P2-WineData

Jeetendra Gan

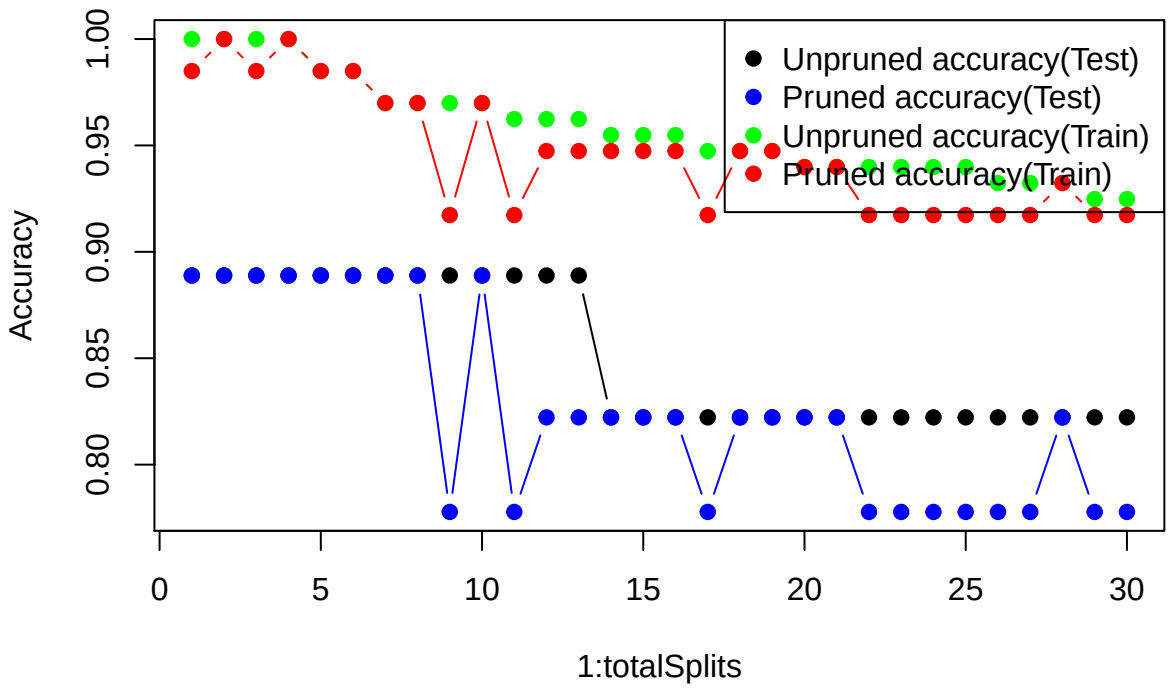
11/30/2019

I have divided the data set into 75% train and 25% test.

```
##   Classes % in Train % in Test
## 1       1   33.08271  33.33333
## 2       2   39.84962  40.00000
## 3       3   27.06767  26.66667
```

The above table shows that the classes are distributed evenly across train and test (i.e. the ratio of class 1(and others) is same in both train and test), although there is a little imbalance which is also present in the main data.

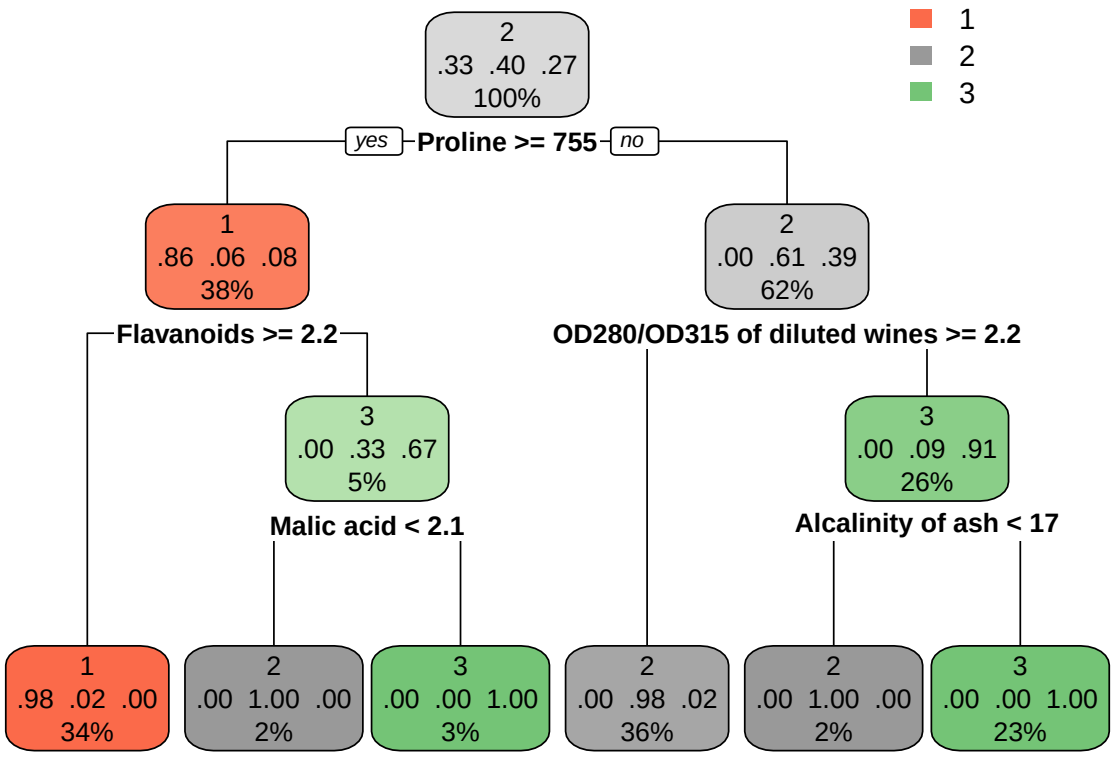
Approach used to select the model.



