IEEE Access

Customer Behaviour Analysis and Predictive Modelling in Supermarket Retail: A Comprehensive Data Mining Approach

Dr. Kavitha Dhanushkodi [1], Akila Bala[2], Nithin Kodipyaka[3] and Shreyas V[4]

[1]Assistant Professor Sr., School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[2]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[3]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
[4]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Corresponding author: Dr. D. Kavitha (e-mail: kavitha.d@vit.ac.in).

This work was supported by Vellore Institute of Technology, Chennai, India.

ABSTRACT In the dynamic landscape of supermarket retail, understanding customer behavior is paramount for optimizing business strategies and enhancing profitability. This paper presents a comprehensive data mining approach to analyze customer behavior and build predictive models within the supermarket retail domain. Leveraging advanced data analytics techniques, our methodology encompasses data preprocessing, exploratory data analysis, feature engineering, model selection, and evaluation. This paper presents a comprehensive approach to customer behavior analysis and predictive modelling within the context of supermarket retail. We delve into the intricacies of data mining methodologies, exploring how retailers can leverage diverse datasets to uncover valuable insights and build predictive models that drive business growth and customer satisfaction. From data preprocessing to model evaluation, each step in the process is meticulously examined, highlighting best practices and key considerations for effective implementation.

INDEX TERMS Customer behavior analysis, Data mining, Predictive modelling, Retail, Sequential pattern mining.

I. INTRODUCTION

In the contemporary landscape of retail, particularly within the dynamic realm of supermarket operations, the quest to understand and predict customer behavior stands as a cornerstone of strategic decision-making. Amidst the proliferation of data sources and the advent of sophisticated analytical methodologies, retailers are compelled to adopt comprehensive data mining approaches to derive actionable insights that drive business success. This paper embarks on a journey through the intricate domain of customer behavior analysis and predictive modelling within the context of supermarket retail, elucidating the pivotal role of data-driven strategies in optimizing operational efficiencies and enhancing customer satisfaction.

The exponential growth of data availability in recent years has paved the way for transformative advancements in retail analytics. With transaction records, customer demographics, loyalty program data, and product information constituting a rich tapestry of insights, retailers are presented with an unprecedented opportunity to unravel the intricacies of consumer behavior. By harnessing advanced data mining techniques, retailers can distil these vast datasets into actionable intelligence, empowering them to make informed decisions across a spectrum of operational domains, including marketing, merchandising, and inventory management.

Against this backdrop, this paper endeavors to elucidate a systematic approach to customer behavior analysis and predictive modelling in the context of supermarket retail. Through a structured methodology encompassing data preprocessing, exploratory data analysis, feature engineering, model selection, and evaluation, retailers can unlock the latent potential of their data assets and derive meaningful insights into customer preferences, purchasing patterns, and brand affinities.

Furthermore, the integration of predictive modelling techniques not only presents an opportunity to forecast forthcoming consumer behaviors but also empowers retailers to proactively adjust their strategies. This proactive approach allows for the customization of product offerings to align with evolving consumer preferences, the tailoring of marketing campaigns to resonate with specific customer segments, and the optimization of resource allocation strategies for improved operational efficiency. Through a meticulous examination of each phase within the data mining process, coupled with insights derived from practical implementations, this paper aims to furnish retailers with the necessary expertise and tools to embrace data-driven decision-making effectively. By doing so, retailers can not only achieve sustained growth but also gain a competitive edge in the dynamic landscape of supermarket retail.

Embracing data-driven strategies as a cornerstone of their operations, retailers can navigate through the complexities of the market, capitalize on emerging opportunities, and foster long-term success in an ever-evolving retail environment.

II. RELATED WORK

Customer segmentation in retail, particularly in online retail, plays a crucial role in tailoring marketing strategies. The RFM model, focusing on Recency, Frequency, and Monetary value, is a significant framework for this purpose [1]. Data mining offers advantages in understanding customer behavior and preferences, enhancing data quality, and enabling streamlined operations in retail. However, privacy concerns must be addressed to maintain consumer trust [2]. The integration of data mining and machine learning in retail, akin to the banking industry, leads to improved customer satisfaction and retention rates through personalized services [3]. Data warehousing aids in customer segmentation and inventory management, facilitating personalized marketing and operational efficiency in retail [4]. E-commerce data mining enables retailers to understand market demands, optimize product assortment, refine pricing strategies, and ultimately drive revenue growth [5]. Data mining has become a pivotal tool in the retail sector, revolutionizing operations with its capacity to derive actionable insights from vast datasets. Its application offers numerous advantages, foremost among them being the acquisition of valuable customer insights. By analyzing purchasing patterns, preferences, and behaviors, retailers can tailor their marketing strategies and enhance customer experiences.

