MEME19803 Group Assignment

Course Code & Course Title: MEME19803 Programming for Data Analytics
Course: Master Programme Department: DMAS

Instructions

1. This is a group assignment with **two** to **three** students including a **group leader** per group.
2. **Group leader** need to submit the following items to `liewhh@utar.edu.my`:
   - a list of members (with signatures)
   - the dataset of interest from the given list
   - a report with proper references .....Thursday of Week 4
3. **Deadline of submission** for **group assignment report** is 5.00pm, 10 Feb 2022 (Thursday of Week 4).
4. In the case of **late submission** for the report and program script, 10% of the maximum marks will be deducted if the work is up to one day late (24 hours) and additional 10% of the maximum marks for each of the subsequent days.
5. **Plagiarism is not allowed**. If the works are found to be plagiarised, no marks will be given and the incident will be reported to the university for further action.
6. The group assignment report **can** contain the **measurement of contribution of each member to the project** in ratio or percentage to prevent any member from not contributing. The penalty will be as follows:
   - If in a two-member group or three-member group, member A is doing all the work while the rest is not doing anything, the member A will put 100% for member A and 0% for member B (and C). Suppose the report scores 17% out of 20%, then the mark distribution will be as follows:
     - member A gets 17%
     - the rest gets 17% $\times$ (1-0.4 $\times \frac{100-0}{100}$) = 10.2%
   - If in a three-member group, member A and member B are doing 40% of the work and member C is doing 30% and the report scores 16% out of 20%, then the mark distribution will be as follows:
     - member A gets 16%
     - member B gets 16%
     - member C gets 16% $\times$ (1-0.4 $\times \frac{40-30}{40}$) = 14.4%

Group Assignment Report (20%)

1. For a **two-member group**, you need to choose three datasets — a group A dataset and two group B dataset for your study.
2. For a **three-member group**, you need to choose five datasets — a group A dataset and four group B dataset for your study.
3. The group A dataset list:
    - `https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection`
    - `https://archive.ics.uci.edu/ml/datasets/Website+Phishing`
    - Challenging: `https://www.cia.gov/the-world-factbook/about/archives/`
    - Challenging: `https://github.com/MoH-Malaysia/covid19-public` (COVID-19 Malaysia Data)
4. The group B dataset list:
    - `https://archive.ics.uci.edu/ml/datasets/Abalone`
    - `https://archive.ics.uci.edu/ml/datasets/Arrhythmia`
    - `https://archive.ics.uci.edu/ml/datasets/Automobile`
    - `https://archive.ics.uci.edu/ml/datasets/Covertype`
    - `https://archive.ics.uci.edu/ml/datasets/Dermatology`
    - `https://archive.ics.uci.edu/ml/datasets/Echocardiogram`
    - `https://archive.ics.uci.edu/ml/datasets/Flags`
    - `https://archive.ics.uci.edu/ml/datasets/Glass+Identification`
    - `https://archive.ics.uci.edu/ml/datasets/Hayes-Roth`
    - `https://archive.ics.uci.edu/ml/datasets/Labor+Relations`
    - `https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient`
    - `https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset`
    - `https://archive.ics.uci.edu/ml/datasets/Steel+Industry+Energy+Consumption+Dataset`

The report should contain the following items:

- Background analysis of features and targets in each dataset with proper references.
- Using Python to read the data array / table and the proper data types are checked to make sure the data have been properly read (which is the first step in a data science pipeline).
- Using Python Numpy array functions and Scipy functions to summarise the statistics (min, max, mean, median, standard deviation, etc.) of each **numeric features** in the dataset and then identify the possible distribution of the data and occationally the outliers (which is the second step in a data science pipeline).
- Explain the sort of business that each dataset is associated with and the sort of pipeline(s) that may be relevant to the dataset.
- Optional: Efforts, distribution of tasks and collaboration in a group project to achieve the goal.