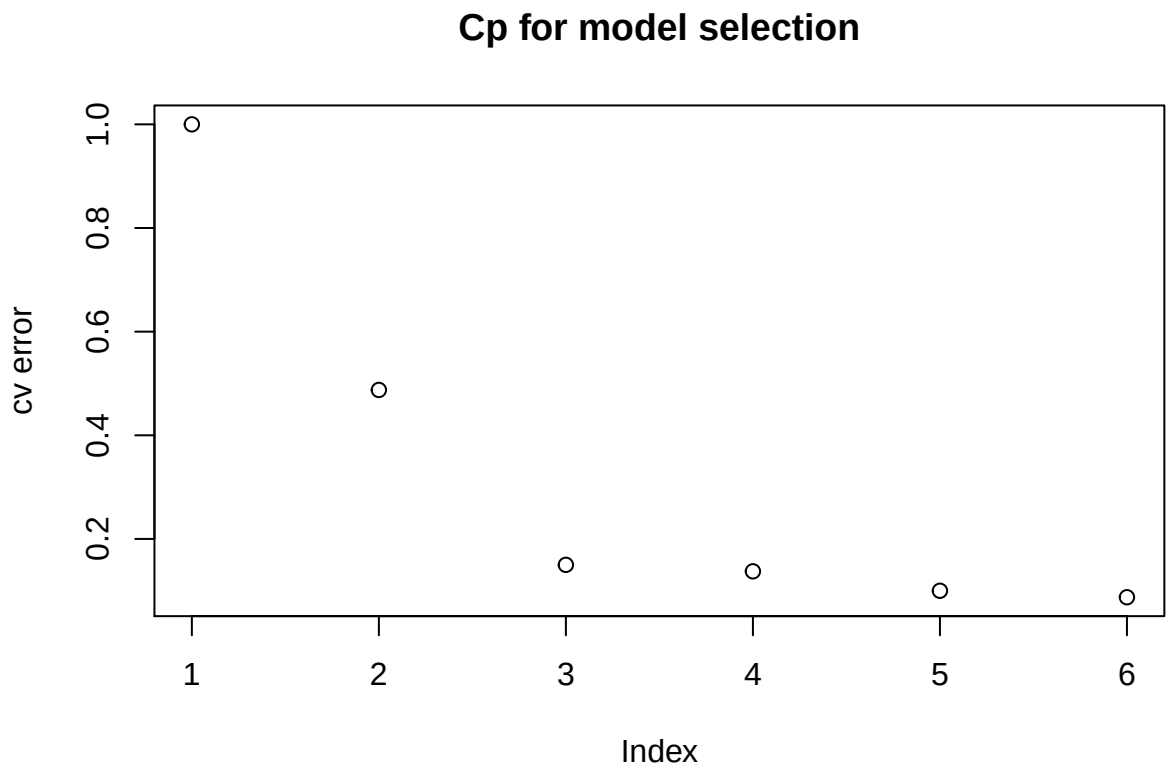
I have trained the model over various values of min-split, ranging from 1 to 30. For each value of min-split, I have calculated the training(Unpruned accuracy-train), and test accuracy(Unpruned accuracy-Test). Also, for every model, the min cp has been found out. The min cp is used to prune the tree. After the tree is pruned, both, test(Pruned accuracy-test) and train accuracy(Pruned accuracy-train) are recorded. Each of these four metrics are shown in the graph above.

