Assignment II

Data Science and Machine Learning

INSTRUCTOR: *William Klement*

CLASS: *AML 1113*

- *Individual task*
- *Due date: Wednesday November 27, 2024 10:00AM*
- *Total marks100: please submit:*
  - *written report answers (including plots)*
    Python code associated with the questions

Instructions:

- For each question, please write a brief report to present your findings
- Present statistical summary and description learned in class
- Support your answers with appropriate data visualization plots especially for each of the demographic features for illustration.
- Use statistical methods and data visualization techniques (implemented in Matplotlib and Seaborn Python Libraries) you learned in class

© LAMBTON COLLEGE

Question 0 [0 marks]:

Download the **Bank Marketing** (https://archive.ics.uci.edu/dataset/222/bank+marketing) dataset from the UC Irvine Machine Learning repository (https://archive.ics.uci.edu/) to use in the questions of this assignment.

Question 1 [30 marks in total]:

Describe the distributions of demographic features for clients (subjects) in the given dataset. The demographic features available in this data set are: Age, Occupation, Marital Status and Education Level. For these features, present your analysis in two settings:

A. [20 marks] as univariate: one feature at a time,
B. [10 marks] as multivariate (all four features together

Hints:

- For this question, use the entire dataset!
- To make your comparison easier for question 4, calculate the mean, median, mode and standard deviation for each demographic feature.

Question 2 [20 marks]:

Repeat the same analysis in **question 1A** but group the clients based on the outcome variable of "***Bank Term Deposit***" being **Yes** or **No** (i.e., the categorical feature "y" in the data set). Please note, this question deals with univariate analysis plus outcome, i.e., each of the demographic features grouped by outcome. For this question, use the entire dataset!

Question 3 [20 marks]:

Repeat the same analysis in **question 2** and group the clients based on the outcome variable of "***Bank Term Deposit***" being **Yes** or **No** (i.e., the categorical feature "y" in the data set). In this question, use two of the four demographic features of your choice plus the outcome, i.e., select two demographic features grouped by outcome. For this question, also use the entire dataset!

Question 4 [20 marks]:

Using stratified sampling implemented in the Pandas library we covered in class, extract a **30%** sample from the dataset stratified over the outcome variable y ("***Bank Term Deposit***") you downloaded above and repeat the analysis from **question 1A** on it. The idea is to compare the results of your analysis form each of the four demographic features using mean, median, mode and standard deviation as metrics supported by your plots and graphs.

Hints:

- For each demographic feature, compare the values of mean, median, standard deviation, and mode calculated on the 30% stratified sample to those calculated form the original dataset above.
- OPTIONAL!!! These answers get much better if grouped by outcome (YES or NO)
- You will need to sample using Pandas (https://www.geeksforgeeks.org/stratified-sampling-in-pandas/)

Question 5 [10 marks]:

In your opinion, what percentage of data sample will give you a reasonably good representative sample of the original dataset? Briefly and in a short report, explain and possibly support your conclusions with evidence (statistical metrics and plots) from your analysis above.