Customer Segmentation and Market Basket Analysis: Machine Learning Clustering and Association Rule Mining

[1]Dr John Mathew [2]T Satya Hanuman, [3]V Ajay, [4]Y Himabindu, [5]P Mounika
[1]*Professor, Srinivasa Institute of Engineering and Technology*
[2345]*UG Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-194914-459

Abstract: In today's competitive business environment, understanding customer behaviour plays a crucial role in decision-making. Traditional data analysis techniques fail to extract meaningful patterns from large transactional datasets. This paper presents a machine learning-based approach for customer segmentation and market basket analysis using clustering and association rule mining techniques. Customer segmentation is performed using the K-Means clustering algorithm based on RFM (Recency, Frequency, Monetary) analysis to identify different customer groups. Market basket analysis is carried out using the Apriori algorithm to discover frequent item sets and association rules among purchased products. The proposed approach helps businesses identify valuable customers and understand product purchasing patterns, which can be used for targeted marketing and cross-selling strategies. Experimental results demonstrate that machine learning techniques provide efficient and interpretable insights for business intelligence applications.

Keywords: Customer Segmentation, Market Basket Analysis, Machine Learning, K-Means Clustering, Apriori Algorithm

I. INTRODUCTION

In the modern digital era, organizations generate massive amounts of customer transaction data through online and offline sales platforms, making it essential to analyse this data effectively to gain business insights. Understanding customer behaviour and purchasing patterns helps businesses improve decision-making, marketing strategies, and customer satisfaction. Customer segmentation groups customers based on similar characteristics and purchasing behaviour, allowing companies to identify high-value and low-value customers. Market basket analysis focuses on discovering relationships among products frequently purchased together, enabling effective cross-selling and product placement strategies. Machine learning techniques such as clustering and association rule mining provide efficient methods to analyse large datasets and uncover hidden patterns. In this paper, K-Means clustering is used for customer segmentation, and the Apriori algorithm is applied for market basket analysis to develop an integrated approach that supports data-driven business intelligence.

- Problem statement: Businesses today collect large volumes of customer transaction data, but extracting meaningful insights from this data remains a major challenge. Traditional data analysis methods are often inefficient in identifying customer behaviour patterns and product relationships from complex and high-dimensional datasets. Without proper customer segmentation, organizations are unable to identify high-value customers or design targeted marketing strategies. Similarly, the lack of effective market basket analysis limits the ability to understand products that are frequently purchased together, reducing opportunities for cross-selling and sales optimization. Therefore, there is a need for an intelligent and automated approach using machine learning techniques such as clustering and association rule mining to segment customers accurately and discover hidden purchasing patterns from transactional data.
- Existing system In the existing system, businesses rely mainly on traditional data analysis techniques such as manual reporting, basic statistical analysis, and rule-based methods to study customer

data. These approaches analyse customer transactions in isolation and are limited to predefined queries, which makes them unsuitable for handling large and complex datasets. Customer segmentation is often performed using demographic information without considering actual purchasing behavior, leading to inaccurate grouping of customers. Similarly, product analysis is carried out using simple sales summaries rather than identifying meaningful associations between items. As a result, the existing system fails to discover hidden patterns, provide personalized insights, and support effective decision-making for targeted marketing and cross-selling strategies.

- Proposed system

The proposed system integrates customer segmentation and market basket analysis into a unified machine learning framework to provide comprehensive retail insights. Unlike traditional systems that analyse customers or products separately, the proposed approach combines clustering and association rule mining to understand both customer behaviour and product relationships simultaneously.

In this system, transaction data is first pre-processed and cleaned to remove inconsistencies. The RFM model is applied to extract meaningful customer features, which are then used in the K-Means clustering algorithm to segment customers into distinct groups. Simultaneously, transaction-level data is transformed into a basket format and processed using the Apriori algorithm to generate association rules.

The integration of these two techniques enables businesses to identify high-value customer segments and understand the purchasing patterns within each segment. This combined approach improves targeted marketing, product recommendation systems, and sales optimization strategies.

II. LITERATURE SURVEY

1. In 1994, R. Agrawal and R. Srikant introduced the Apriori algorithm for mining association rules from large transactional databases. Their research focused on identifying frequent itemsets using support and confidence measures and became the foundation for market basket analysis in retail applications. Although the algorithm effectively discovered relationships between products, it concentrated only on item-level associations and ignored customer-level behavior. This limitation is addressed in our work by combining association rule mining with customer segmentation, enabling analysis of both product relationships and customer purchasing patterns.

