Week 4 Assignment – 50 Points

# Regression Section

1. Manual work. Do not use R for this exercise  
   Build a regression model to classify the Beverage Sales (10)
   1. Open “Beverage Sales.xlsx” .
   2. Identify target variable.
   3. Plot the Temperature vs Sales. Use the excel tool to compute trendline.
   4. Use the manual method you have learned in class and build the regression model. Save the solution excel file as “Beverage Sales Manual Solution.xlsx”.
2. R. Work. No manual work accepted

Build a regression model to classify the Beverage Sales (10)

* 1. Open “Beverage Sales.xlsx” .
  2. Use the required R code to build the linear regression model.
  3. Compare the XL trendline, your manual model, and R generated model.
  4. Submit your R codes with the R code of the Assignment 4. Make sure that I can identify the R code belonging to this exercise!

# KNN Classification Section

1. Manual work and R work (20 Points)

Classify the employee’s employer (MLB’s Yes or No).

Manual

* 1. Open “KNN Employee Salaries.xlsx” file. As you see the historical data is on the sheet “Data” and New Data” has three new observation.
  2. To apply the KNN algorithm you may copy and paste new data on the sheet containing historical data.
  3. Apply, manually KNN algorithm on data and new data to classify new data observations. ignore irrelevant column if any.
  4. Save the XL file as “Solution KNN Employee Salaries.xlsx”. Submit the solution.

R work

1. Use R to classify the new observations in the “KNN Employee Salaries.xlsx” file.

The following code reads a specific sheet from an XL file.

Library(readxl)

emp\_salary.df <- read\_excel("KNN Employee Salaries.xlsx", sheet = "Data")

New.df <- read\_excel("KNN Employee Salaries.xlsx", sheet = "New Data")

1. Remove irrelevant columns if any.
2. Apply the algorithm (modeless KNN code) on new data. Display the result
3. Submit your R codes with the R code of the Assignment 4. Make sure that I can identify the R code belonging to this exercise!
4. Use R to implement the K-NN on the data in the Boston-Housing-classification.csv

(20 Points)

* Explore data
  + Remove columns 2, 4, 9, and 12
  + Create boxplot for each predictor
  + Create histogram for each predictor
  + Interpret predictors
  + Create a boxplot of Score vs Decision
  + Create a boxplot of Attendance vs Decision.
  + Interpret the visualizations
* Data Mining
  + Standardize both attributes (code for standardizing more than one variable is posted along the assignment)
  + Partition data into training 60%, evaluation 20%, and test 20%
  + Apply K-NN k=3 (model 1), and k=5 (model 2). Make sure to use the right partition for evaluating the model and assess it.

Which model is better? What is the model accuracy, sensitivity, specifity, and F1 for evaluation and test data.

* Apply the selected model
  + Apply the model of your choice to the new students in the “BH\_classification\_New.xlsx” file.
* Submit a XL file with the following on each working sheets
  + Original Dataset
  + Dataset normalized
  + Training dataset
  + Validation Dataset with prediction result as a new column
  + Test Dataset with prediction result as a new column
  + Confusion matrix of validation and result
  + New Data with prediction

Submit R codes of all exercise of this assignment in one text file. Make sure that I can identify the R code belonging to each exercise!

Do not submit any screenshot!