Customer segmentation is a critical strategy in online retail, allowing businesses to tailor marketing efforts effectively. The RFM model, which analyzes Recency, Frequency, and Monetary value, provides a comprehensive framework for this purpose [1]. Data mining offers significant advantages in retail, including understanding customer behavior, enhancing data quality, and improving operational efficiency. However, addressing privacy concerns is essential to maintaining consumer trust and compliance with regulations [2]. The integration of data mining and machine learning techniques in retail, drawing parallels from the banking sector, has shown promising results in enhancing customer satisfaction and retention rates through personalized services [3]. Data warehousing plays a vital role in retail operations, facilitating customer segmentation and inventory management. This enables retailers to optimize marketing strategies and streamline inventory levels for improved efficiency [4]. E-commerce data mining provides valuable insights into market trends, enabling retailers to make informed decisions regarding product assortment, pricing strategies, and promotional activities. This ultimately drives revenue growth and ensures competitiveness in the online retail landscape [5]. Data mining has emerged as a pivotal tool for enhancing competitive advantage in both banking and retail sectors. In banking, data mining empowers institutions to sift through vast troves of customer data encompassing transactional histories, demographics, and behavioural patterns. This enables the segmentation of customers according to their distinct needs, preferences, and risk profiles, thereby facilitating tailored services and targeted marketing strategies. Moreover, predictive analytics and anomaly detection techniques bolster risk management efforts by enabling early identification of potential threats and fraudulent activities, thus mitigating financial losses. Similarly, in the retail industry, data mining plays a crucial role in customer segmentation based on purchase behaviors, browsing habits, and demographic insights. This segmentation aids retailers in crafting personalized shopping experiences and devising targeted promotional campaigns. Furthermore, data mining serves as a potent tool in fraud detection, helping retailers combat revenue losses stemming from theft, shoplifting, or the circulation of counterfeit goods. Overall, the integration of data mining techniques equips both banking and retail sectors with valuable insights, enabling them to optimize operations, enhance customer satisfaction, and fortify against financial risks and fraudulent activities [6]. Data mining techniques have become integral to the retail industry, offering diverse applications that enhance operational efficiency and strategic decision-making. One such crucial application is customer segmentation, where clustering and classification algorithms categorize customers based on demographics, purchase behavior, and preferences. This segmentation enables retailers to tailor marketing strategies effectively, thereby optimizing customer engagement and satisfaction [7]. Market basket analysis, a subtype of association rule mining, further aids retailers in uncovering patterns among products frequently purchased together. This insight facilitates optimized product placement, promotions, and cross-selling opportunities, ultimately boosting sales and profitability [7]. Moreover, data mining techniques play a vital role in demand forecasting, employing methods such as time series analysis and regression modelling to predict future demand accurately. By leveraging historical sales data, seasonal patterns, trends, and external factors, retailers can optimize inventory management and ensure adequate stock levels to meet customer demand [7]. In the realm of e-commerce, personalized online sales leveraging web usage data mining have gained significant attention for enhancing user experience and increasing sales. This process begins with comprehensive data collection, encompassing user interactions such as clicks, views, searches, and purchases. Subsequent data preprocessing ensures data quality by removing noise, handling missing values, and transforming data into an analyzable format. User profiling, based on extracted web usage data, enables tailored recommendations and marketing strategies, fostering customer engagement and loyalty [8]. The study titled "Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining" investigates various classification algorithms' effectiveness in analyzing consumer online shopping attitudes and behavior. Following a structured approach involving data collection, preprocessing, and feature engineering, the study conducts unbiased model training and evaluation through data splitting. Rigorous optimization of classification algorithms through model training and hyperparameter tuning is performed, followed by thorough model evaluation using techniques such as cross-validation to assess generalization capabilities. The study's findings provide valuable insights into the strengths and weaknesses of each algorithm in predicting consumer behaviour, offering valuable guidance

