UECM3993 Group Assignment

Course Code & Course Title: UECM3993 Predictive Modelling
Course: AM, AS, FM Department: DMAS

Instructions

1. This is a group assignment with **four** to **six** students including a **group leader** per group.
2. The **group leader** need to submit the following items through email (`liewhh@utar.edu.my`) or MS Teams Chat:
   - a list of members (with signatures)
   - group title/name (cannot be too bizarre or offensive)
   - the dataset of interest from the given list

   for documentation before the start of assignment (Week 4).
3. Towards the deadline, the **group leader** is responsible to submit the following documents for the group assignment through email (`liewhh@utar.edu.my`) or MS Teams Chat:
   (a) "Group Name" Report.pdf ..... Wednesday of Week 11
   (b) "Group Name" program code(s) ..... Wednesday of Week 11
4. **Deadline of submission** for **group assignment report** and **group programming code** is 4.00pm, 24 August 2022 (Wednesday of Week 11).
5. **Group Presentation** will be scheduled in the weeks 11 to 13, date and time to be announced. Each presentation is limited to a maximum of 25 minutes (4 groups per 2-hour lecture).
6. 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.
7. **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.
8. The group assignment report **should** contain the information on the **contributions of members to the project** in ratio or percentage.

Marks

- Marks will be equally distributed by default. If the group assignment report has a section on **individual contributions** (in the first page or second page or the appendix), each member will receive

$$\textbf{teamwork marks} \times (1 - 0.4 \times \frac{\textbf{max IC} - \textbf{IC}}{\textbf{max IC}})$$

  where **IC** = individual contribution. For example, (Note: the same contribution can be applied to programming code.)

  - A group with 7 members with contributions (30%, 30%, 30%, 2.5%, 2.5%, 2.5%, 2.5%) and the report is 10 out of 12
    - 3 members will get $10 \times (1 - 0.4 \times \frac{30-30}{30}) = 10$ marks
    - 4 members will get $10 \times (1 - 0.4 \times \frac{30-2.5}{30}) = 6.33$ marks
  - A group with 4 members with contributions (A:10%, B:20%, C:30%, D:40%) and the report is 10 out of 12:
    - member A gets $10 \times (1 - 0.4 \times \frac{40-40}{40}) = 10$ marks
    - member B gets $10 \times (1 - 0.4 \times \frac{40-30}{40}) = 9$ marks
    - member C gets $10 \times (1 - 0.4 \times \frac{40-20}{40}) = 8$ marks
    - member D gets $10 \times (1 - 0.4 \times \frac{40-20}{40}) = 7$ marks

  The rational for the mark adjustment is to prevent individual members from doing nothing in the group. Any member who does nothing will only receive 60% of the teamwork marks.
- Each member will receive equal marks for the group programming code well unless the **group leader** wants to have a different weights for the group members.
- Each member will receive equal marks for the group oral presentation with extra marks for members who present really well unless the **group leader** wants to have a different weights for the group members.
- A group leader can be **re-elected** if more than half of the members are not happy with the group leader one week before the submission of the assignment.

Group Assignment Report (18%)

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

   - Census Income Data Set (`https://archive.ics.uci.edu/ml/datasets/Census+Income`)
   - Internet Advertisements Data Set (`https://archive.ics.uci.edu/ml/datasets/Internet+Advertisements`)
   - SMS Spam Collection (`https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection`)
   - Student Performance Data Set (`https://archive.ics.uci.edu/ml/datasets/Student+Performance`)
   - Non Verbal Tourists Data Set (`https://archive.ics.uci.edu/ml/datasets/Non+verbal+tourists+data`)
   - Mushroom Data Set (`https://archive.ics.uci.edu/ml/datasets/mushroom`)
   - Shoulder Implant X-Ray Manufacturer Classification Data Set (`https://archive.ics.uci.edu/ml/datasets/Shoulder+Implant+X-Ray+Manufacturer+Classification`)
   - Statlog (Australian Credit Approval) Data Set (`https://archive.ics.uci.edu/ml/datasets/Statlog+%28Australian+Credit+Approval%29`)
   - Website Phishing Data Set (`https://archive.ics.uci.edu/ml/datasets/Website+Phishing`)

2. $\max\{0, 0.5 - 0.1n\}$ mark will be awarded to the data with are investigated by n groups (in order of registration).

3. By using appropriate statistical software framework (R, Python with Scikit-learn, WEKA, C++, etc.), build models with the different statistical learning approaches (both unsupervised and supervised learning methods) which are covered in this course (and those which are not covered, but descriptions and documentations are required for methods not introduced in lecture with good references), find the "best" model for the data set selected and make sure that the objectives are met.

Group Programming Code (10%)

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 group 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 software. The default is R (and Python). The group who try other free and legal open source software (such as Java, C++) which are cross-platform and does not have too much dependencies, i.e. the program can run on Microsoft Windows (of various versions), GNU/Linux platform, MacOS/X, etc., will receive extra marks.

5. The programming code(s) need to demonstrate the appropriate use of **supervised** and **unsupervised** learning with the free and legal statistical software tool.

Group Oral Presentation (10%)

1. Prepare presentation slides which summarises the group assignment report and possible future improvements.

2. An oral presentation which involves every member or a presentation by just one or few representative member(s) are allowed.

3. The oral presentation should cover the following aspects:

   - A good description of the problem and a systematic use of unsupervised and supervised learning methods to discover important information from the dataset.
   - A good illustration of results and conclusions.
   - **Explain the algorithm** behind the best model with respect to following aspects:
     - The mathematical/statistical idea behind best supervised learning model for the problem.
     - Explain how the model would be updated if new data comes in.
   - Well-timed and interesting presentation (heavy marks may be deducted for over-time presentation)