**Assignment: Tree Based Methods, Regression Trees**

BUAD 5082 – Spring 2019

1. **Objectives**

The purpose of this assignment is to provide you with an opportunity to investigate some of the ideas regarding tree based methods and how to create regression trees.

1. **What You Will Need**

* Access to a Windows computer with R, and to the following files, which can be downloaded from the Class Schedule page of the course web site:

1. **What You Will Hand In**

Nothing, it is not expected for you to submit this assignment.

1. **Due Date**

March 10th, 2019: just before midnight.(Have a wonderful spring break but do not forget to work!) After this date we will post the solutions to the problem set.

1. **Note on Collaboration**

This is not an official assignment. You may work with others solve this problem set.

1. **Preliminaries:**

To get set up for the assignment, follow these steps:

1. As the first statement in your script file, enter rm(list=ls())
2. Each question and part in the assignment should begin with the following three comment lines where n is the question number and p is the part letter:

###########################

####QUESTION n PART p####

###########################

1. Do not create console output other than what is asked of you explicitly.
2. We suggest that you read the entire assignment before starting – there are often notes and suggestions at the end of assignment documents.
3. **Assignment Tasks:**

**Question 1: Predicting Myweather data: Tree Based Methods using Classification Trees**

1. Install needed packages (rpart, rattle, pROC, verification)
2. Read the dataset and discard unwanted variables ("Date", "Location","RISK\_MM")
3. Build 90% training vector of indices and 10% test data frame (do not forget to set seed)
4. Build rpart decision tree. Use “class” as method and “information” as split.
5. Print the model information as well as cptable of the model (cp stands for complexity parameter. use *printcp* )
6. Plot the classification tree
7. Use fancyRpartPlot function to plot a better looking tree
8. Building a maximal model. Note that now the model is overfitting
9. Pruning the maximal tree to minimize xerror(misclassification error)
10. Extract the cp value for the minimal xerror
11. Specifying a Loss Matrix with loss argument in the control parameter within the rpart function with an 0s on true positive and true negative.
12. Use your model to predict on the test set

bonus question:

Build a ROC curve

**Question 2: To Gain More Understanding about Regression Classification**

1. Understand how to grow a decision tree
2. Why do we use cross-entropy and gini index?
3. What are the equations of cross-entropy and gini index?
4. What are the differences of cross-entropy and gini index?
5. Why do we need to prune a tree and how exactly do we prune?
6. While calculating the expected entropy, why do we need to include the weight parameter?