R·I·T	**Rochester Institute of Technology** **Golisano College of Computing and Information Sciences** **Department of Information Sciences and Technology**

Lab 5 (3 points)
Tree-based Models

This lab consists of two parts, which use R to build and test-based models.

Part I

This problem involves the OJ data set, which is part of the ISLR package.

1. Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
2. Fit a tree to the training data, with Purchase as the response and the other variables except for Buy as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?
3. Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.
4. Create a plot of the tree, and interpret the results.
5. Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
6. Apply the cv.tree() function to the training set in order to determine the optimal tree size.
7. Produce a plot with tree size on the **x** -axis and cross-validated classification error rate on the **y** -axis.
8. Which tree size corresponds to the lowest cross-validated classification error rate?
9. Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.
10. Compare the training error rates between the pruned and un-pruned trees. Which is higher?
11. Compare the test error rates between the pruned and un-pruned trees. Which is higher?

Part II

We will use Carseats data and seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

1. Split the data set into a training set and a test set.
2. Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?
3. Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

4. Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important.
5. Use random forests to analyze this data. What test MSE do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.