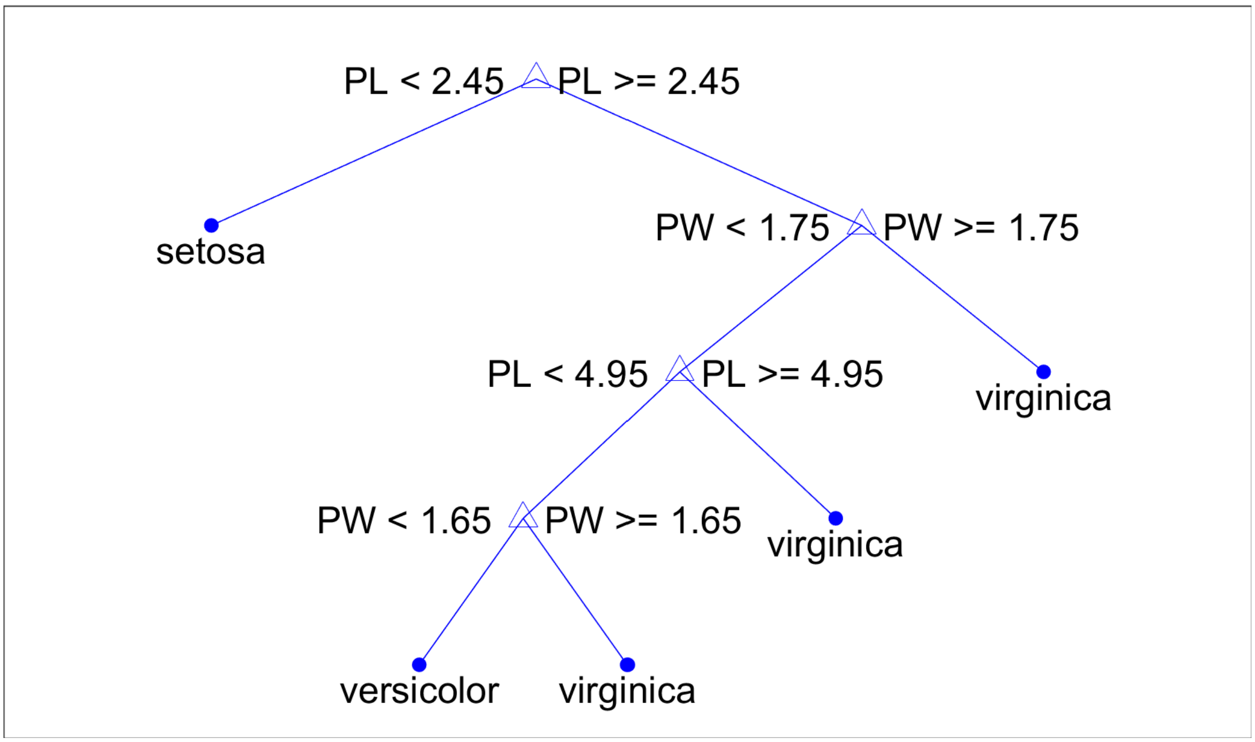
Homework 12

Tamoghna Chattopadhyay

Question 1

(a) Before Pruning, dtCVErr = 0.2933
After Pruning, dtCVErr = 0.2467

(b)



dtCVErr = 0.0533

(c)

The depth of the unpruned modified tree of (b) (based on full feature set) is smaller than the unpruned original tree of (a) (based on the original feature set). This is because when we use the full feature set, we find a better feature to better divide the dataset into the two classes. This division is better and easier than the one done by the two features in case (a). Hence, the depth of the tree is smaller.

