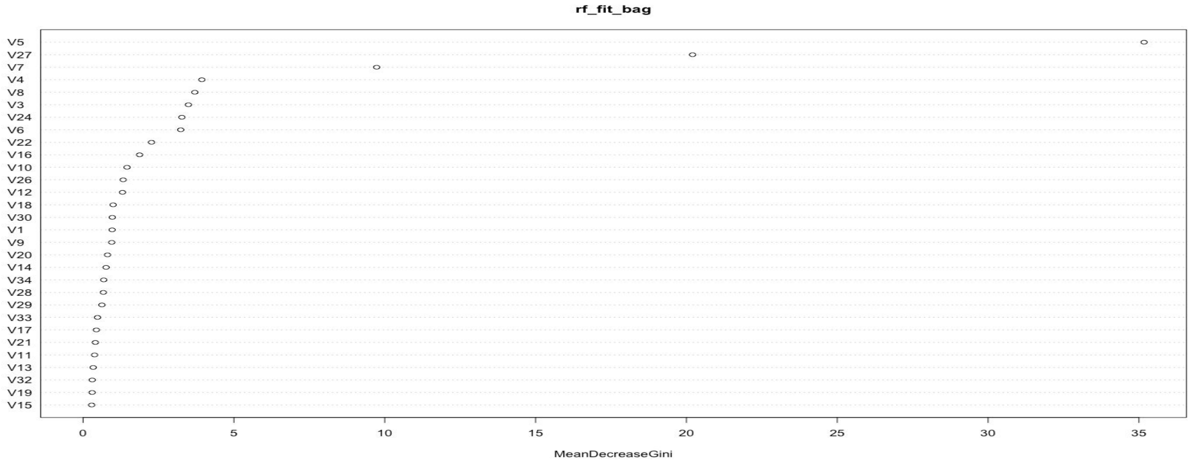
	Train Accuracy	Test accuracy
Full Tree	1	0.8714
Pruned Tree	0.9722	0.8714

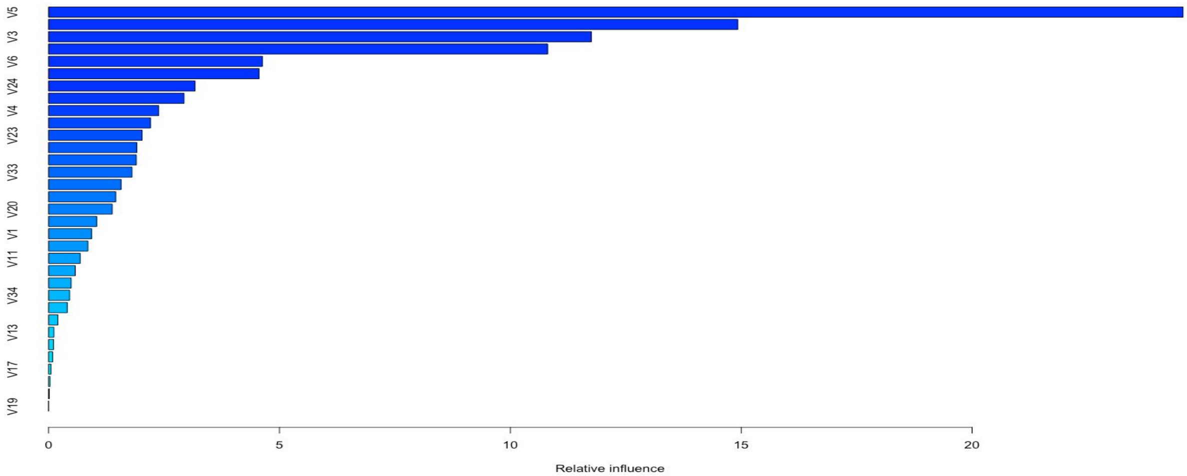
Q3. The Ionosphere dataset has been chosen to test the performance of ensemble and non-ensemble methods. The dataset has 35 variables out of which 1 is removed as the whole column has only 0 values. The dependent variable is Class which takes the values good or bad with a split of 36%. The following are the results of the various models (logistic, knn, RandomForest, bagging and Boosting) on the dataset.

