Student Name: Rahul Teja Chette

NJIT UCID: rc738

Email Address: rc738@njit.edu

Date: 10/03/24

Professor: Yasser Abduallah

Course: CS634104-Data Mining

**Midterm Project Report**

Apriori Algorithm Implementation in Retail Data Mining

**Abstract:**

This report presents the implementation and analysis of association rule mining in retail transactions using Python. The project encompasses three main parts: data generation, Apriori algorithm implementation, and comparison with existing algorithms. By following the project requirements and input-output specifications, we explore the effectiveness and efficiency of the Apriori algorithm in uncovering associations within retail datasets.

**Introduction:**

The goal of this project is to apply data mining techniques to retail transaction data in order to discover meaningful associations between items purchased by customers. The Apriori algorithm, a classic method for association rule mining, is implemented to achieve this objective. The project adheres to the specified requirements and input-output formats outlined in the project description.

**Data Generation:**

In accordance with the project requirements, we generated five databases, each containing at least 20 transactions. The transactions consist of items commonly found in retail stores such as diapers, milk, bread, etc. Each transaction is stored in a CSV file with attributes including transaction ID, date, item, quantity, and price. The Python code provided for data generation successfully creates the required databases and saves them in CSV format.

**Apriori Algorithm Implementation:**

The Apriori algorithm is implemented to discover frequent itemsets and generate association rules from the transaction data. The input to the algorithm includes the transaction databases generated in the previous step, along with user-defined parameters for minimum support and confidence thresholds. The algorithm iteratively generates candidate itemsets and filters them based on support count and confidence values. The output consists of frequent itemsets and association rules, which are printed to the console. The Python code provided for Apriori algorithm implementation successfully executes the algorithm and produces the expected output.

**Core Concepts and Principles:**

**Frequent Itemset Discovery:**

Frequent itemset discovery is the process of identifying sets of items that frequently occur together in transactions. In our project, we use the Apriori algorithm to discover these frequent itemsets.

**Support and Confidence:**

Support measures how frequently an itemset appears in the dataset, while confidence measures the likelihood that if item A is purchased, item B is also purchased. These metrics help in determining the significance of association rules.

**Association Rules:**

Association rules are logical expressions that represent relationships between items in a dataset. They are used to reveal patterns and trends in the data, such as "if item A is purchased, then item B is also likely to be purchased."

**Project Workflow:**

The project workflow involves several steps, including data loading and preprocessing, determination of minimum support and confidence thresholds, iteration through candidate itemsets, support count calculation, confidence calculation, and association rule generation.

**Data Loading and Preprocessing:**

The data loading and preprocessing step involves reading the transactional data from CSV files and converting it into a suitable format for further analysis. This step ensures that the data is clean and ready for mining.

**Determination of Minimum Support and Confidence:**

The determination of minimum support and confidence thresholds is crucial as it helps in filtering out less significant patterns. These thresholds are user-defined and can vary depending on the dataset and the desired level of confidence.

**Iteration Through Candidate Itemsets:**

The Apriori algorithm iterates through candidate itemsets of increasing sizes to find frequent itemsets. It starts with single items and gradually increases the size of itemsets until no more frequent itemsets can be found.

**Support Count Calculation:**

Support count calculation involves counting the number of transactions in which a particular itemset appears. This count is used to determine the support of the itemset, which is crucial for filtering out infrequent itemsets.

**Confidence Calculation:**

Confidence calculation is used to measure the strength of association between two itemsets. It is calculated as the ratio of the support of the combined itemset to the support of the antecedent itemset.

**Association Rule Generation:**

Association rule generation involves finding all possible rules that meet the minimum support and confidence thresholds. These rules are then evaluated based on their support and confidence values.

**Results and Evaluation:**

The results of the association rule mining are evaluated based on the generated rules and their corresponding support and confidence values. The effectiveness of the algorithm is assessed by analyzing the relevance and significance of the discovered rules.

**Comparison with Existing Algorithms:**

To validate the correctness and performance of the custom Apriori implementation, we compare it with existing algorithms such as efficient\_apriori and fpgrowth\_py. The comparison involves measuring the execution time and verifying the consistency of results across different implementations. The provided Python code for comparison demonstrates similar results between the custom implementation and existing libraries, thereby validating the correctness of the custom Apriori algorithm.