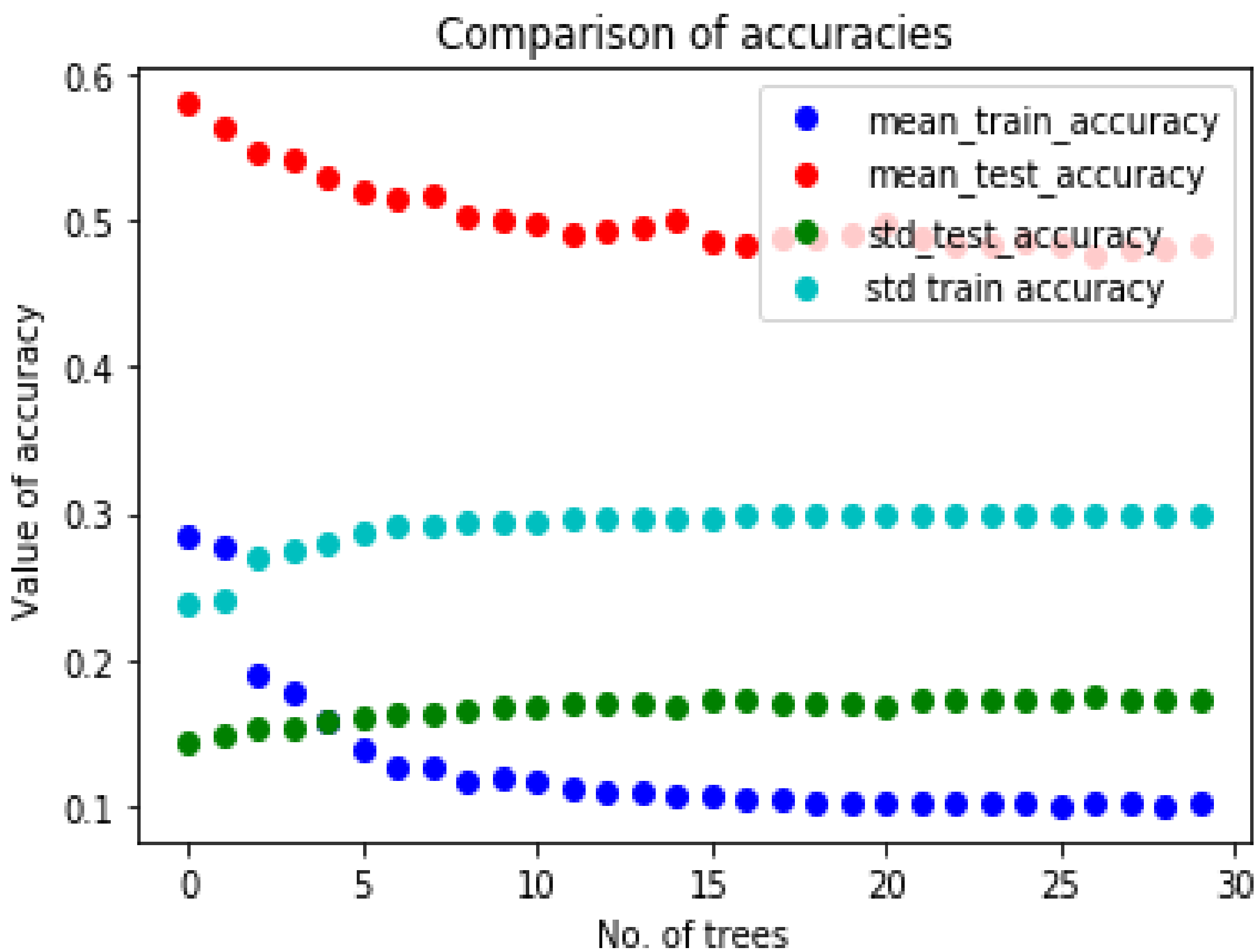
(d)

