Association Rule Mining using Online Retail Dataset

Submitted as part of Data Mining / Machine Learning Assignment

Objective: To analyze customer purchasing behavior using Association Rule Mining techniques.

1. Introduction

Association Rule Mining is a data mining technique used to discover interesting relationships between variables in large datasets. It is widely applied in market basket analysis to identify products that are frequently purchased together. This assignment applies association rule mining on the Online Retail dataset to extract meaningful patterns that can help businesses improve marketing strategies and customer experience.

2. Dataset Description

The Online Retail dataset contains transactional data from a UK-based online retail store. Each record represents a purchased item along with invoice number, quantity, date, and customer details. For this assignment, invoice numbers and product descriptions are used to construct transaction baskets.

3. Data Preprocessing

Before applying association rule mining, the dataset was preprocessed to ensure data quality. Missing values in essential columns such as InvoiceNo and Description were removed. Cancelled transactions were excluded to avoid misleading patterns. Product descriptions were cleaned by removing extra spaces and inconsistencies.

4. Methodology

The Apriori algorithm was used to identify frequent itemsets. Transactions were grouped by invoice number and transformed into a one-hot encoded format. Minimum support, confidence, and lift thresholds were applied to extract meaningful association rules.

5. Association Rule Metrics

Support: Measures how frequently an itemset appears in the dataset. It helps identify commonly purchased products.

Confidence: Indicates the likelihood of purchasing item Y when item X is purchased. Higher confidence means a more reliable rule.

Lift: Evaluates the strength of an association by comparing it to random chance. A lift value greater than 1 indicates a positive association.

6. Analysis and Interpretation

The generated rules revealed several strong associations between products. Rules with high lift values indicate product combinations that occur more often than expected. These insights can be used for

cross-selling, promotions, and recommendation systems.

7. Interview Questions and Answers

1. **What is Association Rule Mining?**
   Definition, purpose, and use in market basket analysis.

2. **Explain the Apriori Algorithm.**
   Step-by-step explanation, pruning principle, and why it is efficient.

3. **What is Support?**
   Definition, formula, and significance.

4. **What is Confidence?**
   Definition, formula, and interpretation.

5. **What is Lift and why is it important?**
   Meaning of Lift > 1, Lift = 1, Lift < 1, and why it is better than confidence alone.

6. **Difference between Support, Confidence, and Lift**
   Tabular / conceptual comparison (commonly asked in viva).

7. **What are the limitations of Association Rule Mining?**
   Computational complexity, large rule generation, no causality.
8. **What is Lift and why is it important?**

   Lift measures the strength of a rule over random occurrence and helps filter meaningful rules.

9. **What are Support and Confidence?**

   Support shows frequency, while confidence shows rule reliability.

10. **What are limitations of Association Rule Mining?**
    Large number of rules, high computational cost, and lack of causal inference.

8. Conclusion

Association Rule Mining provides valuable insights into customer buying behavior. The Apriori algorithm successfully identified meaningful product relationships that can support business decision-making.