Model	Train Accuracy	Test Accuracy
Logistic	1	0.84
knn (k=3)	0.95	0.92
Knn (k=21)	0.88	0.87
Random Forest	1	0.92
Bagging	1	0.91
Boosting (d=3,s=0.005)	0.95	0.93

The important variables for the random forest, bagging and boosting modes are shown as shown below:

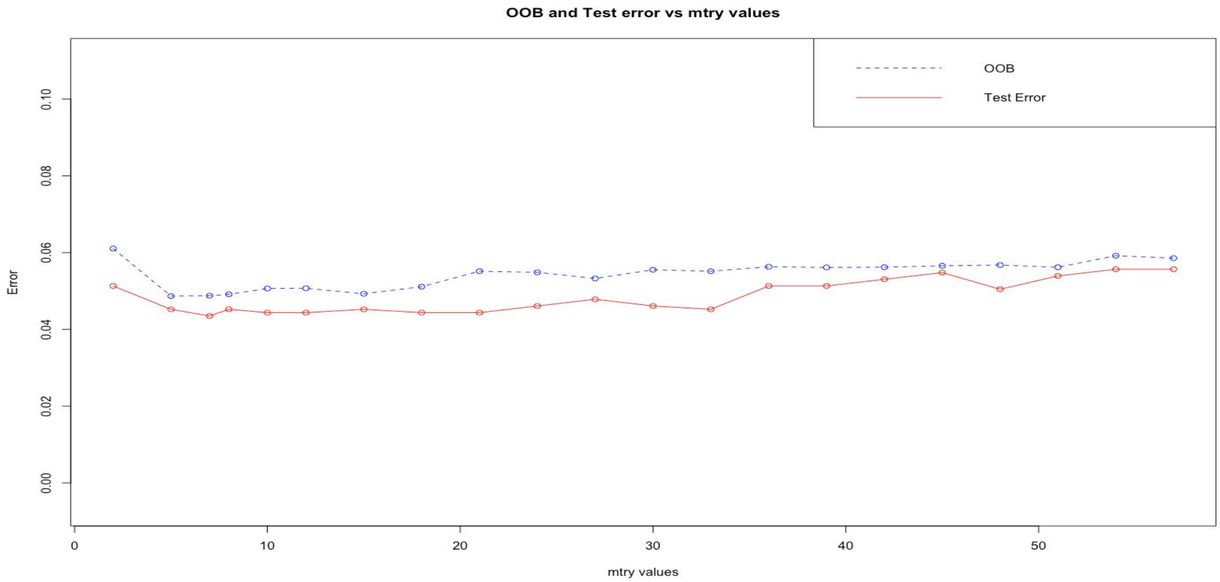






The advantage of the committee machines(ensemble methods) is that they generalize better on the test data compared to simplistic models such as logistic regression. Knn does generalize well for test data and its performance is comparable to committee machines in high complexity(low k values) and its performance decreases for higher k values compared to committee machines. Also, we can see from the summary of logistic regression that none of the variables has come out significant. The disadvantage of committee machines is that we lose in terms of interpretability. Although we can see the relative importance of various independent variables, we can’t calculate the weight of each variable on the dependent variable.

Q4. We load the SPAM dataset and store the test error and OOB for each value of mtry. We can see that the value of mtry for which test error is low is 7 which is the floor(sqrt(#variables)). Also, we can see from the graph that oob and test error are virtually the same so we can see that oob is a valid estimate of the test error.



Q5. The randomness of random forest arises from the way the decision tree is built. Each time a split is considered in a tree, a random sample of m predictors from p predictors are chosen as the split candidates and a fresh sample of m predictors are chosen at each split. This helps in decorrelating the trees and makes the average of trees less variable. If we chose m=#predictors it will be nothing but bagging.