#### Data Mining – Assignment 2

Due on Sunday 28th April 2024 at 23:55

# Group Assignment

You will be working in groups of two to complete this assignment. I suggest that you work through One Drive with the sample document being shared 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. Comprehensive report (please refer to the template as a guideline) (**ONLY pdf**)
2. Jupyter Notebook OR Orange file OR both that contain 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).

1. Video presentation
2. 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.
* It should not be used by another group.

## Deliverables

1. By the end of this assignment, you are expected to produce a comprehensive report (max 15 pages) that covers all aspects of data mining as discussed in the module. You must identify a testable, answerable, non-trivial research/business 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**.

See the sample document/report (on BrightSpace). This document contains generic sections for you to complete for each part of this assignment. These sections are just suggestions! You can modify or use your own section headings. Whatever makes sense to you for completing the assignment and telling your story.

Only include details and images which are important for your findings and narrative. **Do not** fill your report with reports/charts/etc, which do not add to your discussion.

Include a References section for each Task. Each reference should be cited in your text. **Use a referencing/citation** method you are used to or select APA or Numeric Style.

**[75 marks]**

1. Detailed work in a Jupyter Notebook / Orange file.

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

1. Video presentation[[1]](#footnote-1)
2. 10 minutes max
3. All team members should participate

**[10 marks]**

# Part 2– Individual Learning Journal (Individual Submission)

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

In 300 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]**

# Marking Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Achievement** | Excellent | Satisfactory | Basic | Unsatisfactory |
| **% of Marks Available** | >75% | 55-75% | 40-55% | <40% |

|  | Weight | Excellent | Satisfactory | Basic | Unsatisfactory |
| --- | --- | --- | --- | --- | --- |
| Problem Definition and presentation | 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. Presentation has 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 |
| Application of  Algorithms | 15 | Suitable algorithms  selected, good details  on these and why, good  details on  experimentation,  insights from  experimentation,  reflections, and  discussion | Suitable algorithms  selected, some detail 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 |
| 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 |
| Learning for  work | 15 | Excellent level of  insights, brings  together details  through work, 4-8  citations comparing  related work in each task, clear  identification of  improvements,  reflection on learning  outcomes from tasks | Good level of  discussion of results,  identify some areas for  improvement, 3-5  citations used to  compare results in each task, identification of  some improvements  with limited discussion | Some discussion and evaluation of work,  some comparison with related research,  a limited number of  citations used | Little or no discussion or work, no  comparison with  related research |
| Action plan also considering some ethics implications | 20 | Excellent level of  discussion and insights for both case studies.  Clear well defined action plan, including some  ethical and legal issues.  Clear well defined  impact on role and  tasks | Good level of  discussion of the action plan including some ethical and legal issues for  both case studies.  Clearly reflection on  these and the roles and tasks | Some discussion of  the action plan including some  ethical and/or legal  aspects. Limited  discussion with little or  no reflection, and  impact on roles and  tasks | Little or no discussion,  simple overviews  given, Little or no  action plan including some ethical and legal  aspects considered |

1. I suggest using MS Teams, and your camera should be on. [↑](#footnote-ref-1)