for researchers and practitioners in online consumer behaviour analysis [9]. Srikant and Agrawal's seminal paper titled "Mining Sequential Patterns" has significantly influenced the data mining landscape, particularly in sequential data analysis. Their work offers crucial insights and methodologies for extracting meaningful patterns from sequences of data, introducing concepts like the Apriori-based algorithm for sequence pattern mining. Over the years, researchers have built upon this foundation, exploring various generalizations and performance enhancements to advance sequential pattern mining techniques. This paper remains a cornerstone in the literature, serving as a benchmark for subsequent research endeavors aimed at harnessing insights within sequential data structures [10]. Magnini, Honeycutt Jr, and Hodge explore data mining applications in the hotel industry, highlighting its role in extracting actionable insights from vast data reservoirs. They emphasize benefits like enhanced customer segmentation and demand forecasting, alongside challenges such as data quality and privacy concerns [11]. Ritbumroong investigates online analytical mining (OLAM) methodologies, combining data mining and online analytical processing (OLAP) to extract actionable insights. OLAM's integration allows businesses to understand customer preferences, anticipate trends, and tailor strategies effectively [12]. Hemalatha's study focuses on market basket analysis in Indian retailing, aiming to understand consumer purchasing behaviors and associations between products. Leveraging data mining methodologies, this research aids in targeted marketing and product placement decisions [13]. Huang, Chang, and Narayanan delve into customer behaviour in dynamic markets, using data mining techniques to comprehend and anticipate shifts in preferences. Their study highlights the importance of adapting to changing market dynamics for effective forecasting and response [14]. Punpukdee et al. conducted a comprehensive investigation into determinants of consumer behaviour in Thailand's offline marketing sphere, synthesizing insights from existing literature and employing data mining techniques to uncover patterns and connections among factors shaping consumer decisions [15].

III. DATASET

A. COLUMNS

- **BasketID**: Unique identifier for each basket or transaction.
- **BasketDate**: Date and time of the basket or transaction.
- **Sale**: Sale amount for each product in the basket.
- **CustomerID**: Unique identifier for each customer.
- **CustomerCountry**: Country of the customer.
- **ProdID**: Unique identifier for each product.
- **ProdDescr**: Description of the product.
- **Qta**: Quantity of each product in the basket.

B. DATA ENTRIES

Each row represents a product purchased within a specific basket or transaction. The dataset contains information about various transactions, including the date, customer details, product details, and quantities purchased.

C. DATA TYPES

- **BasketID**: numeric identifier.
- **BasketDate**: Date and time format.
- **Sale**: Numeric
- **CustomerID**: Numeric
- **CustomerCountry**: Categorical, representing the country name.
- **ProdID**: alphanumeric identifier.
- **ProdDescr**: Text description.
- **Qta**: Numeric, representing quantity.

D. DATA QUALITY

To assess the quality of the data, the process began by eliminating duplicate entries, amounting to 5232 instances, which represented approximately 1.11% of the entire dataset. This left the analysis with 466,678 rows for further examination. Upon visual inspection of the numerical attributes, it became apparent from the plots in Figure 1a that both attributes exhibited notably high outliers, both positive and negative. However, the presence of negative values was inconsistent with the semantics of the attributes, as they should inherently be positive. From Figure 1 and Figure 2, further investigation revealed that negative values in the "Qta" attribute likely represented refunds, supported by the symmetric behaviour of the attribute. Additionally, nearly all records with negative "Qta" values were associated with BasketIDs starting with "C," indicative of cancellations. Notably, some records with negative "Qta" values lacked a corresponding "C" BasketID prefix, yet analysis of their respective "ProdDescr" suggested they pertained to errors or damaged items. Regarding negative "Sale" values, only two records exhibited this property, which upon examination of the "ProdDescr" ("ADJUST BAD DEBT") were attributed to errors. Importantly, all rows identified as errors were associated with null CustomerIDs. Subsequently, the analysis proceeded by removing entries corresponding to the 65,073 null CustomerID values, constituting approximately 13.94% of the dataset. This action was deemed necessary as the primary objective was to analyze customer behaviour, rendering entries with null CustomerIDs irrelevant. This removal process also eliminated the previously identified errors. Additionally, ProdIDs that did not conform to the defined format, consisting solely of letters such as 'POST', 'D', 'C2', 'M', 'BANK CHARGES', etc., accompanied by respective ProdDescrs such as 'POSTAGE', 'Discount', 'CARRIAGE', 'Manual', 'Bank Charges', etc., were eliminated from the dataset. Consequently, 1273 entries were dropped from the dataset.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3407151

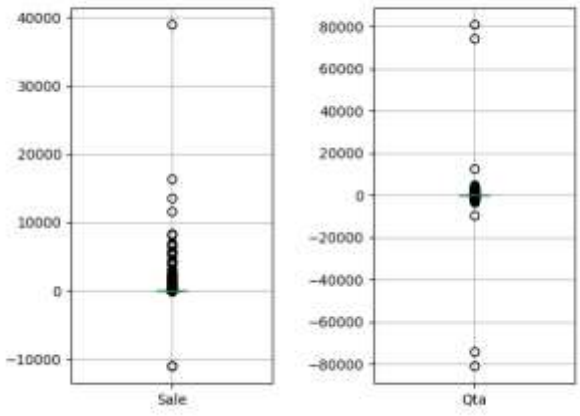

FIGURE 1. Boxplot before Outlier Removal.

