CDS6314 DATA MINING

Trimester 2530

ASSIGNMENT (20%)

INSTRUCTIONS:

1. This assignment carries 20% of the coursework assessment.
2. This is a group project, 2-person group.
3. Deliverables for this assignment include Python code (*.ipynb*) and a report (*.pdf*).
4. Submission deadline: **Week 8, 26th Dec 2025 (Friday), 11.59pm.**
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:

Online platforms collect large volumes of customer data, including demographics, membership, usage patterns, purchases, and feedback. Analyzing this data can reveal hidden patterns useful for marketing, recommendations, and customer retention. This assignment focuses on association rule mining, a technique for discovering relationships between attributes in large datasets. Your task is to perform data mining on structured customer data, designing and implementing the full pipeline from data exploration and preprocessing to association rule mining and knowledge visualization.

OBJECTIVE:

To perform association rule mining on Amazon Prime User data and find interesting associations between user profile, experience and rating.

Dataset: *Amazon_Prime_Users (attached with the assignment file).*

A. PYTHON TASK:

Based on the provided dataset, your group is required to formulate four exploratory questions that can be answered through association rule mining. You will then implement a full data mining pipeline in Python that covers the following stages:

1. ***Data Exploration*:** Analyze the dataset to understand its structure and content. Include summary statistics, distributions, and visualizations that highlight key patterns or anomalies.
2. ***Data Preprocessing*:** Prepare the data for mining by cleaning, handling missing values, transforming variables, and creating meaningful features. Explain each step and justify your choices.
3. ***Data Mining*:** Apply association rule mining (e.g., Apriori or FP-Growth) to identify frequent itemsets and generate association rules. Clearly indicate the antecedent (if) and consequent (then) of each rule.
4. ***Knowledge Evaluation*:** Evaluate the discovered rules using interesting measures. Interpret the findings in context, explaining their significance and potential actionable insights.

In your notebook, clearly indicate which code cells correspond to each stage of the pipeline to help readers follow your workflow. Ensure your code is reproducible and well-documented

B. 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

Title, group number, and members.

2. Introduction

Provide the background and motivation for this study. Review at least four related research papers that have used Amazon datasets. Discuss the similarities and differences between the datasets and data mining techniques of those studies compared to your work.

3. *Formulating Exploratory Questions*

You must create four exploratory questions that can *reasonably* be answered through association rule mining. Each question must be justified: why is it relevant, what pattern is expected, and

why the result may be meaningful for business decision-making. (*Sample questions are provided later in this document.*)

4. ***Data Exploration***

Conduct a comprehensive exploration that may include:

- Attribute descriptions and dataset structure
- Summary statistics and distribution plots
- Detection of unusual patterns or anomalies
- Identification of variables potentially suitable for association mining

Your exploration must be supported by meaningful visualisations.

5. ***Data Preprocessing***

You must implement at least two preprocessing strategies. Explain the pre-processing steps and how these would affect the resulting association rules later. Describe the transformations performed to prepare the data for mining, such as discretization, feature selection, or generation of concept hierarchies, and provide reasoning for each choice.

6. ***Association Rule Mining***

Detail the application of association rule mining on the preprocessed dataset, including the selection of support, confidence, lift, or other interestingness measures. Present the rules generated and highlight any particularly notable or meaningful patterns.

7. ***Results Discussion***

Discuss the outcomes of the association rule mining, explaining how the results answer the exploratory questions formulated earlier. Comment on whether specific preprocessing choices influenced the rules and highlight insights or trends observed.

8. ***Conclusion***

Summarize the overall findings of the analysis and discuss their practical implications or potential applications. Suggest directions for future work, such as addressing limitations, exploring additional patterns, or applying alternative techniques.

9. References

List of References.

10. Group Contribution Declaration

To ensure fair and transparent allocation of marks, each group MUST include a detailed Group Contribution Declaration in the technical report (last page). Please note that a 5% deduction will be applied if the contribution declaration is NOT included in the submitted report. This declaration must:

1. List all major tasks performed in the assignment.
2. State each member's contribution percentage (total must sum to 100%).
3. Assign contribution percentages for EACH task, showing who did what.
4. Be signed by both students (typed signatures allowed).
5. Reflect genuine contribution. Misrepresentation may be treated as academic misconduct.

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 files to eBwise:

- Python notebook codes (.ipynb)
- Technical report (pdf)

Note:

Please name your submission files using your Tutorial Section and Group Number using these format: TT1L_Group01.ipynb and TT1L_Group01.pdf for example for Group 1 in TT1L. You do not need to resubmit the raw dataset.

Sample Exploratory Questions (for Student Reference ONLY)

These are examples; you must create your own questions.

Q1. Are long-term Prime members more likely to give higher satisfaction ratings for video streaming services?

- *Why it's interesting:* Helps understand retention and identify loyal user traits.

Q2. Does high viewing frequency of movies associate with higher likelihood of renewing Prime membership?

- *Why it's interesting:* Reveals behavioural patterns linked to subscription renewal.

Q3. Are young adult users (18–30) more likely to give positive feedback when they consume multiple content types (movies, series, sports)?

- *Why it's interesting:* Helps design age-targeted recommendations.

MARKS DISTRIBUTION:

Code (12%)	Data Exploration	3
	Data Preprocessing	3
	Association Rule Mining	5
	Visualizations	2
Report (8%)	Introduction and Literature Review	2
	Questions Formulated	2
	Analysis of Findings	2
	Conclusion	1
Total		**20%**