**IE 425**

**Homework 1 (due March 26 23:59pm)**

1. Consider the dataset called “heart.csv”. There are 14 attributes that are described below, where the last one is the output attribute.

age: age in years

sex: (1 = male; 0 = female)

cp: chest pain type

trestbps: testing blood pressure (in mm Hg on admission to the hospital)

chol: serum cholestoral in mg/dl

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg: resting electrocardiographic results

thalach: maximum heart rate achieved

exang: exercise induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment

ca: number of major vessels (0-3) colored by flourosopy

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

target: 1 or 0

a. Partition the dataset heart.csv into training and test sets with 70% going into the training set by using CaTools package and a seed value of 425.

b. By playing with the parameter values of the “tree” function of the “tree” package, find the best tree in terms of the performance on the test set. Give the number of leaf nodes of the best tree and provide the summary of the tree using the summary(tree\_name) function with its corresponding performance (accuracy, error rate) on the training as well as test set.

c. Provide the Confusion Matrix along with sensitivity, specificity, precision and recall. Comment on which input attributes are important in making predictions.

d. Repeat (a)-(c) with rpart package.

2. Consider the dataset called “suicide-rates.csv”. There are 11 attributes that are given below, where the last one is the output attribute (suicide rate).

country

year

sex

age

suicides\_no

population

country-year

gdp\_for\_year ($)

gdp\_per\_capita ($)

generation

suicides/100k pop

Our goal is to predict the suicide attribute.

a. Partition the data set into training and test sets with 70% going into the training set by using a seed value of 492.

b. By playing with the parameter values of the “tree” function of the “tree” package and using the training set, find the best tree in terms of the performance on the test set given in RMSE. Give the number of leaf nodes of the best tree and provide the summary of the tree using the summary(tree\_name) function with its corresponding performance (RMSE) on the training as well as test set.

c. Comment on which input attributes are important in making predictions.

d. Perform 10-fold cross validation using the whole dataset by cv.tree and rpart functions. What is the best size of the trees given by these two functions?