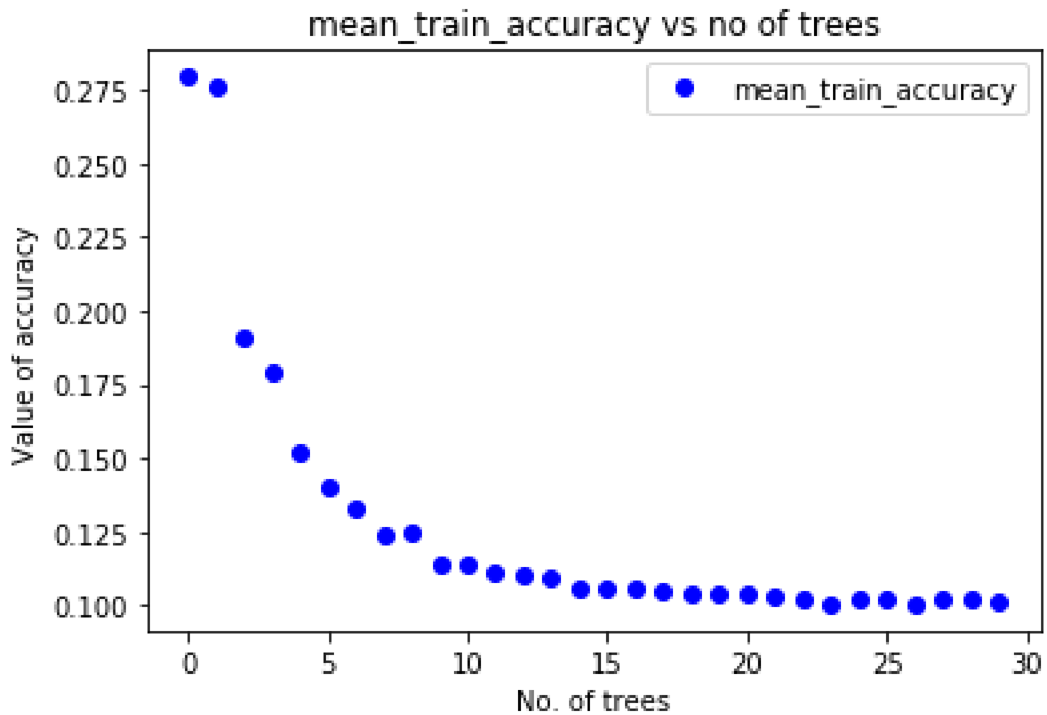
When we compare the cross validation error of modified tree in part (b) to the original tree in part (a), both before and after pruning, we find that the value of cross validation error is much lesser in the case of (b). This is because when we select more features, we can create more regions to separate which decreases the error. As we have more features, we get more dimensions to decide on the regions, so the cross validation error goes down.