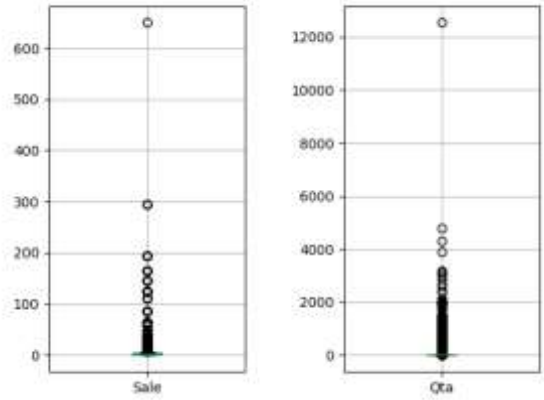

FIGURE 2. Boxplot after Outlier Removal.

E. VARIABLE DISTRIBUTION

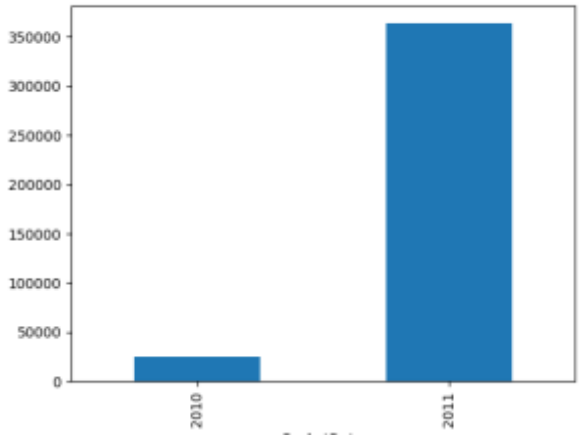

FIGURE 3. Yearly Variable Distribution.

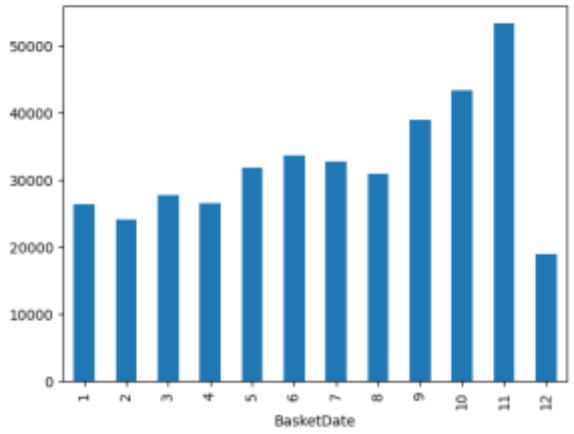

FIGURE 4. Monthly Variable Distribution.

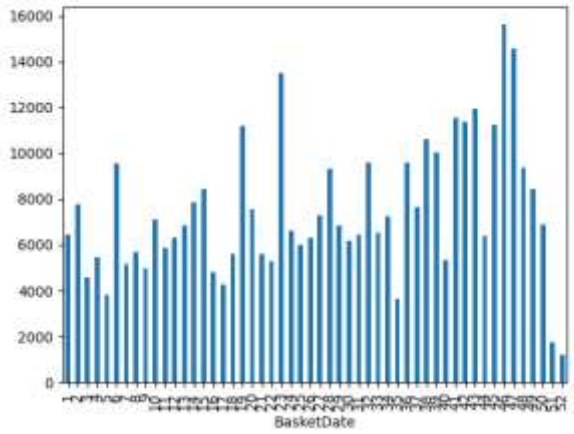

FIGURE 5. Weekly Variable Distribution.

