COMP6930 - Assignment #1
(Due Date: 6th March 2018)

Instructions: You are to submit a **.zip** folder named using your UWI ID number to inzamam.rahaman@outlook.com. This folder should contain two subfolders - one per part named accordingly. Moreover, at the top of each file, you should write (using comments) your UWI ID number. Note that the datasets to be used for both parts A and B are called data_a.csv and data_b.csv respectively.

Part A - Linear Regression (20 marks)

The price of a diamond is dependent on several of its qualities. Some of these qualities include:

- carat: weight of the diamond (0.2–5.01)
- cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
- color: diamond colour, from J (worst) to D (best)
- clarity: a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
- x: length in mm (0–10.74)
- y: width in mm (0–58.9)
- z: depth in mm (0–31.8)
- depth: total depth percentage $= z/mean(x, y) = 2 * z/(x + y)$ (43–79)
- table: width of top of diamond relative to widest point (43–95)

In data_a.csv, you are given a table with diamonds with these qualities and their corresponding price. Design a linear regression model that accepts the above qualities as input, with appropriate selection, encoding, and feature engineering, and measure the performance of your model using 5-fold cross validation using Mean Average Percentage Error. Note that Mean Average Percentage Error is not provided by sklearn and you will have to code it yourself.

Part B - Logistic Regression (20 marks)

Medical diagnosis and prognosis is one of the most important applications of machine learning and data mining. Given measurements of a breast cancer tumor, we will like to be able to predict the likelihood of recurrence of breast cancer in a patient. You are given a dataset of measurements of breast cancer tumors in data_b.csv along with information of whether the patient experience recurrence or not. The data is a cleaned version of the noted Wisconsin Breast Cancer dataset. This data contains several inputs for a model (such as radius_mean, texture_mean,perimeter_mean, compactness_worse, etc...) and your target in the diagnosis field. You are to design a model using logistic regression and :

- Use 5-fold cross validation to compute F1 scores of your model. Report both the mean F1 score and the std. in F1 scores over all folds.
- Comment on the most important features of the model. Briefly explain your reasoning on why those features were the most important. You may either analyze your model directly, or use a library such as Skater