MEME19903/MECG11103 Assignment

Course Code & Course Title: MEME19903/MECG11103 Predictive Modelling
Course: MM, MDMA Department: DMAS

Instructions

1. In this assignment, a report and an R script needs to be submitted. The total marks are 20%. The breakdown of marks are: the report (5+7=12 marks) and the R script (8 marks).
2. The **deadline of the submission** is **6:30pm Monday Week 12** (8th August 2022). Both the assignment report (in PDF or in Word document format) and the R script (readable by the base R) can be submitted through email (`liewhh@utar.edu.my`) or MS Teams Chat.
3. In the case of **late submission** for the report and program script, 10% of the 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.
4. **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.

Assignment Report (12%)

1. Pick a dataset from the following list and perform **unsupervised** and **supervised** learning on them:

    - Student Performance Data Set (`https://archive.ics.uci.edu/ml/datasets/Student+Performance`)

2. By using the R statistical software, use different statistical models to analyse the data. You should apply the statistical models for prediction introduced in the lecture. You can use the statistical models for prediction not covered in the lecture, but descriptions and documentations are required with proper citations.

3. The report should have an introduction, appropriate data description, appropriate exploratory data analysis, appropriate theories, the comprehensive use of predictive supervised learning models to the data and a table summarising the performance measurements.

Programming Code (8%)

1. Write a programming code (or nicely structured programming codes) with an appropriate use of libraries which analyse the **raw dataset** which is picked in the assignment report and works in a data science pipeline.

2. Marks will be **deducted** if the programming code is in notebook and/or markdown and/or non-text formats which are not directly executable.

3. Marks may be **deducted** if data processing taught in the practical are not used but the sophisticated techniques from the Internet are copied (such as dplyr, etc.) without proper documentation in the assignment report.

4. The programming code can only use free and legal statistical software such as R, RStudio or MCRAN. The code should have reasonable dependencies and is cross-platform, i.e. the program can run on Microsoft Windows, GNU/Linux platform, MacOS/X, etc.

5. The programming code(s) need to demonstrate the appropriate use of **supervised** and **unsupervised** learning with proper comments.