a)

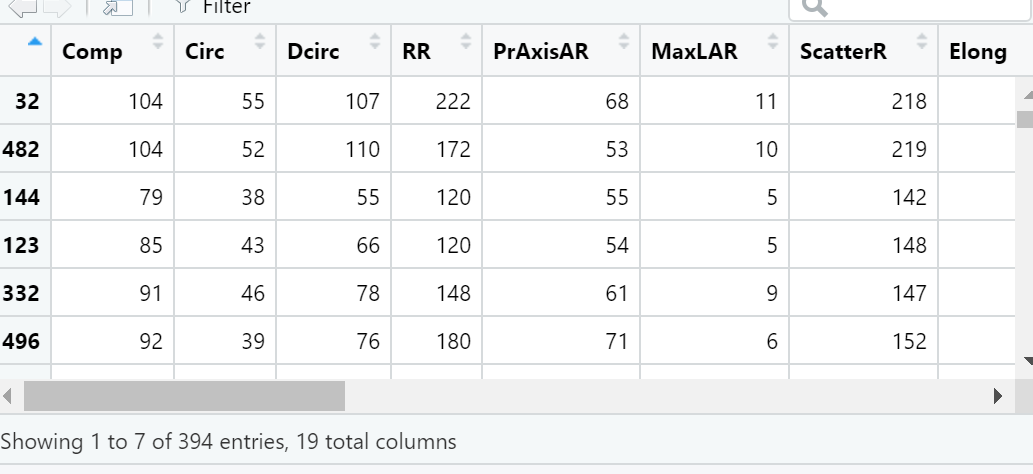


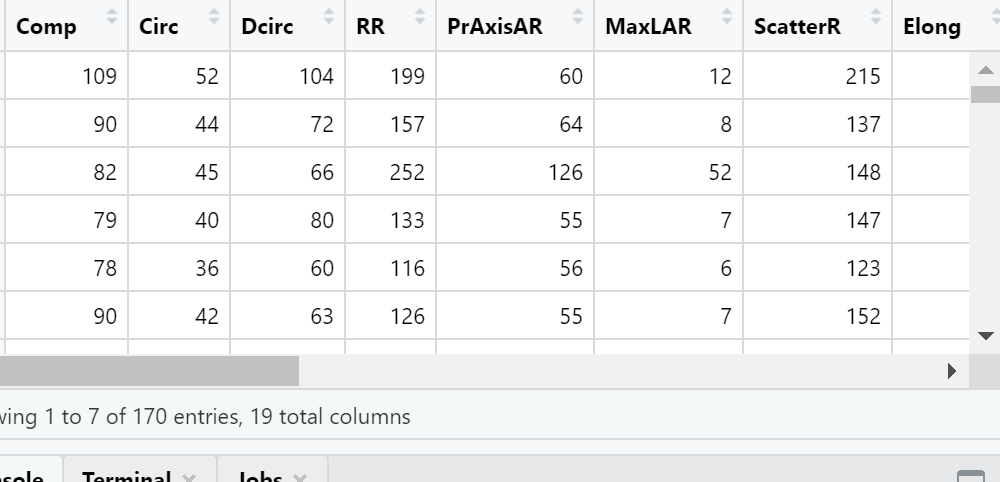
Process:

1. **Read data** : I read ‘vehicle’ dataset.
2. **Made the Train and Test data**: I divided 70% of data into train and 30 % of data into test.
3. **Made a pruned tree model:** I made a tree model and show the cp value. With the cp value, I can findout the tree size that maximizes the model’s explanation. Also, I compared the pruned tree with the full tree.
4. **Compare the pruned tree and the full tree:** I made a model with full variables to compare with the pruned tree model. And I made two confusion matrixs to compare two models.

