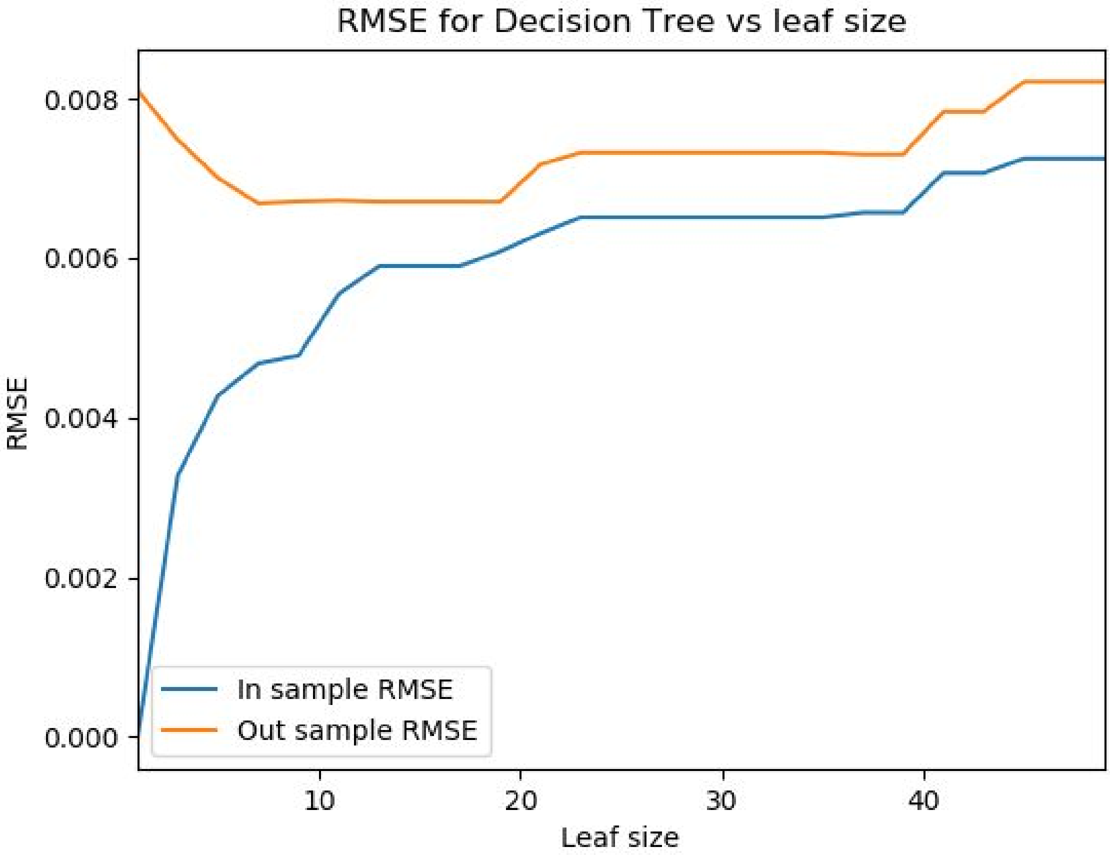
CS 7646 - Project 3 Report:

Abhijeet Gaurav
agaurav@gatech.edu

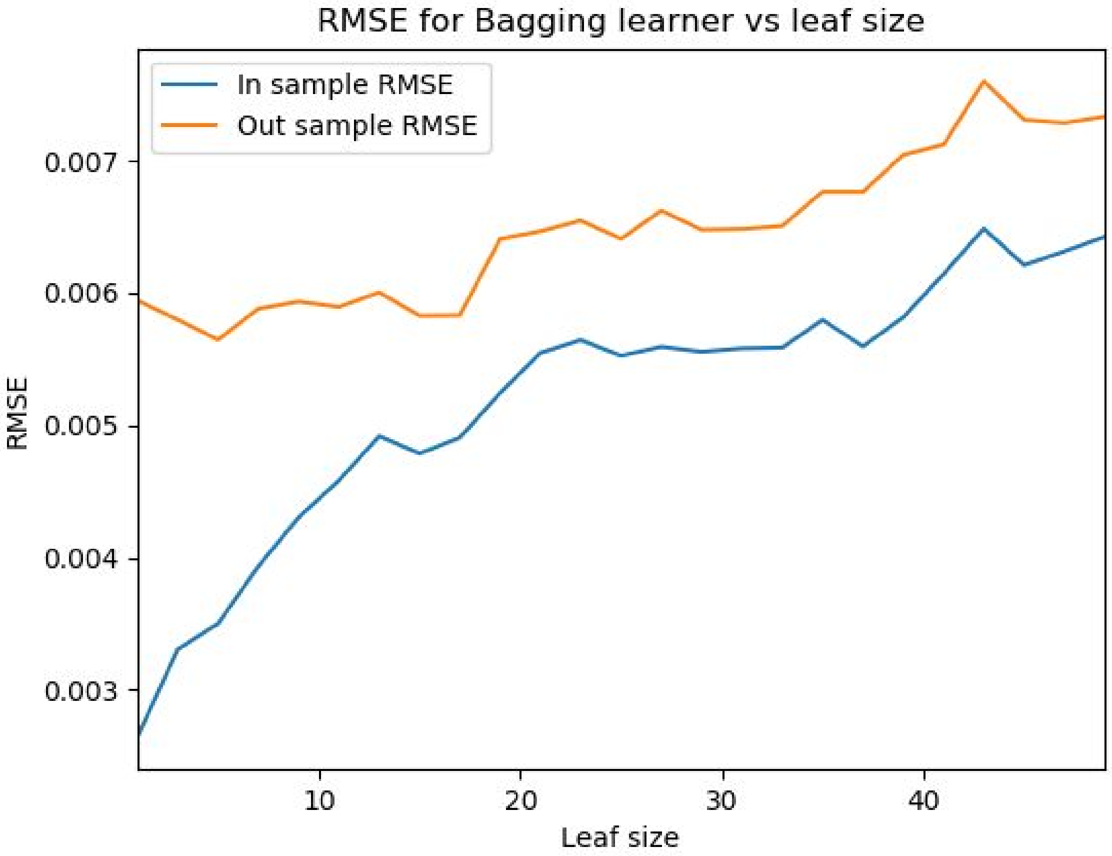
Question 1:



Experiment methodology: Istanbul.csv dataset was used to train and test the Decision tree model with leaf size from 1 to 50. The root-mean-square error (RMSE) is plotted for both the 'In sample' and 'Out sample' for clear comparison.

Clearly, from the above graph it can be seen that overfitting does occur with respect to the leaf size. The RMSE for 'out sample' decreases as we lower the value of leaf size (if you look in the negative x-axis direction). On the far right, at large leaf size = 50; the RSME is large, and as we decrease leaf size, both the 'in sample' and 'out sample' error are decreasing till leaf size = 8. At this point (leaf size = 8), the out sample RMSE starts increasing again but the in sample RSME keeps decreasing proving that beyond this point we have an overfitting case.

Question 2:

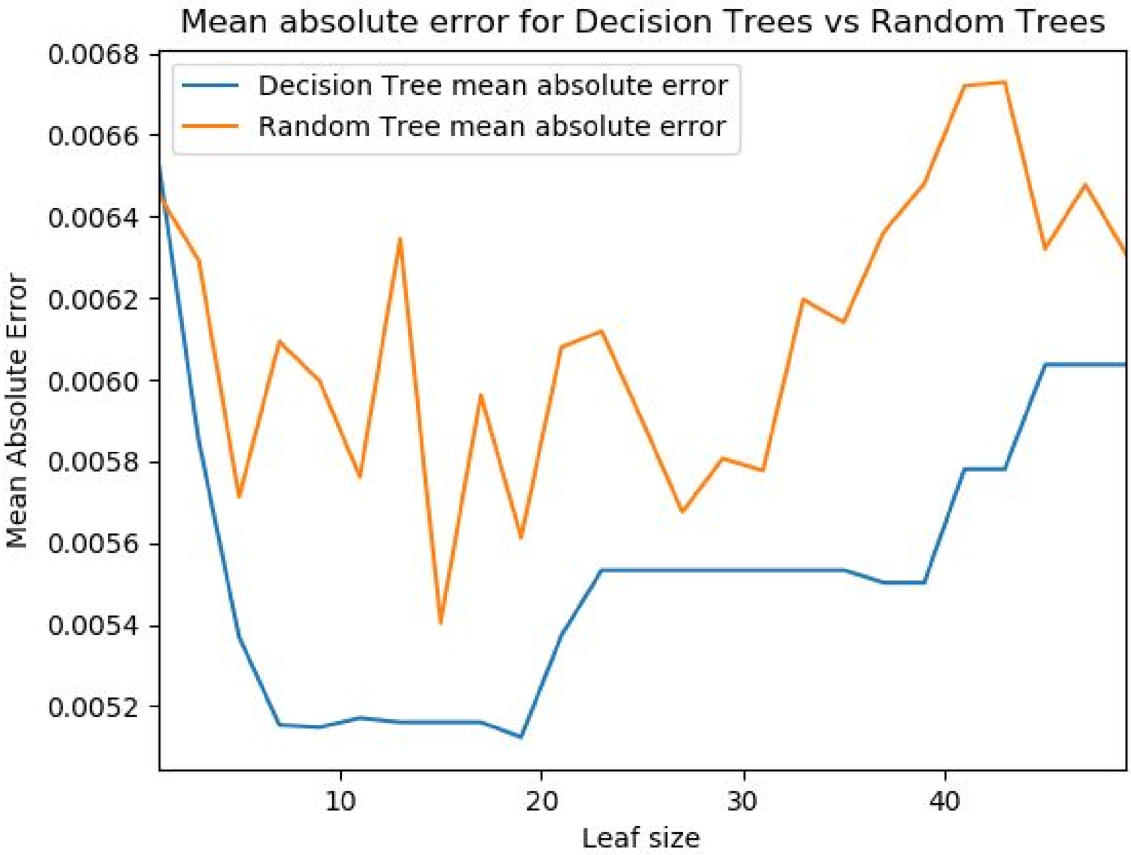


Experiment methodology: With a bag size of 10; Istanbul.csv dataset was used to train and test the Decision tree model with leaf size from 1 to 50. The root-mean-square error (RMSE) is plotted for both the 'In sample' and 'Out sample' for clear comparison.

