MULTIMEDIA UNIVERSITY

CDS6314 DATA MINING

Trimester March/April 2025 (Term 2510)

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) (Do NOT submit any zip file).
4. Submission deadline: **13th May 2025 (Tuesday), 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 marital satisfactions' data and find interesting associations.

Dataset: Marital satisfaction, sex, age, marriage duration, religion, number of children, economic status, education, and collectivistic values: Data from 33 countries.

Reference: Sorokowski, P., Randall, A. K., Groyecka, A., Frackowiak, T., Cantarero, K., Hilpert, P., Ahmadi, K., Alghraibeh, A. M., Aryeetey, R., Bertoni, A., Bettache, K., Błażejewska, M., Bodenmann, G., Bortolini, T. S., Bosc, C., Butovskaya, M., Castro, F. N., Cetinkaya, H., Cunha, D., … Sorokowska, A. (2017). Marital satisfaction, sex, age, marriage duration, religion, number of children, economic status, education, and collectivistic values: Data from 33 countries. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01199

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. **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 Preprocessing**
    a. Verify data quality: How clean/dirty is the data? Document any quality issues.
    b. Describe the data cleaning steps (if any) and justify the approach
    c. Describe the steps taken to transform the raw data into form suitable for data mining including justification

    Note: Possible processes include Data encoding, scaling, normalization, and transformation, data discretization, concept hierarchy generation, feature selection, etc.
5. **Data Exploration**
    a. Describe the dataset
    b. Explain the initial exploration done the data by showing statistics and visualizations
    c. Univariate Analysis: Distributions of the features/ variables , summary statistics
    d. Bivariate Analysis: Correlation Analysis
    e. Multivariate analysis: Correlation analysis
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. **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?

8. 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.)

9. References: APA format v7

10. Appendix

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 the work. If you need to highlight important Python code, put it in an appendix.

SUBMISSION:

Submit the following in a ZIP file via eBwise 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 Group number (e.g. TTxL_G#.ipynb)*
- *You do not need to resubmit the raw data.*

MARKS DISTRIBUTION:

Code (100%)	**Evaluation Criteria**	**Weight (%)**
Code Quality	Code is well-structured, modular, and follows PEP 8. Proper use of functions/classes for reusability. Comments/docstrings are clear and explain logic. Introductory comment describes overall strategy and gives evidence of preliminary planning. Thoughtful problem decomposition breaks the problem into independent pieces that can be solved easily. Code is a pleasure to read, and easy to understand. Code and comments form part of a seamless whole.	15.00%
Data Preprocessing	Missing data is handled appropriately. Categorical data is encoded correctly. Data is normalized/scaled if necessary. Outliers are detected and managed.	15.00%
Data Exploration	Data Exploration & Visualizations Univariate Analysis Bivariate Analysis Multivariate analysis (Intense exploration and evidence of many trials and failures. You have looked at the data in many different ways before coming to your final answer. You have gone beyond what was asked: additional research from other sources used to help understand/explain findings. Your explanation and presentation is creative.)	25.00%
Association Rule Mining	Association Rule Mining & Visualizations	30.00%
Reproducibility	Code runs without errors in a fresh environment. Notebook	5.00%

	runs without issue.	
Technical Proficiency	Expertise in relevant tools, technologies, or methodologies to carry out tasks efficiently and effectively. Mastery of Python vocabulary means that the absolute minimum amount of code is used to get the job done. Code free from duplication. Each function encapsulates a single task, and repeated tasks are performed by functions, not copy and paste.	5.00%
Collaboration and Communication	Collaboration and Communication: Coordinating with cross-functional teams and communicating technical aspects to non-technical stakeholders to ensure alignment and support.	5.00%
Report (100%)	Clarity, structure, language, reference format Clear headings demarcate separate sections. Excellent flow from one section to the next. Tables and graphics carefully tuned and placed for desired purpose.	10.00%
	Abstract Abstract Structure: Introduction: Start with the purpose of the study and its significance. Methods: Briefly describe the dataset, preprocessing, and techniques used. Results: Highlight key findings and comparisons between methods. Impact: Emphasize the practical applications and contributions of the research. Future Work: Conclude with directions for further research.	10.00%
	Introduction and Literature Review: Understanding of the background and literature Motivations and objectives	20.00%

	Questions Formulated Data mining task formulation and justification	20.00%
	Analysis of Findings: Results comparison and analysis Discussion on findings and insights Findings very well organised. (You suggest multiple explanations for a given finding, and use multiple tools to explore surprising results.)	30.00%
	Conclusion	10.00%
Total		**200%/ 10**