MECG15603/MCCG11103
Assignment 3

Course Name: Statistical Learning/Predictive Modelling
Programme: DMC, DAC

Instructions

1. In this assignment, a team with 2 to 3 members will be formed to write **an R script (10%)**, **a report (18%)** and to present your results in **an oral presentation (5%)** with supervised learning models at least the number of members. The following are the course learning outcomes (CLO) relevant to the assessment:
   - CLO2: Compare statistical models through supervised learning for prediction and estimation .... Report
   - CLO4: Demonstrate results from optimised supervised and unsupervised learning models. .... R script, Oral Presentation
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** (4th August 2025). Both the R script (readable by the notepad editor) and the assignment report (in PDF or in Word document format) can be submitted through email (`liewhh@utar.edu.my`) or MS Teams Chat (to Liew How Hui) 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 Lab (10%)

1. Pick **one** dataset from the following list of case studies and perform **unsupervised** and **supervised** learning on the selected dataset for a particular case study:

   - Differentiated Thyroid Cancer Recurrence Dataset (`https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence`). In this case study, you need to analyse the data and build a model which can predict the recurrence of thyroid cancer. This case study is relevant to Malaysia's ambition to continuously maintain a good healthcare system to fight against cancers as well as promoting its health tourism. The relevant Sustainable Development Goal for this case study is SDG 3 (Good Health and Well-being).
   - Higher Education Students Performance Evaluation Dataset (`https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation`) In this case study, you need to analyse the data and build a model which can infer the grade based on students' family background and study habits. It is hoped that good habits can be identified from students with good grades and the family background does not affect students' grade. The relevant Sustainable Development Goals for this case study are SDG 4 (Quality Education) and SDG 10 (Reduced Inequalities).
   - Spambase Dataset (`https://archive.ics.uci.edu/dataset/94/spambase`) The "spam" concept is diverse, it includes advertisements for products/web sites, make money fast schemes, chain letters, pornography. In this case study, you need to analyse the data and build a model which can predict whether a given email is spam or not. We can allow around 7% misclassification error but false positives (marking good mail as spam) are very undesirable. The relevant Sustainable Development Goal for this case study is SDG 8 (Decent Work and Economic Growth).

2. Write a programming code (or nicely structured programming codes) with the use of appropriate libraries which analyse the **raw dataset** which is picked for the assignment report.

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/rationale in the programming code.

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 perform feature analysis, feature-response analysis, pattern analysis (dimension and clustering) and then **compare** at least $\max\{n, 2\}$ supervised learning models for prediction and estimation. Here n is the number of members in a team. The results should be properly organised in the assignment report.

Part 2: Group Assignment Report (18%)

1. You should document the following analysis results derived from the programming lab in Part 1 into your assignment report.

    - the results of unsupervised learning with appropriate feature analysis, feature-response analysis and pattern discovery of the data dimension and possible clustering patterns.
    - the results of supervised learning by comparing at least $\max\{2, n\}$ number of supervised learning models for their performance. Here n is the number of members in an assignment group.

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

    - An introduction is comprehensive (documenting the background of the data) and has appropriate references and problem solving objective(s) for a case study;
    - A chapter with appropriate unsupervised learning with feature analysis and feature-response analysis using methods from exploratory data analysis, appropriate dimension reduction and clustering pattern analysis (based on programming lab results);
    - A chapter with the training of at least $\max\{2, n\}$ supervised learning models and their performance analysis. The trained models should be properly analysed;
    - 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.

Part 3: Presentation (5%)

1. Prepare presentation slides which summarises the group assignment report and the programming scripts.

2. An oral presentation which involves all members or selected member(s) are allowed.

3. The oral presentation should cover the following aspects:

   - A good introduction of the case study with relevant references as well as the sourcing of data, data loading, data transformation (for unstructured data), type checking (for structured data), cleaning and proper univariate analysis.
   - A good illustration of the discoveries of the statistical patterns (statistical distribution or clustering patterns) of each feature with respect to the response as well as the combinations of features with respect to the response with the use of unsupervised learning methods.
   - A good illustration of the trained supervised learning models. If the trained model meets the prediction requirement, explain how the model could be deployed and updated; OR

     If the trained model does not meet the prediction requirement, explain the possible reasons from statistical and/or mathematical perspective.
   - The presentation is concluded with a comparison table for comparing the performance of the trained supervised models and how the objectives have been met as well possible future research work.
   - The presentation is well-timed (less than 18 minutes but not too short, heavy marks may be deducted for the presentation which is over-time) and well-prepared.