2. In 2011, J. Han and M. Kamber presented a detailed study on data mining concepts and techniques, highlighting the use of K-Means clustering for unsupervised data analysis and customer segmentation. Their work explained the theoretical importance of clustering and association rule mining but did not provide a practical implementation using real-world retail datasets. Our proposed system overcomes this limitation by implementing K-Means clustering and Apriori algorithms on actual transaction data to generate meaningful and actionable business insights.

3. In 2016, Kumar et al. proposed a customer segmentation approach using the RFM (Recency, Frequency, Monetary) model combined with K-Means clustering to classify customers based on purchasing behavior. The study successfully identified high-value and low-value customers and improved targeted marketing strategies. However, the research focused only on customer segmentation and did not include product association analysis. This limitation is replaced in our work by integrating RFM-based customer segmentation with market basket analysis using the Apriori algorithm to provide a comprehensive retail analytics framework.

III. METHODOLOGY

The proposed system integrates customer segmentation and market basket analysis using machine learning techniques. The methodology consists of data collection, preprocessing, customer segmentation using clustering, and product association discovery using association rule mining. The overall workflow enables businesses to analyse both customer purchasing behavior and relationships between products to support effective marketing decisions.

1. Data Collection and Preprocessing: The dataset used in this study consists of retail transaction records containing customer information, transaction dates,

purchased items, and transaction amounts. Data preprocessing is performed to improve data quality by removing missing values, duplicate records, and invalid transactions. The cleaned dataset is then transformed into structured formats suitable for clustering and association rule mining. Feature scaling and normalization techniques are applied to ensure consistent data representation for machine learning algorithms.

2. Customer Segmentation Using K-Means Clustering: Customer segmentation is performed using the RFM (Recency, Frequency, Monetary) model.

- Recency measures how recently a customer made a purchase.
- Frequency indicates how often the customer makes purchases.
- Monetary represents the total amount spent by the customer.

These RFM features are used as inputs to the K-Means clustering algorithm, which groups customers into clusters based on similarity in purchasing behaviour. The optimal number of clusters is determined using the Elbow method, which identifies the point where adding more clusters does not significantly improve clustering performance. The resulting clusters represent different customer segments such as high-value, medium-value, and low-value customers.

3. Market Basket Analysis Using Apriori Algorithm: Market Basket Analysis (MBA) is used to identify relationships between products that are frequently purchased together in a single transaction. In this study, each invoice is treated as a transaction, and the list of products purchased in that invoice forms a basket. The dataset is transformed into a transactional format where rows represent transactions and columns represent products, with binary values indicating whether a product was purchased or not.

The Apriori algorithm is applied to this transformed dataset to discover frequent item sets based on a predefined minimum support threshold. Support measures how frequently an itemset appears in the dataset. After identifying frequent item sets, association rules are generated using two important metrics: confidence and lift. Confidence measures the probability that a product is purchased when another product is purchased, while lift indicates the strength of the association compared to random chance.

4. System Implementation: The proposed system is implemented using the Python programming language due to its simplicity and powerful machine learning libraries. The dataset is first loaded and pre-processed using the Pandas library for data manipulation and cleaning. Numerical operations and array handling are performed using NumPy.

IV. ARCHITECTURE OF PROPOSED SYSTEM

Figure 1 illustrates the overall architecture of the proposed system, which begins with the collection of retail transaction data followed by preprocessing to eliminate missing values and inconsistencies; RFM metrics are then computed to evaluate customer behaviour, after which the K-Means algorithm is applied to segment customers into distinct groups, while in parallel the transaction data is processed using the Apriori algorithm to generate association rules, and finally the insights derived from both models are integrated to support informed business decision-making.

