MEME19803 Group Assignment

Course: Programming for Data Analytics (MEME19803/MECG11503/MCCG11503)
Programme: MM, MC 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` / Liew How Hui @ MS Teams Chat:
   - a list of members (with signatures)
   - the dataset of interest from the given list
   - a group report with proper references ..... Wednesday of Week 4
3. Every member need to submit an individual report based on the group report.
4. **Deadline of submission** for **group assignment report** is 5.00pm, 8 Feb 2023 (Wednesday of Week 4).
5. In the case of **late submission** for the report and program script, 10% of the maximum marks may 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.
6. **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.
7. Each member will need to submit a hand-written individual report (scanned and save in PDF format) to `liewhh@utar.edu.my` / Liew How Hui @ MS Teams Chat.

Group Assignment (20%)

1. For a **two-member group**, you need to choose two datasets for your case studies.
2. For a **three-member group**, you need to choose three datasets for your case studies.
3. Datasets for the case studies
    - `https://archive.ics.uci.edu/ml/datasets/Website+Phishing`
    - `https://archive.ics.uci.edu/ml/datasets/AAAI+2013+Accepted+Papers`
    - `https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+%28EPM%29%3A+A+Learning+Analytics+Data+Set`
    - `https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping`
    - `https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression`
    - `https://www.cia.gov/the-world-factbook/about/archives/` (You only need to analyse the data for Malaysia. Challenging)
    - `https://github.com/ricardovvargas/3w_dataset`
    - `https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones` (too many features. Challenging)

The **group report (17%)** should contain the following items:

- Background analysis of features (for text data, simple text analysis should be performed) in each dataset with proper references.
- Using Python to read the original data (from the given URL, data from other sites will not be accepted) and convert them to 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 occasionally 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.

The **individual report (3%)** should contain the following items:

- A **summary** (in own words) and a **comment** on the group report.
- Suggestions for the data (other than those listed in this assignment) you are interested in and how can the analysis of the suggested data will be helpful in your career.