CDS6314 DATA MINING

Trimester 3, 2024/2025

PROJECT (30%)

Deadline: **Week 14, 9th February 2025 (Sunday), 11.59pm**

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*), a report (*.pdf*) and presentation.
4. Late-Day policy applies (20% deduction per day late from deadline).
5. If ***plagiarism*** is detected, the assignment will be granted 0%.

INTRODUCTION:

In this project, your task is to perform data mining on structured data of your domain of interest. You will be required to devise the full pipeline, from dataset search till knowledge visualization to gain hands-on experience on the implementation of a data mining solution on practical data.

DATASET / TOPICS:

1. Based on your tutorial section, your group can only work on dataset within the domain that has been assigned.

Tutorial Section	Domain
TT1L	Agriculture
TT2L	Accommodation and Tourism
TT3L	Financial / Banking
TT4L	Medical
TT5L	Nutrition and Diets
TT6L	Climate and Environment

2. Search for dataset(s) / topic in a domain of interest from online repositories to conduct this project. The topic must not be the same as the FYP of any group member.
3. The selected dataset should be structured and have at least 20 attributes (columns) and 5,000 instances (rows).
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 (http://www.data.gov.my/data/en_US/dataset)
    - DOSM Malaysia (https://open.dosm.gov.my/data-catalogue)
    - World Bank Open Data (https://data.worldbank.org/)
    - The Humanitarian Data Exchange (https://data.humdata.org/)
    - The Omdena datasets (https://datasets.omdena.com/)
    - Other relevant and legitimate sources.

DELIVERABLES:

1. Report
2. Source code (Python notebook)
3. Presentation slide

Report (strictly PDF only):

1. Cover page: Title, Authors, YouTube and dataset links

2. Abstract: Summary of overall project, from motivation to conclusion. (maximum 250 words)

3. Introduction: Introduce the background, motivations, and objectives.

4. Literature review:

Review at least 4 related research papers that used the same / similar / related dataset. Discuss the similarities and differences of those papers with this project.

5. Methodology

Describe the overall data mining pipeline implemented 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. (Optional) Deployment: Dashboard, web tool, etc. (Streamlit, Heroku, etc.)

6. Conclusion

Summarize the overall findings of the work and discuss potential use case 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

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.

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

Source Code (Python Notebook ipynb only):

Source code of application/implementation must be in Python.

Please ensure the code can be used in different machines and in Python Notebook (.ipynb). Submit only ONE notebook per group. If any special instructions are needed for building or running the code, please provide a readme file.

Presentation:

- Present the overall project from motivation to insights.
- Record your group's presentation, upload it to YouTube and kept as unlisted. All members in the group must participate in the presentation and introduce themselves with the camera turn on.
- 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.).
- Maximum duration of presentation: **10 minutes** including demo.
- The YouTube link must be included on the Cover Page of the report.
- Submit your presentation slides in pptx format.

SUBMISSION INSTRUCTIONS:

- Submit all the **deliverables** to **eBwise** according to your registered group before the deadline. ONE submission per group.
- **Do NOT submit any zip file.**
- Late submission is acceptable with penalty of 10 marks per hour. Zero marks will be awarded for submission after 4 hours.

PENALTIES:

- 10 marks will be deducted for each hour late after the deadline.
- 0 mark will be awarded for this Project if the content of this Project is plagiarized from any sources
- 0 mark will be awarded for this Project if the group submit the Project 4 hours late.
- 3 marks will be deducted for the video that exceeds 10 minutes
- 3 marks will be deducted for submitting a slide that exceeds 10 pages.
- 5 marks deducted for not having a Cover Page.
- 20 marks will be deducted for working with data in different domain other than being assigned per tutorial section

PROJECT RUBRICS:

Deliverable		Marks
Report (10%)	Cover page	2
	Clarity, structure, language, reference format	2
	Understanding of the background and literature	2
	Motivations and objectives	2
	Data mining task formulation and justification	2
Technical Analysis (15%)	Correctness	3
	Depth	3
	Complexity	3
	Results comparison and analysis	3
	Discussion on findings and insights	3
Presentation (5%)		5
Total		**30%**