Data Mining and Machine Learning

Project Assignment

Course	MSCBD - MSCC
Stage / Year	1
Module	Data Mining and Machine Learning
Semester	1
Assignment	Project Assignment
Date of Title Issue	07 Nov 2015
Assignment Deadline (Graded)	- Paper, Artefacts and Video Presentation: Thursday 11th Dec @ 2 pm - Learning Journal Friday 12th Dec @ 2 pm
Assignment Submission	Upload to Moodle
Assignment Weighting	60% of the module

Group Assignment

You will be working in groups of two to complete this assignment. I suggest that you work through OneDrive so that you can both work 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. A report (3500 – 5000 words).
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.

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.
- You are encouraged to select a problem that is aligned to one of the UN sustainability goals.

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

Deliverables

1. By the end of this assignment, you are expected to produce a report 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 report, 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 (Hyperparameters tuning).

   [75 marks[1]]

2. Detailed work in a Jupyter Notebook file.

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

3. Video presentation[2]
   i. 10 minutes max
   ii. All team members should participate

   [10 marks]

[1] Subject to random weekly checks on the progress including the repository.
[2] I suggest using zoom, and your camera should be on.

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, for example:

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

[15 marks]

[Total 100 marks]

Penalties:

1. **Standard late submission will apply.**

Marking Rubric

Achievement	Excellent	Satisfactory	Basic	Unsatisfactory
% of Marks Available	>70%	55-70%	40-55%	<40%

	Weight	Excellent	Satisfactory	Basic	Unsatisfactory
Problem Definition	10	Well defined problem definition, justifications for selections. Excellent presentation and clarity.	Problem definition, medium to well defined. Clear and defined presentation.	Problem definition missing details and Impact. There are a few mistakes.	Poorly defined problem, and poor presentation.
Data Insights & Data Preparation	20	Good, focused insights from dataset select, good explanation, no trivial data analysis, selection of appropriate data preparation, explanations given,	Useful insights with explanations, impact of these on problem solution, selection of some appropriate data preparation, explanations given	Some insights given, limited details, limited data preparations, appropriateness of data prep, minimum explanations given	No or poorly selected data insights, limited or no data preparation
Algorithms selection and application	15	Suitable algorithms selected, good details on these and why, good details on experimentation, insights from experimentation, reflections, and discussion	Suitable algorithms selected, some details of selection and why, some details of algorithm experimentation, some discussion of experimentation	Suitable algorithms selected, limited details of selections given, limited details of application of algorithms given, limited details of algorithm settings and tuning	Limited or no details of selection and application of algorithms for data and problem. No explanations

	Weight	Excellent	Satisfactory	Basic	Unsatisfactory
Analysis of Results	20	Excellent detailed analysis of results and excellent insights of these. Clearly demonstrates impact and outcomes	Well detailed analysis of results, good level of insights on these, what then mean, their impact and outcomes	Some discussion of results, at a basic level with little insights	Little, no or very limited analysis of results and outcomes from tasks
Ethics implications and conclusions and further actions	15	Excellent level of discussion and insights. Clear well defined ethical and legal issues.	Good level of discussion of the action plan including some ethical and legal issues.	Some discussion of the action plan including some ethical and/or legal aspects.	Little or no discussion, simple overviews given, Little or no action plan including some ethical and legal aspects considered
Documentation writing, referencing and presentation	20	Excellent writing with minor errors. Excellent use of citation and reference is appropriate and relevant.	Good level of writing, with some grammatical and styling issues. The use of citation lacks precision and references need improvement.	The report does not reflect a good level of problem understanding, Styling and grammatical issues and references are not relevant nor well cited.	The documentation does not satisfy the minimum requirement, citation is very limited.