CDS6314 DATA MINING

Trimester 1, 2024/2025

ASSIGNMENT (20%)

INSTRUCTIONS:

1. This assignment carries 20% 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*) and a report (*.pdf*).
4. Submission deadline: **29th December 2024 (Sunday), 11.59pm, Week 8.**
5. Late-Day policy applies (20% deduction per day late from deadline).
6. If ***plagiarism*** is detected, the assignment will be granted 0% with no negotiation.

INTRODUCTION:

This project combines **association rule mining** and **classification** to uncover patterns and predict if the candidates are likely to further to higher studies or to start working based on the dataset.

OBJECTIVE:

1. **Association Rule Mining**: Uncover interesting associations between user profiles, experiences, and ratings.
2. **Classification**: Build a model to predict if the candidates are likely to further their study or to start working.

Dataset: Career dataset (attached with the assignment and also available via the provided Link)

PYTHON TASK:

Based on the stated dataset, devise a minimum of 4 **exploratory questions** to be answered by the following tasks.

Association Rule Mining Task:

Devise a pipeline for preprocessing, mining, and knowledge evaluation, then implement the python codes for the process. Steps should include, but not limited to the following:

1. Data Exploration (statistics and visualization)
2. Data Preprocessing (cleaning, transformation)
3. Data Mining (association rule mining) (e.g., Reading/Writing Skills, Memory Capability, Hard/Smart Worker, Worked in Teams Ever, Leadership Experience, etc.)
4. Knowledge evaluation (interestingness measure)

Classification Task:

Implement at least 2 classification models based on the dataset to predict if the candidates are likely to further their study or to start working.

Note:

- *Do NOT create separate python notebooks for the* **association rule mining and classification** *tasks. You should design your processing pipeline carefully and explain your steps/results whenever necessary in your notebook.*
- *Please include a reference list at the end of the notebook(s) of any tutorials, GitHub codes, websites, videos, etc. used for learning and reference to complete the tasks.*

TECHNICAL REPORT:

Write a technical report to introduce the domain including related research, and compile the details and results of the python task. The report should include the following items:

1. **Cover page:**
    a. Title
    b. Group number and members
    c. Contribution of each member
2. **Introduction**
    a. Introduce the background and motivations
    b. Review at least 3 related research papers that used the same / similar / related dataset
        - Discuss about the differences and similarities of the dataset / data mining of those papers with this work
3. **Formulating Exploratory Questions**
    a. Create questions to explore and find out potentially frequent patterns in the dataset
    b. Explain and justify the potential subjective interestingness of the outcomes from answering the formulated questions
4. **Data Exploration**
    a. Describe the dataset
    b. Explain the initial exploration done the data by showing statistics and visualizations
5. **Data Preprocessing**
    a. Describe the data cleaning steps (if any) and justify the approach
    b. Describe the steps taken to transform the raw data into form suitable for data mining including justification
    c. Note: Possible processes include data discretization, concept hierarchy generation, feature selection, etc.
6. **Association Rule Mining**
    a. Details on the application of association rule mining on the processed data, including choices of interestingness measures
    b. Compile the rules generated from mining and identify interesting patterns

7. Classification

Describe the classification algorithm applied.

8. Results Discussion

a. Discuss the results generated from association rule mining.

b. How do the results answer the formulated exploratory questions?

c. Are there certain factors in the preprocessing that influence the rule generation?

d. What are the attributes affecting the classification result, listed in the descending order? Justify your findings.

9. Conclusion

a. Summarize the overall findings of the work

b. Discuss potential use case or importance of the findings

c. Suggest potential future directions of the work (e.g. how to overcome limitations, other dimensions of exploration, etc.)

10. References

SUBMISSION:

Submit the following files to Ebwise:

- Python notebook codes (.ipynb)
  - Please ensure the codes can reproduce the results given the raw data
  - Include a Readme.txt of instructions to navigate the notebooks (if necessary, especially if multiple notebooks)
- Technical report (.pdf)

Note:

- *Name the submission files with your Group number (e.g. Group01.ipynb, Group01.pdf).*
- *Only group leaders need to submit. Submission by group members will not be marked. (The first name in the group list is the group leader).*

MARKS DISTRIBUTION:

Code (12%)	Data Exploration	2
	Data Preprocessing	2
	Association Rule Mining	4
	Classification	3
	Visualizations	1
Report (8%)	Introduction and Literature Review	2
	Questions Formulated	2
	Analysis of Findings	3
	Conclusion	1
Total		**20%**