# **BS806 --- Homework 11**

**Reading**. Chapter 4 of (ISLR) Introduction to statistical learning.

**Question 1.** The City of Somerville sends out a happiness survey to a random sample of Somerville residents asking them to rate their personal happiness and their satisfaction with City services every second year since 2011. Every year, the survey is refined. We will use data of year 2015. The dataset includes six features (X1 – X6), each of which take value 1 to 5, and one decision attribute (D) with values 0 (unhappy) and 1 (happy).

**Variable Information**:

D = decision attribute (D) with values 0 (unhappy) and 1 (happy)

Attributes X1 to X6 have values 1 to 5.

X1 = the availability of information about the city services

X2 = the cost of housing

X3 = the overall quality of public schools

X4 = your trust in the local police

X5 = the maintenance of streets and sidewalks

X6 = the availability of social community events

We will use this dataset, **Somerville\_Happiness.csv**, to perform classification analysis.

1. **[10 points]** Randomly select 100 observations for training and the remaining observation will serve as test data. Please use set.seed(2) to set the random number seed.
2. **[54 points]** Conduct the following classification analysis and then identify the best classifier based on area under the curve (AUC)
   1. Logistic regression (12 points)
   2. Classification Tree (12 points)
   3. Linear Discriminant Analysis (10 points)
   4. Quadratic Discriminant Analysis (10 points)
   5. Compile a table to compare the AUC for these approaches and write a report which method you would recommend based on your analysis (10 points)
3. **[36 points]** Perform the classification analysis using the whole dataset (i.e. without splitting as training and test) by implementing the approach you recommended in Question 2 with default threshold or majority vote (i.e. probability of 0.5 as threshold). Then
   1. Construct a confusion matrix. A confusion matrix is a table describing the performance of a classification model (or “classifier”), i.e. a 2x2 Table in this example. (8 points)
   2. What’s misclassification rate and accuracy? (8 points)
   3. What’s your sensitivity and specificity? (8 points)
   4. What’s your positive predicted value and negative predicted value? (8 points)