HW6 (Due April 1st)

You may work in groups of two. Both names must appear on the turned-in assigment. Note that you may not work with someone who was in your group previously in this class.

1) Consider the cps_income dataset found in the 0.Data folder, which contains data from "current population survey". Look through the Description document to get a sense for the variables in the dataset, what they represent, and the values those variables can take on. Note that the goal is to predict whether an individual makes MORE or LESS than 50k per year.

   DOWNLOAD THE DATASET, IMPORT INTO Rstudio AND RUN THIS CODE FIRST:
   YOUR_DATA_NAME = data.frame(YOUR_DATA_NAME)
   THEN IDENTIFY AND USE TWO QUALITATIVE and THREE QUANTITATIVE (LABELED-CONTINUOUS) VARIABLES.

Note: You must include code and any output.

(a) State the variables you picked and create a decision tree using rpart() using the variables you picked. **Note:** Also include a plot of the decision tree here.

(b) Try to obtain a better (resulting in smaller prediction error) control parameter than the one in (a).

(c) Use the tree in (b) to estimate the probability that the first individual in the dataset earns " >=50K".

(d) Use the tree in(b) to predict whether the individual in (c) earns more or less 50K.

(e) Randomly assign 30000 observations to a training set and the rest to a testing set. Build the decision tree (with default control parameter) using the training set and obtain the misclassification rate on the testing set.

(f) Obtain the misclassification rate for 10-fold cross validation when using the decision tree model (with default control parameter). Note you cannot use the cv.glm() function and will need to do this in a loop.

2) Now identify and use all 5 available quantitative predictor variables only. Using the same training and testing sets as in 1(e) for the K-nearest neighbors (with K=20) algorithm with Euclidean distance to predict the INCOME column on all the observations. **Don't forget** to standardize the columns first! You must include the code used and output generated.

3) Obtain an optimal value of K to use in K-nearest neighbors algorithm with Euclidean distance for this data after trying at least 50 possible values. Include a plot of the misclassification rates for different values of K.