**Instructions**

- This homework assignment is worth 66 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: September 15, 2023 by 11:59 pm.**

# Exercise 1

(4 points) What type of algorithm would you use to segment a company customers database into multiple groups?

# Exercise 2

(4 points) Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem? Explain.

# Exercise 3

(4 points) What is a test set, and why would you want to use it?

# Exercise 4

(4 points) You are training a classification model with 100 variables/features that achieves 95% accuracy in the training dataset. However, when you run it in the test dataset, you only get 60% accuracy. Which of the following options are valid approached to solve this problem?

(a) Reduce the number of input variables/features

(b) Add extra variables/features

(c) Implement cross-validation during the training process

(d) Select another algorithm

(e) (a) and (c)

(f) (a) and (d)

(g) (b) and (c)

(h) (b) and (d)

(i) None of the above

# Exercise 5

Consider the `Real_Estate.csv` data file posted on Blackboard (under the In-Class 3f assignment link). This file contains information related to 414 houses. The goal is to predict the house price of unit area. In **Python**, answer the following:

(a) (3 points) Using the pandas library, read the csv data file and create a data-frame called `house_price`.

(b) (3 points) Drop the `No` and `transaction_date` columns.

(c) (4 points) Split the data-frame into two data-frames: train (80%) and test (20%).

(d) (10 points) Using the train data-frame, build a linear regression model in which: `house_age`, `distance_to_the_nearest_MRT_ station`, `number_of_convenience_stores`, `latitude`, and `longitude` are the input variables, and `house_price_of_unit_area` is the target variable. After the model is built, predict the house price of unit area of a the houses in the test data-frame. Report the MSE.

(e) (10 points) Using the train data-frame, build a linear regression model in which: `house_age`, `number_of_convenience_stores`, `latitude`, and `longitude` are the input variables, and `house_price_of_unit_area` is the target variable. After the model is built, predict the house price of unit area of a the houses in the test data-frame. Report the MSE.

(f) (5 points) Using the results from parts (d) and (e), what model would use? Explain.

(g) (15 points) Repeat steps (c) to (e) 100 times, and visualize the MSE of each of the models at each iteration. Which of the two model has better performance on the test datasets? Explain.