TDS3301 DATA MINING

Trimester 1, 2023/2024

ASSIGNMENT (20%)

INSTRUCTIONS:

1. This assignment carries 20% of the coursework assessment.
2. This assignment is to be completed individually.
3. Deliverables for this assignment include Python code (.ipynb) and a report (.pdf).
4. Submission deadline: **20th December 2023 (Wednesday), 11.59pm**.
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:

Association rule mining finds common patterns in data. Often used for market basket analysis to understand customer patterns, it is also possible to be applied on other types of datasets to find interesting associations and relationships between attributes.

OBJECTIVE:

To perform association rule mining on higher education students' data and find interesting associations.

Dataset: Higher Education Students Performance Evaluation (Link)

Reference: Yılmaz, N., & Sekeroglu, B. (2019, August). Student performance classification using artificial intelligence techniques. In *International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions* (pp. 596-603). Cham: Springer International Publishing.

PYTHON TASK:

Based on the stated dataset, devise a **minimum of 4 exploratory questions** to be answered via association rule mining. Following that, 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)
4. Knowledge evaluation (interestingness measure)

Note:

- *You may create separate python notebooks for the tasks if necessary.*
- *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. 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

2. 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

3. Data Exploration

a. Describe the dataset

b. Explain the initial exploration done the data by showing statistics and visualizations

4. 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.

5. 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

6. 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?

7. 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.)

8. References

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.

SUBMISSION:

Submit the following in a ZIP file via MMLS assignment:

- 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 file using your student ID (e.g. 1001101010.zip)*
- *You do not need to resubmit the raw data.*

MARKS DISTRIBUTION:

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