It can be seen that for most of the split values less than 5, the training accuracy is 1 for both pruned and unpruned trees. Whereas the test accuracy is less for both pruned and unpruned trees. This is a clear indication of overfitting. This is one reason to choose a value greater than or equal to 5. After 6, the training accuracy decreases(for both pruned and unpruned trees), so does the test accuracy. From 7 to 12, the test accuracy for the pruned tree fluctuates and hence unreliable. It goes down below 83% for larger values of min-split.

It is better to choose the min-split value between 5, and 6. Below, I have selected the min-split value for 5.

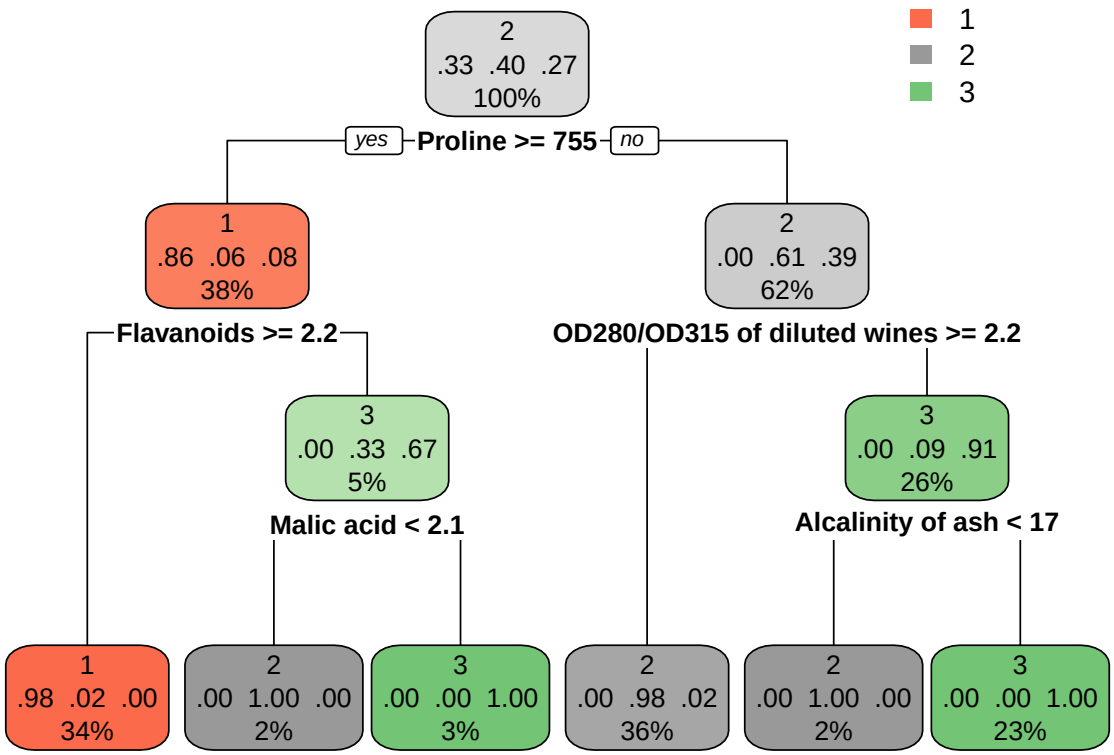


This is an unpruned tree with a min-split value of 5.




```
## [1] "The minimum cp value from the graph is: 6"
```

This value can be used to prune the tree. The tree shown below is result after passing into the prune function. It can be seen that no pruning has been done.



How many testing samples fall into each node?

There are 6 left-nodes in all. There is 1 node that classifies 44 1's correctly. It also classifies a 2 as a one. THere are three nodes that are designated for 2's. Two of those are pure nodes, with two correct classifications. The third class-2 node classifies 47 2's correctly and one three incorrectly as a two. There are two nodes with class-3. Both of them are pure nodes. One of them classifies 4 three's correctly and the other classifies 23% data correctly.

The main seperator node is Proline. If proline is greater than 754, then flavanoids are used to split the data. If flavenoids are greater than 2.2, then the data point belongs to class 1. Else, malic acid is used as a seperator. If malic acid is greater than 2.1, then the data is classified in class 2, else in class 3.

If proline is less than 755, OD280/OD315 is used as a separator. If it is greater than 2.2, then the data point is classified to be 2. Else Alcalinity of ash is used as a separator. If it is less than 17, then we can be 100% certain that the data point belongs to class 2. If it is greater or equal to 17, the data point belongs to class 3.