CDS6314 DATA MINING

Trimester 2530

PROJECT (30%)

INSTRUCTIONS:

1. This project carries 30% of the coursework assessment.
2. This is a group project, with a maximum of 4 members.
3. Deliverables for this assignment include Python code (.ipynb / .py), a report (.pdf) and presentation.
4. Timelines:
   - Proposal deadline: **2nd January 2026 (Friday), 11.59pm**
   - Final submission deadline: **6th February 2026 (Friday), 11.59pm**.
   - Presentations schedule: **9th – 13rd February 2026**
5. Late-Day policy applies (10% deduction per day late from deadline).
6. If plagiarism is detected, the assignment will be granted 0% with no negotiation.

INTRODUCTION:

In this project, you will perform data mining on structured data from a domain of your choice and build a simple interactive application that visualizes patterns and also provides actionable insights or decision-support capabilities. You will design the full pipeline from dataset selection and preprocessing to analysis, modeling, and translating results into meaningful visualizations and actions within the interface. This project will be split into 2 phases, the proposal phase, and the implementation phase.

DATASET / TOPICS:

1. Search for dataset(s) / topic in a domain of interest from online repositories to conduct this project.
2. Dataset(s) / topics can be something your group is interested in with the condition that the topic is not exactly the same as the FYP of any group member.
3. The selected dataset should be structured, have at least 15 attributes (columns) and 5,000 instances (rows), and must be verified real data.
4. Sources:
    - UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
    - Kaggle (https://www.kaggle.com/data)
    - Malaysia's Open Data Portal (https://data.gov.my/data-catalogue)
    - World Bank Open Data (https://data.worldbank.org/)
    - The Humanitarian Data Exchange (https://data.humdata.org/)
    - Other relevant and legitimate sources.

DESCRIPTION OF TASKS:

Phase 1 (Proposal) [1 week].

Prepare a simple write-up of the selected dataset and preliminary details of the project.

The proposal should include the following details:

1. Cover page: Project members and project title
2. Project description:
    - *Dataset*:

      What is the domain of the dataset? What is the source? Provide the link of the source, introduce the dataset (attributes, records, etc.) and the collection process of the original source. (Optional: can include preliminary data exploration, statistics, etc.)
    - *Data Mining Task*:

      What is the data mining task you intend to execute? Explain the motivation, aims, and expected outcomes of the project.
    - *Application Mockup:*

      Simple draft of your idea for the application interface. Who is the target user and how do you expect them to interact with the application?
    - *Related Works*:

      Which reading will you examine to provide context and background? Are there similar work or applications? List the sources in APA citation format and briefly state why you chose them.
    3. Preliminary Task Distribution
    4. Declaration of FYP
    5. References

Phase 2 (Final) [5 weeks].

Final report, source code submission, presentation, and peer evaluation.

Suggested structure for **Final Report**:

1. **Cover page:** Title, Authors
2. **Abstract:** Summary of overall project, from motivation to conclusion. (maximum 250 words)

3. **Introduction:** Introduce the background, motivations, and objectives.
4. **Related works:**

   Review or summarize works (research papers or applications) that used the same / similar / related dataset or have a similar purpose.

   Discuss the similarities and differences of those works with this project.
5. **Methodology**

   Describe the overall work done to achieve the project objectives and provide justification on the steps and methods selected. The report should include, but not limited to, the following items:

   a. *Overall framework*: introduce the overall pipeline and data mining task.
   b. *Dataset*: source, collection process, volume, attributes, etc.
   c. *Data Preprocessing*: EDA, feature selection, data transformation, etc.
   d. *Data Mining*: Techniques used, parameters, etc.
   e. *Evaluation*: Experiments conducted, evaluation metrics, comparisons, etc.
   f. *Results and Discussion*: Compile results generated (tables, charts, etc.) and discuss the outcomes and findings.
   g. *Application*: Dashboard, web tool, etc. (Streamlit, Heroku, Flask, etc.)
6. **Conclusion**

   Summarize the overall work and discuss potential use cases or importance. Suggest potential future directions of the work (e.g. how to overcome limitations, other dimensions of exploration, etc.)
7. **References:** APA format
8. **Appendix**

Note: It is not necessary to screenshot and show the python codes in the technical report. Instead, use text descriptions, algorithms, visualizations/flowcharts, or others to explain the work.

Source Code preparation:

Source code of application/implementation must be in Python.

Please ensure the code can be used in different machines and they can be in Python scripts (.py) or Python Notebook (.ipynb). If any special instructions are needed for building or running the code, please provide a readme file.

You may create separate python scripts/notebooks for each key step, if necessary.

Please include as Appendix in the Final Report of any tutorials, GitHub codes, websites, videos, etc. used for learning and reference to complete the project.

Presentation guidelines:

Present the overall project from motivation to insights and demo of application.

- All members of the group must attend the face-to-face presentation, and each should talk about one part of the work done.
- Every group must prepare slides for the presentation (max. 10 slides including title slide with group details and ending slide).
- Additional tools can be used to make the presentation more effective (figures, tables, animations, etc.).
- Prepare a short demonstration of the application during the presentation.
- Maximum duration of presentation: 15 minutes including demo.

Peer Evaluation:

Part of the final project marks will be based on the average marks given by group members.

The following is the calculation of the peer evaluation:

Number of group members who submitted peer evaluation form $= n$

Peer evaluation marks given to student by another member $= P_i \in [0,20]$

Peer evaluation score for a student,

$$PE = \frac{\frac{1}{n}\sum_{i=1}^{n} P_i}{10}\%$$

$PE = 0\%$ for the member who did not submit peer evaluation form.

SUBMISSION CHECKLIST:

All submissions to be made via eBwise.

1. Phase 1: Proposal submission (.pdf).

- Name the submission file following this format:
  2530_CDS6314_Project_Proposal_Group_P[Number].pdf
  (e.g. *2530*_CDS6314_*Project_Proposal_Group_P01.pdf*)
- Deadline: **2nd January 2026 (Friday), 11.59pm**

2. Phase 2: Final report (.pdf), Source code (.py / .ipynb), Presentation slides (.pptx)

- Please ensure the codes can reproduce the results and run the application given the raw data in any machine.
- Include a Readme.txt of instructions to navigate the dataset and codes (if necessary, especially if multiple data files and scripts)
- Compile everything into a .zip for submission. Name the folder following this format:
  2530_CDS6314_Project_Final_Group_P[Number].pdf
  (e.g. *2530_CDS6314_Project_Final_Group_P01.pdf*)
- Deadline: **6th February 2026 (Friday), 11.59pm**

3. Presentation:

- Schedule: **9th – 13rd February 2026**
- Specific slots and location to be announced later.

PROJECT RUBRICS:

Proposal (3%)		3
Report (8%)	Clarity, structure, language, reference format	1
	Understanding of related works	1
	Motivations and objectives	1
	Methodology	2
	Analysis	3
Technical (12%)	Correctness	3
	Depth	3
	Complexity	3
	Application Utility	3
Presentation (5%)		5
Peer Evaluation (2%)		2
Total		**30%**