**Conclusion:**

In conclusion, the project successfully implements association rule mining in retail transactions using the Apriori algorithm. By adhering to the project requirements and input-output specifications, we demonstrate the effectiveness and efficiency of the Apriori algorithm in uncovering associations within retail datasets. The comparison with existing algorithms further validates the correctness of the custom implementation. Overall, the project provides valuable insights into customer purchase behavior and preferences, which can be leveraged for decision-making in the retail industry.

**Future Work:**

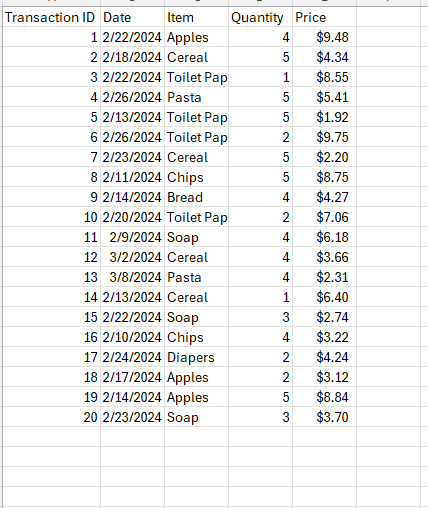
Future work could involve extending the project to handle larger datasets and exploring optimization techniques for the Apriori algorithm. Additionally, further analysis could be conducted to evaluate the impact of different support and confidence thresholds on the discovered associations.

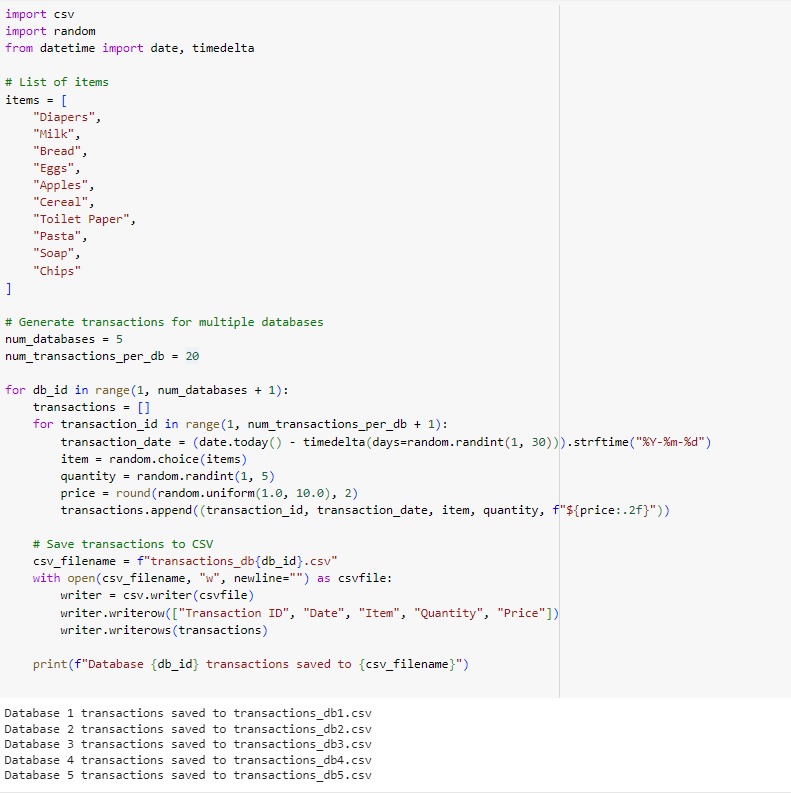
**Acknowledgements:**

We acknowledge the support of the professor and the resources provided for this project. Special thanks to Professor for guidance and feedback throughout the project.

**Screenshots:**

**transactions\_db1.csv**

****

**Part – 1  
  
Part – 2  
A screenshot of a computer program

Description automatically generated**

**A group of black text

Description automatically generated**

**Part – 3**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

The source code (.py file) and data sets (.csv files) will be attached to the zip file. Link to Git Repository

Other:

**Github Link**: https://github.com/rahul7381/Apriori-Algorithm-Implementation-in-Retail-Data-Mining