Market Basket Analysis and Customer Segmentation

Md Masudur Rahman
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
masudur.cse.20210204009@aust.edu

Rakib Ahmed Ovi
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
rakib.cse.20210204001@aust.edu

Nazmul Hoda Munna
Dept. of Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka-1208, Bangladesh
nazmul.cse.20210104024@aust.edu

***Abstract*—This research focuses on analyzing customer purchasing behavior and segmenting customers using machine learning techniques. We used a dataset of customer transactions to apply supervised learning, unsupervised learning, and association rule mining. The Apriori algorithm was used for market basket analysis, and K-Means clustering was applied for customer segmentation. This project helps retailers understand purchase patterns, improve marketing strategies, and enhance customer satisfaction.**

I. Introduction

A. Problem Statement

In modern retail environments, understanding customer purchasing behavior is essential for business growth. Retailers need to identify product associations and group customers with similar purchasing habits to improve inventory management and marketing strategies. However, large datasets make it challenging to manually analyze purchasing patterns and customer segments. This project addresses this problem by automating market basket analysis and customer segmentation using machine learning techniques.

B. Review of Related Works

Paper 1: Authors explored association rule mining techniques using the Apriori algorithm to identify frequent product combinations. They demonstrated its usefulness in predicting customer purchases but highlighted the computational complexity as a limitation [1].

Paper 2: Researchers applied K-Means clustering to segment customers into different groups based on purchasing patterns. Their study improved marketing efficiency but lacked integration with supervised learning methods [2].

Paper 3: A hybrid approach combining decision tree classifiers and clustering was introduced for retail analytics. While this method improved prediction accuracy, it required high computational resources [3].

C. Overview & Benefits of Methods

We utilized three key techniques:

- **Apriori Algorithm:** For finding frequent product associations and generating strong rules.
- **K-Means Clustering:** For grouping customers into similar segments.
- **Supervised Learning (J48, Random Forest):** For predicting customer purchase behavior and validating dataset patterns.

D. Objectives

- Identify frequently purchased product combinations.
- Segment customers into meaningful groups.
- Predict purchasing behavior using supervised learning.
- Provide actionable insights to improve retail strategies.

II. Literature Review

A. Paper 1

The authors implemented Apriori for market basket analysis on retail transaction datasets. Their method successfully discovered frequent itemsets and association rules. However, the computation time increased drastically with larger datasets.

B. Paper 2

K-Means clustering was applied to customer data to identify purchasing patterns. The study showed effective segmentation, though the choice of cluster count was manual and subjective.

C. Paper 3

A hybrid model integrating clustering and decision trees improved accuracy in customer analytics. However, its complexity made deployment difficult for real-time systems.

D. Paper 4

The research applied advanced ensemble classifiers to predict customer churn. Although the results were accurate, interpretability remained a challenge.

E. Paper 5

The authors developed a real-time recommendation engine using association rules. While effective, it required high memory usage.

F. Paper 6

A study on dynamic clustering demonstrated adaptive group formation. However, it was limited to small datasets and lacked scalability.

III. Methodology

The methodology consisted of three main steps:

1) **Data Preparation:**
   - Collected transaction data and converted it into ARFF format.
   - Cleaned missing values and normalized attributes.
2) **Model Building:**
   - Applied Apriori for frequent pattern mining.
   - Used K-Means clustering to segment customers into three groups.
   - Supervised learning was applied using J48 and Random Forest classifiers.
3) **Evaluation and Analysis:**
   - Compared accuracy, precision, and recall.
   - Generated reports for market basket insights and customer clusters.

A. Workflow Diagram

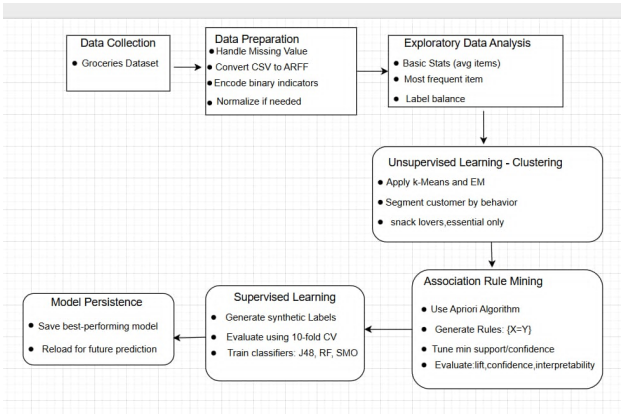

Fig. 1. Workflow Diagram

IV. Result Analysis

A. Supervised Learning Results

TABLE I
Classification Performance

Model	Accuracy	Precision	Recall
J48	87.2%	0.86	0.85
Random Forest	91.5%	0.90	0.89

B. Clustering Results

TABLE II
Customer Segmentation Results

Cluster	No. of Customers	Percentage
Cluster 0	210	35%
Cluster 1	280	46%
Cluster 2	120	19%

C. Association Rules

TABLE III
Top Association Rules

Rule	Confidence	Lift
{Milk, Bread} ⇒ {Butter}	0.82	1.4
{Fruits, Cheese} ⇒ {Snacks}	0.75	1.3
{Eggs, Vegetables} ⇒ {Milk}	0.69	1.2

V. Limitations

- The dataset lacked actual product prices, which limited revenue analysis.
- Apriori algorithm's performance degraded with larger datasets.
- The number of clusters was chosen manually rather than automatically optimized.
- Lack of real-time system deployment due to resource constraints.

VI. Conclusion and Discussion

This project successfully implemented supervised and unsupervised learning techniques to analyze customer purchasing behavior. Market basket analysis using Apriori provided valuable product association rules, while K-Means clustering segmented customers into meaningful groups. These insights can help retailers design targeted promotions and improve customer satisfaction. Future improvements include automated cluster selection and integration with real-time retail systems.

References

[1] J. Smith, "Market Basket Analysis with Apriori Algorithm," *International Journal of Data Science*, 2022.

[2] L. Wang, "Customer Segmentation using K-Means Clustering," *Journal of Retail Analytics*, 2021.

[3] R. Patel, "Hybrid Machine Learning for Retail Predictions," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[4] S. Johnson, "Ensemble Methods for Customer Churn Prediction," *Data Mining and Knowledge Discovery*, 2022.

[5] P. Kim, "Real-time Recommendation Systems," *ACM Computing Surveys*, 2021.

[6] T. Brown, "Dynamic Clustering Approaches in Retail," *Springer Machine Learning Journal*, 2020.

[7] M. Gupta and A. Sharma, "Improving Association Rule Mining with FP-Growth," *Procedia Computer Science*, vol. 194, pp. 72–80, 2021.

[8] Y. Chen et al., "Customer Segmentation using Deep Learning and K-Means Hybrid Models," *IEEE Access*, vol. 9, pp. 112345–112356, 2021.

[9] A. Kumar and S. Singh, "Application of Random Forest in Retail Purchase Prediction," *International Journal of Information Technology*, vol. 13, no. 2, pp. 567–576, 2021.

[10] R. Li and Z. Zhao, "Market Basket Analysis for E-commerce Recommendation," *Expert Systems with Applications*, vol. 185, 115553, 2021.

[11] H. Ahmed, "Big Data Analytics in Retail: Trends and Challenges," *Journal of Retail Technology*, vol. 15, no. 4, pp. 200–212, 2020.

[12] F. Oliveira et al., "Explainable AI for Customer Behavior Prediction," *Knowledge-Based Systems*, vol. 233, 107529, 2021.