# MEME19903/MECG11103/MCCG11103
# Assignment

Course Name: Predictive Modelling
Course: MAC, DMC, DAC Department: DMAS

## Instructions

1. In this assignment, a team with 2 to 4 members will be formed to write an R script and a report with predictive models at least the number of members. The assignment will contribute to the total marks of 20% in the assessment. The breakdown of marks are: the R script (5 marks, CO2: Compare statistical models through supervised learning for prediction and estimation), the report (15 marks, CO4: Demonstrate results from optimised supervised and unsupervised learning models.)
2. The team leader is responsible for combining all contributions from the team members and write down the contributions of each member in quantitative measurements. Any member who contributes nothing to the assignment will only receive 60% of the group assignment marks.
3. The **deadline of the submission** is **9:30pm Monday Week 12** (7th August 2023). Both the R script (readable by the base R) and the assignment report (in PDF or in Word document format) can be submitted through email (`liewhh@utar.edu.my`) or MS Teams Chat by the team leader.
4. 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.
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.

# Part 1: Programming Code (5%)

1. Pick **one** dataset from the following list of case studies and perform **unsupervised** and **supervised** learning on it:

   - Case study in identifying spam and scam messages with the Spambase dataset (`https://archive.ics.uci.edu/dataset/94/spambase`)
   - Case study in health science with the Chronic Kidney Disease dataset (`https://www.kaggle.com/datasets/mansoordaku/ckdisease`)
   - Case study in customer service with Airline Passenger Satisfaction dataset (`https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction`)

2. 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.

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

4. 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.

5. 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.

6. The programming code(s) need to **compare** at least $n$ statistical models through **supervised learning** for prediction and estimation as well as unsupervised learning which will be included in the report. Here $n$ is the number of members in a team.

# Part 2: Group Assignment Report (15%)

1. By using the analysis resuls from the programming code. You should demonstrate

   - the results of unsupervised learning with appropriate feature analysis and feature-response analysis.
   - the results of supervised learning by comparing statistical learning models and obtain the optimal model.

2. The report should be written in a proper report format with the following components:

   - an introduction to the background of the data;
   - a chapter with appropriate unsupervised learning with feature analysis and feature-response analysis using methods from exploratory data analysis, dimensionally reduction and clustering (from the software);
   - a chapter applying the statistical models for supervised learning prediction introduced in the lecture. Statistical models for prediction not covered in the lecture can also be applied for comparison but descriptions and proper academic citations are required.
   - a conclusion

   Note that a good report should have appropriate theories and academic references for appropriate unsupervised learning and supervised learning as well as logical and appropriate presentations (using either tables or statistical diagrams) for the demonstration of results.