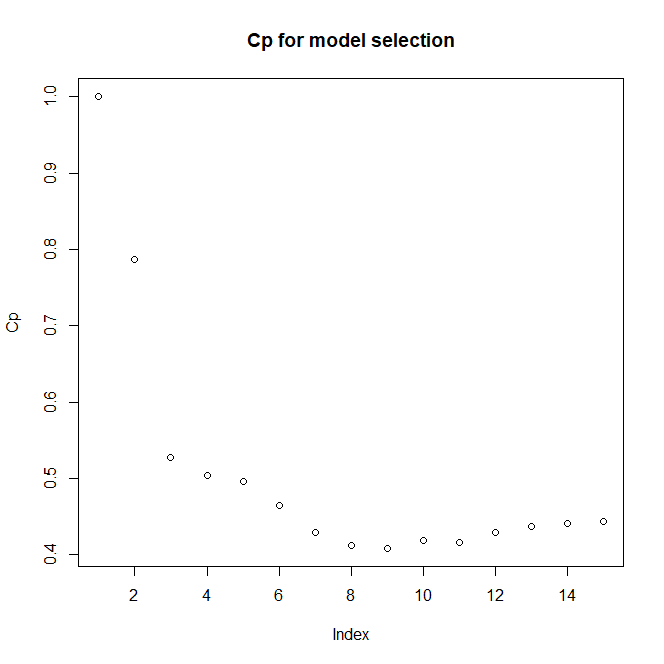
Output:

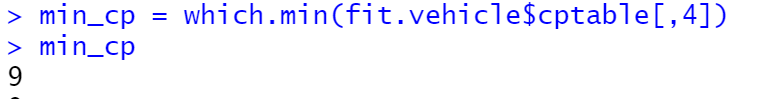
train



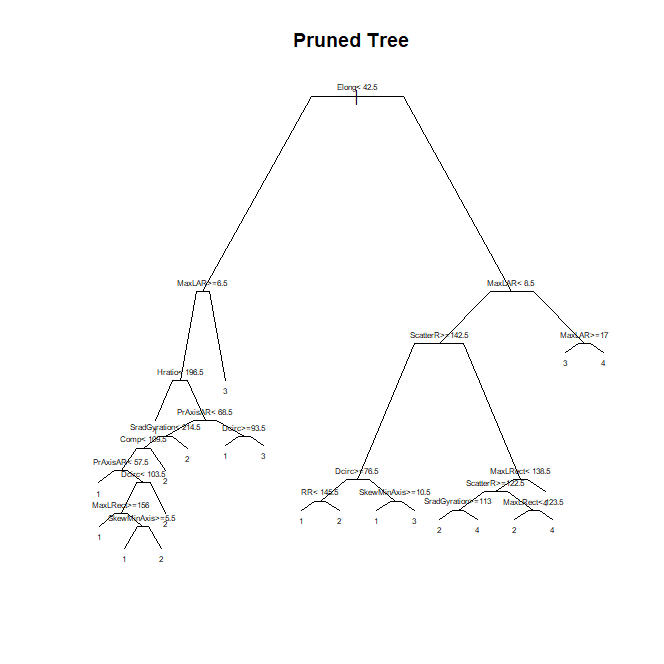


**Cp for model selection with the best subset selection model.**

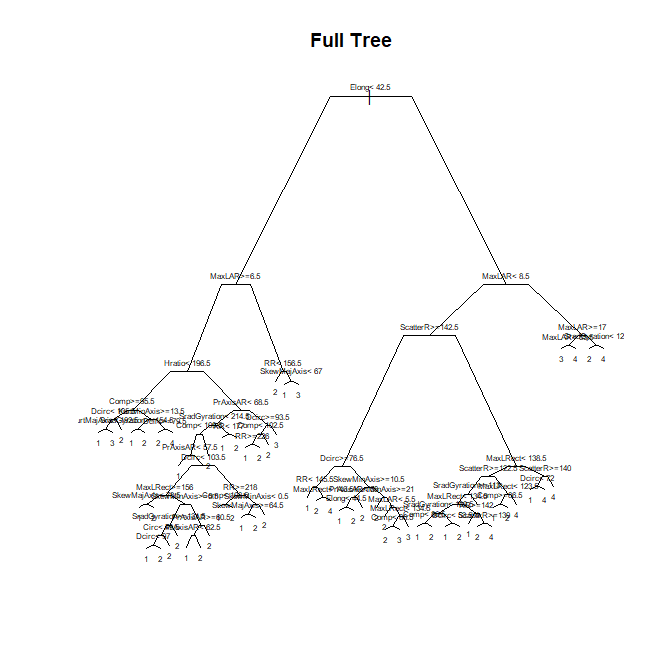




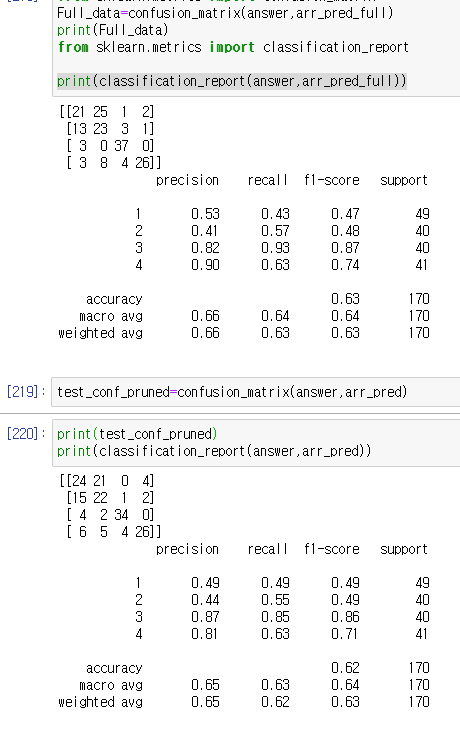
**Pruned Tree shape**



**Full tree shape**



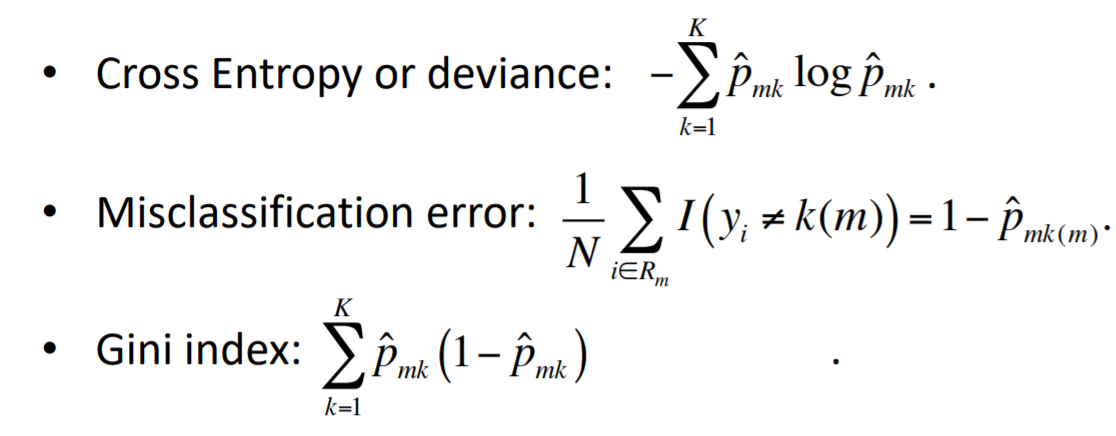
Confusion matrix for the full data and the pruned data



Discussion:

***Partitioning***

The tree model classifies the data, so that the same class of the data supposed to be at the same area. The tree model tries to reduce the impurity of each area. In other words, a low impurity area consists of objects in the same class. To calculate the impurity, the model uses some indicators such as ‘Entropy’, ‘Gini’ and ‘misclassification error’.



The tree model recursivly devide the data with creteria that minimize the area’s impurity. Finally, we got the Full tree.

***Pruning***

The full tree can show good performance in the training data; however, it is possible that the model is too overfitted in the train data. To generalize the model, we need to prune the tree model. To find out the fitness pruned tree, we used minimum cp value in the model which maximizes the model’s explanation.

***Accuracy***

When we compare the accuracy of full and pruned tree, it is quite similar. 63% for the full tree, and 62% for the pruned tree. However, the pruned tree is the more general model, it is possible that the pruned tree model can make better performance in a totally different data set. (Train and Test set comes from same data set)