UECM3993 Group Assignment (2024.06)

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

Instructions

1. This is a group assignment with **four** to **seven** 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. The lecturer reserves the right to assign remainder students who are not part of any assignment group to an assignment group with less than **seven** members. Those who cannot form a group before Week 4 may be penalised.
4. 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
5. **Deadline of submission** for **group assignment report** and **group programming code** is 6pm, 28 August 2024 (Wednesday of Week 11).
6. **Group Presentation** will be scheduled in weeks 11 to 13, date and time to be announced. Each presentation is limited to a maximum of 18 minutes (5 groups per 2-hour lecture).
7. 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.
8. **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 (you may be suspended if you are found guilty).
9. The group assignment report **is recommended to** 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** (either 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 (20%, 20%, 20%, 20%, 4%, 3%, 3%) and the report is 13 out of 18
    - 4 members will get $13 \times (1 - 0.4 \times \frac{20-20}{20}) = 13$ marks
    - 2 members will get $13 \times (1 - 0.4 \times \frac{20-3}{20}) = 8.84$ marks
    - 1 member will get $13 \times (1 - 0.4 \times \frac{20-4}{20}) = 8.58$ marks
  - A group with 4 members with contributions (A:10%, B:20%, C:30%, D:40%) and the report is 15 out of 18:
    - member A gets $15 \times (1 - 0.4 \times \frac{40-40}{40}) = 15$ marks
    - member B gets $15 \times (1 - 0.4 \times \frac{40-30}{40}) = 13.5$ marks
    - member C gets $15 \times (1 - 0.4 \times \frac{40-20}{40}) = 12$ marks
    - member D gets $15 \times (1 - 0.4 \times \frac{40-10}{40}) = 10.5$ marks

  The rationale for the mark adjustment is to prevent individual members from doing nothing in the group. Any member who does absolutely 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 based on individual contributions.
- 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 a **case study** on the dataset by applying the **unsupervised learning** and **supervised learning** on the dataset:
   - National Poll on Healthy Aging (NPHA) Dataset (`https://archive.ics.uci.edu/dataset/936/national+poll+on+healthy+aging+(npha)`): In this case study, you will try to predict (and possibly infer) two different target variables, namely Mental Health and Trouble Sleeping.
   - Audit Dataset (`https://archive.ics.uci.edu/dataset/475/audit+data`): This is an interesting case study because you will need to perform data preparation to combine the two data in the zip file. Try to work with the zip file without unzipping it.
   - Census Income Data Set (`https://archive.ics.uci.edu/dataset/20/census+income`): This is a classic case study in which the income exceeding \$50K/yr is predicted based on census data.
   - Spam filtering with YouTube Spam Collection Data Set (`https://archive.ics.uci.edu/dataset/380/youtube+spam+collection`): There are 5 different datasets in this case study, you need to work with at least two sets of data to compare the effectiveness of your best spam filtering predictive model. Note that you need to perform data preparation for the "comments" of the YouTube videos.
   - e-Commerce with the Online Shoppers Purchasing Intention Dataset (`https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset`): The data is rather large ($> 10k$ rows) and you will try to predict and infer the customer's shopping intention.
   - Dry Bean Data Set (`https://archive.ics.uci.edu/dataset/602/dry+bean+dataset`): Models should be developed to identify 7 different registered classes of dry beans in this case study.
   - Self-proposed dataset from UCI Machine Learning website and Kaggle should not be from earlier assignments or earlier than 2020 and must be relevant to the case studies of particular sectors in Malaysia and **requires** approval from lecturer by submitting the data and url of the data to the lecturer.
2. 0.5 out of 18 marks are allocated for the registration of data for the case study. 0.5 will be awarded to the first and second groups that register the same title. The third group to register the same title will receive 0.4, the fourth group will receive 0.2 while later groups will only get 0. This is to prevent too many groups going for the same set of data. Self-proposed data may be allowed upon lecturer's approval but will also receive 0 out of 18 marks. Those who change data after week 4 will also receive 0.
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 (the objectives cannot include 'trying out predictive models' but should be the usefulness of the data in dealing with real-world

problem). Your program script must work on the **original data** from the URL or a small portion of the marks will be deducted.

4. The report should be written in appropriate report format which is simple, neat and easy to refer and containing the following contents:
    - An introduction with appropriate references and objectives.
    - The understanding of the data is presented nicely with appropriate academic citations.
    - Unsupervised learning with EDA of the features and various advanced methods to identify interesting patterns from the data.
    - Supervised learning with various predictive models related to the data. The performance measures should be summarised appropriately for comparison.
    - A conclusion.

Group Programming Code (10%)

1. Write a programming code (or nicely structured programming codes) with an appropriate use of libraries which analyse the **original dataset** from the URL which is picked in the group assignment report and works in a data science pipeline. Marks are deducted if Excel or other expensive software (e.g. SAS, SPSS) is used to pre-process the data rather than automating the data processing in the programming code.

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 or the programming code.

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 free and legal statistical software tool and appropriate comments.

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 unsupervised learning and supervised learning results.
   - **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.
   - The presentation is well-timed (heavy marks may be deducted for over-time presentation), has a proper conclusion and is interesting (not reading literally from the slides).