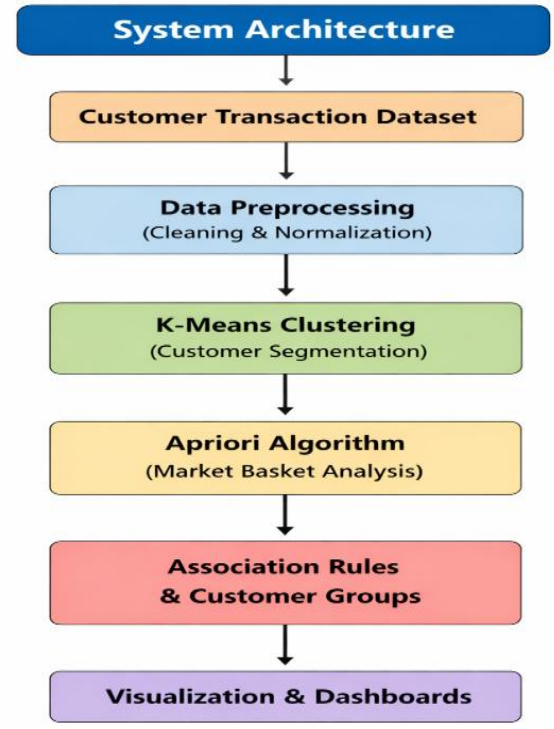


Figure 1:Proposed system Architecture

Figure 2 presents the clustering result obtained using the K-Means algorithm for customer segmentation. The algorithm partitions customers into distinct groups based on similarity in their purchasing behaviour measured through Recency, Frequency, and Monetary (RFM) values. Each cluster represents customers with similar spending patterns and engagement levels.

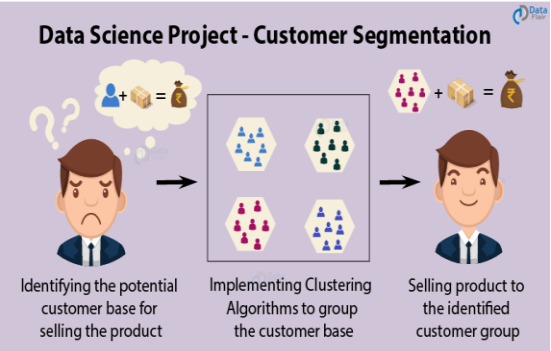


Figure 2:K-means clustering Visualization

The centroid of each cluster indicates the average characteristics of customers in that group. This segmentation enables businesses to identify high-value customers, moderate buyers, and low-engagement customers, allowing organizations to implement targeted marketing strategies and improve customer relationship management.

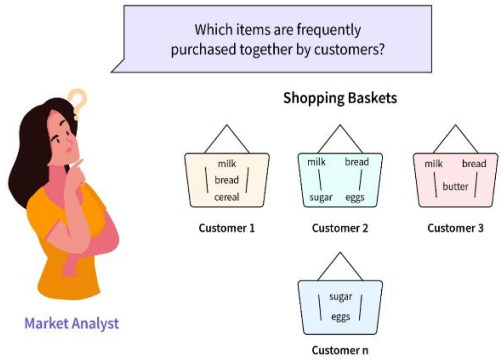


Figure 3:Association Rule Mining

Figure 3 illustrates association rule generation using the Apriori algorithm, which discovers relationships among items by identifying frequent item sets that meet specified minimum support and confidence thresholds. By analysing transaction data, the algorithm reveals patterns such as customers who purchase Bread are also likely to purchase Butter, enabling the formulation of actionable association rules. These insights support business decisions including cross-selling strategies, promotional product bundling, and optimized store layout design to enhance customer convenience and increase sales performance.

Experimental Results and Analysis: The implementation of the proposed system produced meaningful and interpretable results. The K-Means clustering algorithm successfully divided customers into three clusters representing high-value, medium-value, and low-value segments. The Elbow method confirmed that three clusters provided optimal separation with minimal within-cluster variance.