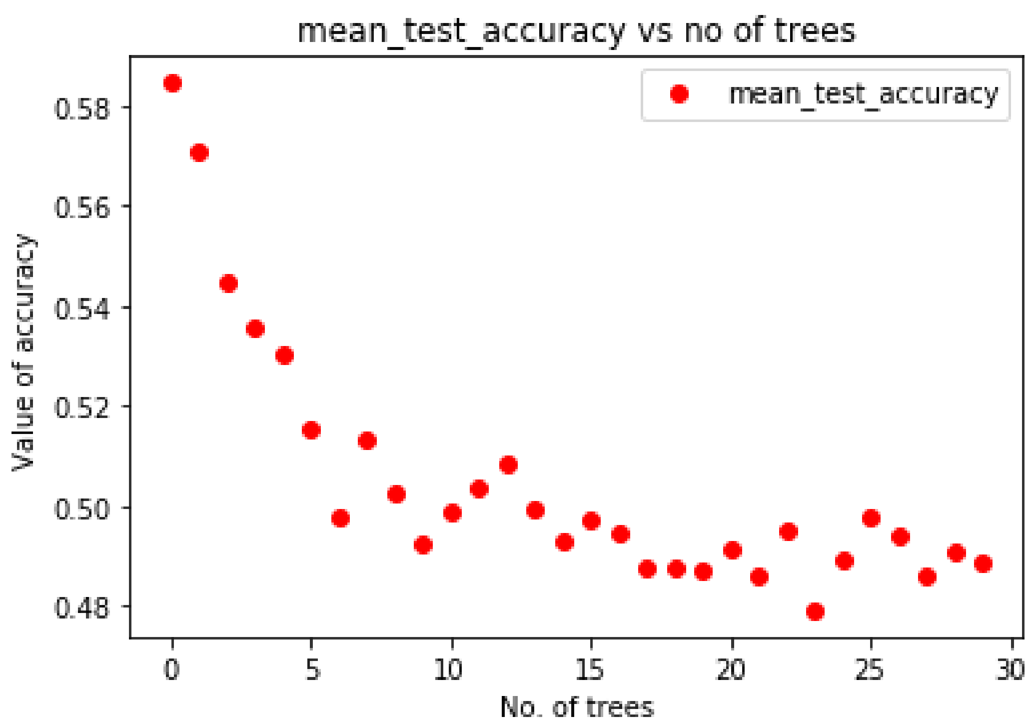
Question 3

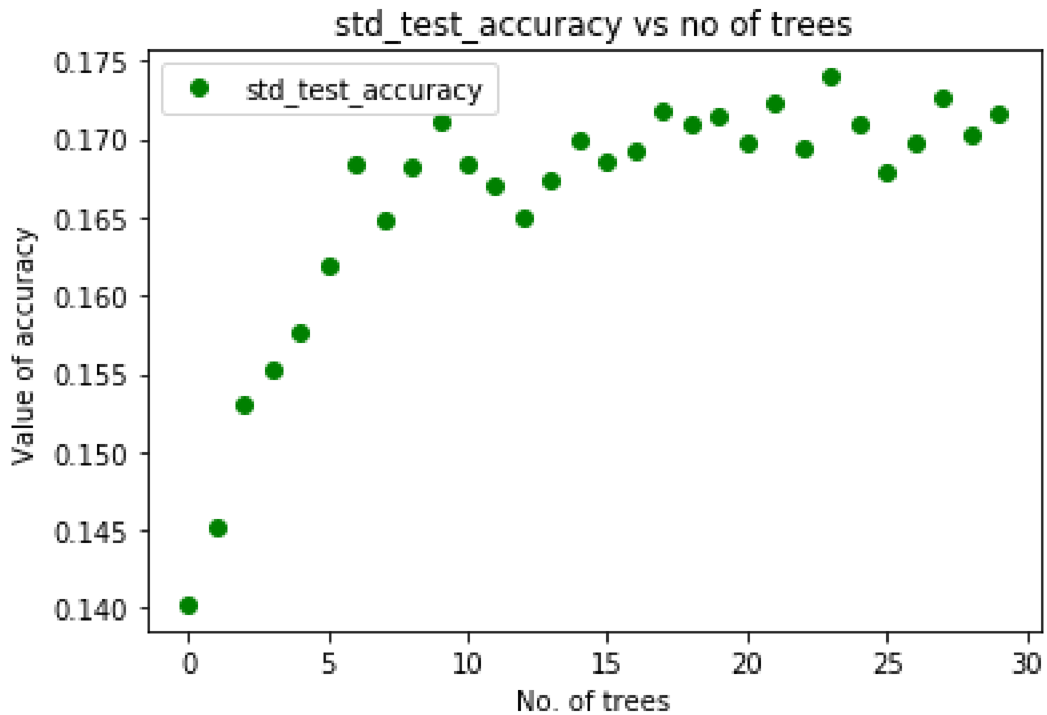
(a) I used Python for coding the problem.