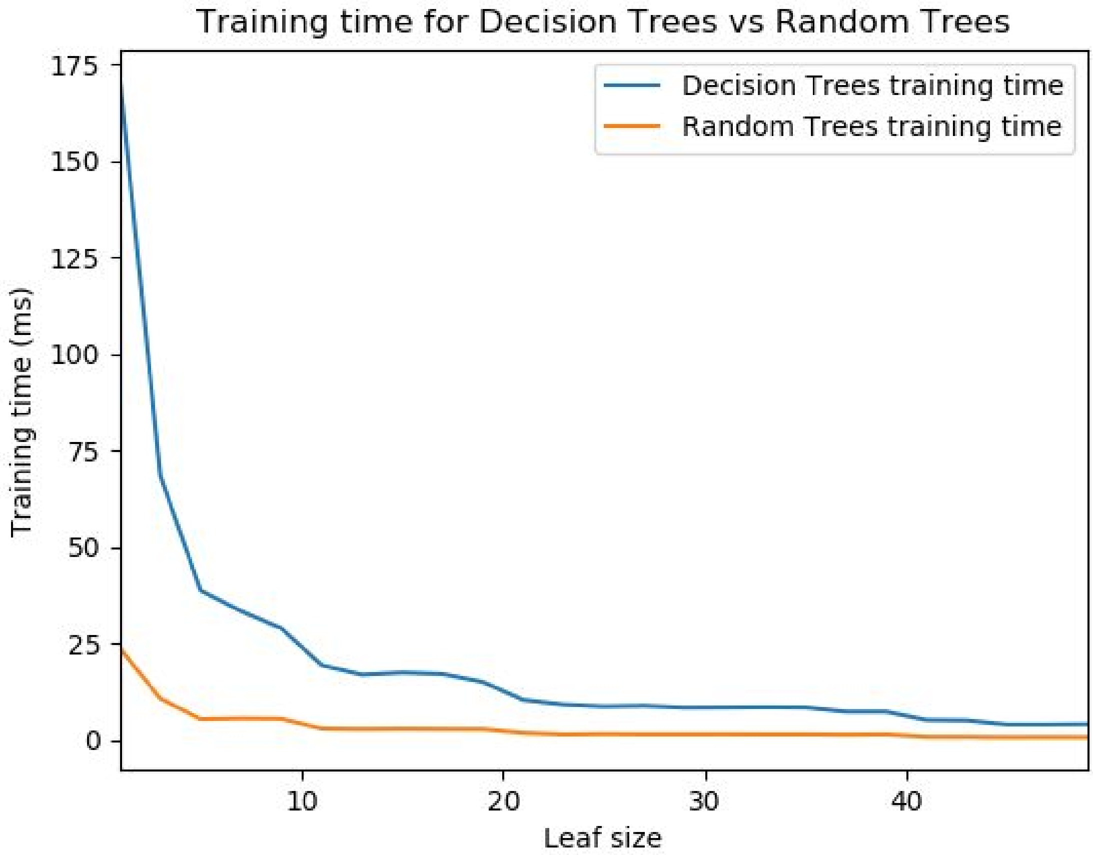
Clearly, from the above graph it can be seen that bagging can be helpful in decreasing the overfitting. In the above example with bagging the RMSE didn't increase for 'Out sample' as sharply as in Question 1's case where the RMSE increased as we decreased leaf size below 8. Infact, at leaf size < 8, the RSME is quite stable. This proves that Bagging helped in avoiding overfitting. Also, the RSME is consistently lower for bagging compared to without bagging.

Question 3:

Experiment methodology: Istanbul.csv dataset was used to train and test the Decision tree model and Random Tree model with leaf size from 1 to 50. The mean absolute error and training time for each learner is compared.



The above figure indicates that the Decision tree model gives a better accuracy than the Random tree model. The mean absolute error for the decision tree is consistently lower than the Random tree model (excluding leaf size = 1).



When compared on the training time, it can be clearly seen that random trees are consistently better than decision trees. As the leaf size decreases, there is much faster increase in training time for the decision tree as compared to the random tree. This can be explained on the basis of the fact that in the decision tree, we calculated correlation value every time for all the points in the dataset but on the other hand, for the random tree we decided to just do a random split.