DMAT – Assignment

Course	MSCBD-DMAT
Stage / Year	1
Module	Data Mining Algorithms & Techniques
Semester	2
Assignment	Assignment
Date of Title Issue	27th March 2024
Assignment Deadline (Non-graded)	7th April 2024 (Proposal + Peer Feedback)
Assignment Deadline (Graded)	5th May 2024 (Conference Paper + Code Notebook + Video presentation + individual learning journal)
Assignment Submission	Upload to Moodle
Assignment Weighting	50% of the module

Group Assignment

You will be working in groups of two or individually to complete this assignment. I suggest that you work through Google Drive with this document being stored as a Google Doc so that you can both works together. If you prefer using LATEX, then you can use www.overleaf.com for shared LATEX environment.

Objectives

1. This assignment involves selecting a topic and a relevant dataset; defining the aims and objectives of mining; designing and implementing the right mining techniques and reporting the results.
2. To successfully apply a set of data mining skills imparted in this module to a previously unseen dataset to achieve **knowledge discovery**.
3. Conduct an extensive and comprehensive **literature review** related to the selected problem.

Deliverables (Link will be provided for each)

1. **Conference style paper** for Part 1 (an 6- 8 page report in IEEE conference format).
2. **Jupyter Notebook file** that contains all your working. You should clearly use the **headings** for clarity with your notebook. Include the **dataset** or **a link to the dataset**. (If you used an API to retrieve your data, submit your **script for retrieving the data** as well as the data itself).
3. **Video presentation**
4. **Individual learning journal**. (Not applicable for individual project)

Part 1 - Classification/Association/Clustering/Time series or combination of them

Choosing Your Dataset

□ Your dataset should concern a real-world problem that lends itself to easy understanding by your classmates.

□ It should not be used by another group.

* Please refer to additional materials section in moodle for datasets links and suggested list of APIs.

* Please post to the student discussion forum “Dataset Selection” clearly indicating which dataset you are using so that other students do not select the same dataset.

Deliverables

1. **Proposal (including peer feedback)**
   You are required to submit a one-page description of your proposed problem, with clear question and hypothesis. Your proposal should include the source of at least one potential dataset and at least 4 references to support your proposal. "Peer feedback" refers to the exchange and evaluation of proposal between you and other students.
2. **Conference paper**
   By the end of this assignment, you are expected to produce an IEEE conference style paper (max 8 pages) that covers all aspects of data mining as discussed in the module. You must identify a testable, answerable, non-trivial research question and then formulate a methodology to answer that question, using one of the data mining frameworks (KDD or CRISP-DM).
   You are expected to do an extensive literature review that comprehensively covers all related work to the dataset(s) (problem) of your choice. You should critically evaluate your sources, describing the relation to your proposed solution. Your literature review should inform the choice of your problem and the suggested solution. Your resources should satisfy the three R’s rule: Related, Recent and Reputable.

   The suggested paper structure:

   i. Abstract
   ii. Introduction
   iii. Related Work
   iv. Methodology
   v. Evaluation and results
   vi. Conclusions and Future Work
   vii. References

   Within your paper, you should be able to cover the following points:

   i. Description of your dataset
   ii. Preprocessing and EDA
   iii. Training, testing and validation sets
   iv. Classifier(s) used / Association / Clustering
   v. Optimisation.

   [75 marks[1]]
3. **Detailed work in a Jupyter Notebook file.**

[1] Subject to random weekly checks on the progress including the repository.

[Will be checked to support the paper, if not present 20% will be deducted]

4. **Video presentation**[2]
   1. 10 minutes max
   2. All team members should participate

[10 marks]

Part 2– Individual Learning Journal (Individual Submission)

This is to be submitted in a separate moodle submission link.

In 500 words, you are required to reflect on your work within the group, what you did and what you learn within the process. You should also evaluate your contribution and your colleagues' contribution as well:

- Name (Member 1): 70%
- Name (Member 2): 30%

[15 marks]

[Total 100 marks]

Penalties:

1. Failure to submit a proposal	**-10%**
2. Failure to do Peer Feedback for the proposal	**-10%**
3. Failure to submit the Jupyter Notebook	**-20%**
4. Failure to use git repository (gitlab.griffith.ie)	**-100%**
5. Non active gitlab repository	**-10 to -20%**
6. Standard late submission will apply.	

[2] I suggest using zoom, and your camera should be on.