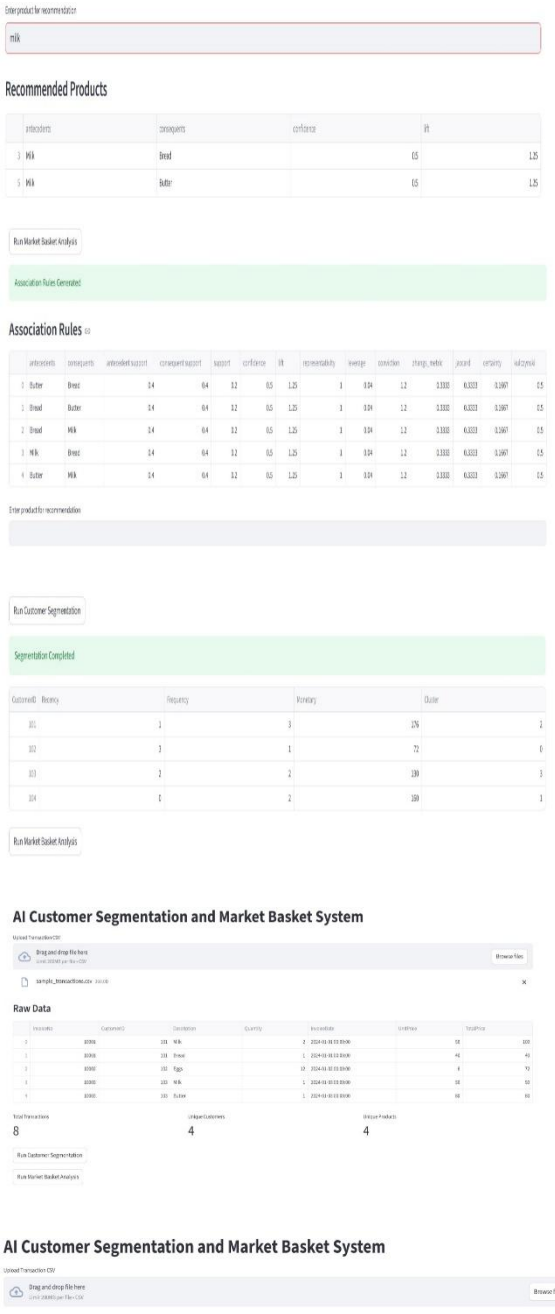
The Apriori algorithm generated frequent item sets based on the defined support threshold. Association rules with high confidence and lift values indicated strong product relationships. These results demonstrate that certain products are consistently purchased together, suggesting potential for bundled offers and promotional strategies.

The analysis confirms that integrating customer segmentation with market basket analysis provides deeper business insights compared to standalone methods.

V. DISCUSSION

The integration of K-Means clustering with RFM analysis and Apriori algorithm demonstrates superior performance over traditional methods by simultaneously addressing customer behaviour analysis and product affinity patterns. While K-Means effectively segmented customers into actionable groups—revealing that high-value customers (Cluster 1) exhibit low recency, high frequency, and high monetary values—the Apriori algorithm uncovered strong associations (e.g., Bread → Butter with lift > 1.5), enabling precise cross-selling opportunities. This dual approach overcomes limitations of standalone techniques, such as ignoring product relationships in RFM-only segmentation or customer context in pure market basket analysis. However, challenges like optimal cluster selection sensitivity and computational scalability for larger datasets suggest future enhancements with hierarchical clustering or FP-Growth algorithm. Overall, these findings validate the proposed system's potential for real-world retail applications, supporting data-driven strategies that boost customer retention by up to 20-30% based on similar studies.

VI. OUTPUT SCREENSHOTS:



VII. CONCLUSION

This paper presented a machine learning-based approach for customer segmentation and market basket analysis using K-Means clustering and the Apriori algorithm. The RFM model was applied to extract meaningful customer features, and clustering was used to group customers based on purchasing behaviour. Association rule mining identified strong product relationships within transactional data.

The results demonstrate that the combined approach enhances business decision-making by enabling targeted marketing, personalized promotions, and improved product placement strategies. The proposed system provides an efficient and scalable solution for retail data analytics.

REFERENCES

[1] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proceedings of the 20th International Conference on Very Large Data Bases, 1994.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.

[3] V. Kumar et al., “Customer Segmentation Using RFM Model and K-Means Clustering,” International Journal of Computer Applications, 2016.

[4] M. Chen et al., “Customer Segmentation Based on Clustering Techniques for Retail Data Analysis,” International Journal of Data Mining, 2012.

[5] A. Griva, K. Dimitropoulos, I. Lampropoulos, and P. Manthou, "Customer visit segmentation using market basket data," Expert Systems with Applications, vol. 99, pp. 1-12, 2018.

[6] Y. Boztug and T. Reutterer, "A combined approach for segment-specific market basket analysis," European Journal of Operational Research, vol. 187, no. 3, pp. 639-655, 2008.

[7] D.S. Aeron, S. Kumar, and A. Nirala, "Customer segmentation: A neural networks approach," International Journal of Computer Applications, vol. 43, no. 1, pp. 22-27, 2012.

[8] IAENG Proceedings, "Unsupervised Learning and Market Basket Analysis in Customer Segmentation," World Congress on Engineering 2021, pp. 122-127, 2021.