Unsupervised Machine Learning Project

Naveen Kavitha Gunasekaran, Raghu Ram Baskaran, Anjana Deivasigamani, Priyadharshan Sengutuvan

Abstract:

This project aims to explore and analyze a retail transaction dataset using dimensionality reduction and association rule mining techniques covered in our course modules. The dataset contains detailed information about customer purchases, including 522,065 rows and 7 attributes: BillNo (6-digit transaction ID), Itemname (product description), Quantity, Date (timestamp), Price, CustomerID (5-digit), and Country from the public kaggle website - https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis/data.

In the first half of our analysis, we will apply Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while preserving its key characteristics. PCA will help us identify the most important features driving purchasing patterns and customer behavior. We will explore both linear and non-linear PCA approaches, including Kernel PCA, to capture complex relationships in the data.

For the second half, we will employ association rule mining algorithms, specifically the Apriori and FP-Growth algorithms, to uncover frequent itemsets and generate meaningful association rules. These rules will reveal patterns in customer purchasing behavior, such as which products are frequently bought together.

Our exploratory data analysis goals include:

1. Customer Segmentation and Behavior Analysis
    - Apply K-means clustering to group customers based on purchasing patterns
    - Use association rule mining to identify frequently co-purchased items
    - Analyze the relationship between customer geography and product preferences
2. Temporal Trends and Seasonality
    - Employ time series analysis to detect patterns in sales volume and product popularity
    - Explore potential seasonal trends in purchasing behavior
3. Product Performance and Inventory Optimization
    - Conduct Pareto analysis to identify top-selling products and their contribution to overall revenue
    - Use frequent itemset mining to inform inventory management decisions

We anticipate that some of these approaches may not yield significant results due to data limitations or unforeseen complexities. However, the diversity of our analytical approach ensures that we will gain valuable insights to inform business decisions. By leveraging these data-driven insights, we aim to demonstrate the value of advanced analytics in retail operations and provide a foundation for ongoing data-driven decision-making processes.