(b)




mean_train_accuracy vs no of trees
mean_train_accuracy
Value of accuracy
No. of trees
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
0
5
10
15
20
25
30


mean_test_accuracy vs no of trees
mean_test_accuracy
Value of accuracy
No. of trees
0.58
0.56
0.54
0.52
0.50
0.48
0
5
10
15
20
25
30



As the number of trees increases, the mean training error rate decreases and becomes very close to 0.1 %. This decrease in mean error rate for the training data is much smoother than the one for test data. The mean error rate for test data also decreases, as the number of trees increases, but it.s value is much higher than the training data mean error and also there's a lot of fluctuations, i.e., it's not a smooth curve. Also, the initial decrease when we increase the number of trees from 1 to 7, is much higher than the later decrease when we increase the number of trees. The standard deviation of the test error increases for me but the increase is very miniscule from 0.14 to 0.175. This is higher than the mean train error but lower than the mean test error. It also fluctuates a lot and is not a smooth curve.

PROBLEM ON READING

Q1. (a) It is a case of semi supervised learning problem. Since we use only given training sample, its transductive.

(b) As we are out of original training data, and use new data, it is inductive.

(c) As all the data availaible is labeled, it is a type of supervised learning.

(d) It is a case of semi supervised learning problem. It's of inductive type.

Q2. CART applied to regression.

(a) $$\text{Cost}\,\{(\underline{x}_i, y_i) \in R_{m'}\} = \sum_{\underline{x}_i \in R_{m'}} (y_i - w_{m'})^2$$

When we differentiate wrt $w_{m'}$,

$$\frac{\partial\, \text{Cost}}{\partial\, w_{m'}} = \sum_{\underline{x}_i \in R_{m'}} -2(y_i - w_{m'}) = 0$$

$$\therefore \sum_{\underline{x}_i \in R_{m'}} -2y_i + 2w_{m'} = 0$$

$$\therefore \sum_{\underline{x}_i \in R_{m'}} w_{m'} - y_i = 0$$

$$\therefore \; N_{R_{m'}} \cdot w^*_{m'} - \sum_{x_i \in R_{m'}} y_i = 0$$

$$\boxed{\therefore \; w^*_{m'} = \sum_{x_i \in R_{m'}} \frac{y_i}{N_{R_{m'}}}}$$

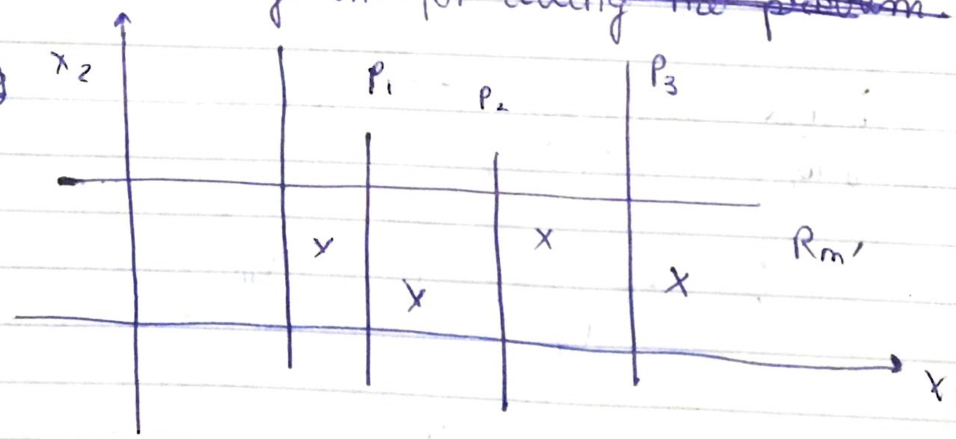
(b) For a given region $R_{m'}$, containing $N_{R_{m'}}$ data points, and a given feature threshold x_j, if we want to find an optimal value threshold t_k by trying different values, then we need to try

atmost $N_{R_{m'}} - 1$ values of t_k.

This is because, given a feature to threshold in $R_{m'}$, we can atmost get $N_{R_{m'}}$ different outputs splitting a region if the points are not aligned.

Q3.

~~(a) I used Python for coding the problem.~~



$(N_{R_{m'}} - 1)$ possible decision boundaries within $R_{m'}$.