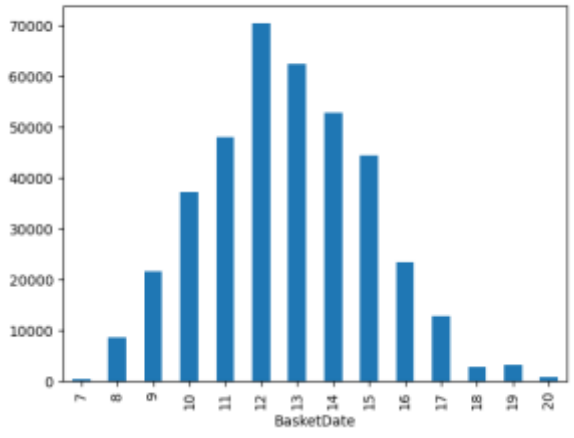

FIGURE 6. Daily Variable Distribution.

From Figure 3, we can see that the attribute is highly unbalanced; in fact, we have that almost all the records are related to transactions of 2011, while the objects from 2010 are very few. Indeed, the rows of 2011 represents about 93% of the whole dataset.

Furthermore, Figure 4, represents an estimation of the probability density function of BasketDate
divided by year, we can appreciate the two different distributions.
In fact, for 2010, we have a very uneven plot, which indicates that the records are not uniformly distributed
concerning the days in a month as shown in Figure 5. This is because, for the majority of the months in 2010, there were registered only transactions from a single day from Figure 6, the 12th; this is the value for which the plot shows the peak. On the other
hand, the distribution for 2011 is much more homogeneous, meaning that the transactions were registered
for most days in the months of that year.

F. SUMMARY

The dataset is transactional sales data for a retail business, l for an online store.
Each transaction consists of multiple products.
It includes details such as the date of purchase, customer information, product details, and quantities sold.
The dataset covers sales across different countries, as indicated by the "CustomerCountry" column.

IV. EXPLORATORY DATA ANALYSIS

A. DATASET CLEANING

1) REMOVAL OF NULL CUSTOMERID ENTRIES
Identify rows in the dataset where the CustomerID variable is null.
Remove these rows from the dataset as they lack essential customer identification information.

2) IDENTIFICATION OF PECULIAR TRANSACTIONS
Examine the ProdID variable to identify transactions characterized by codes consisting solely of letters.
Investigate these transactions to understand their nature, such as port charges, bank charges, or discounts.

3: IDENTIFICATION OF CANCELED TRANSACTIONS
Check for entries in the BasketID variable prefixed with "C," indicating cancelled transactions.
Ensure that for each cancelled transaction, there exists a corresponding entry with a negative quantity (Qta) representing the cancellation.

B. FEATURE EXTRACTION

In the project, RFM (Recency, Frequency, Monetary) analysis is employed as a fundamental technique for customer segmentation and understanding consumer behaviour in the context of online shopping [4]. At its core, RFM analysis operates on the premise of the Pareto Principle, often referred to as the "80-20 rule," which suggests that approximately 80% of a company's revenue is generated by around 20% of its customers. By segmenting customers according to RFM criteria, businesses can identify this vital subset of customers who contribute significantly to their bottom line. The first component of RFM analysis is recency, which measures how recently a customer has made a purchase or interacted with the business as shown in Table I. Customers who have engaged with the business more recently are often considered more valuable, as they may be more likely to make additional purchases in the near future. Frequency refers to the number of transactions or interactions a customer has had with the business within a specific period. Customers who make frequent purchases or engage with the business regularly are typically seen as more loyal and valuable.

Monetary value assesses the total amount of money a customer has spent on purchases or transactions with the business. Customers who have spent more money over time are generally considered high-value customers, as they contribute more to the company's revenue. RFM analysis enables the categorization of customers based on their transactional history, focusing on three key metrics:

- **Recency Feature:** This metric measures the time elapsed since a customer's last purchase. Customers who have made recent purchases are likely to be more engaged with the business and may exhibit different behaviour compared to those who have not made purchases in a while.
  Calculate the number of days between the present date and the date of the customer's last purchase.
  This feature quantifies how recently a customer made a purchase, indicating their engagement level with the business.
  Example: *recency = present_date - last_purchase_date*

- **Frequency Feature:** Frequency refers to the number of transactions made by each customer within a given period. Customers who make frequent purchases may represent loyal or high-value segments, while those with lower frequencies may require targeted marketing efforts to increase engagement.
  Count the number of orders or transactions for each customer.
  This feature represents how frequently a customer makes purchases, indicating their loyalty or engagement with the business.
  Example: *frequency = total_number_of_orders*

- **Monetary Feature:** The monetary metric represents the total amount spent by each customer over a specified timeframe. This metric helps identify high-spending customers who contribute significantly to revenue generation and can inform strategies for personalized marketing and promotions. Calculate the total purchase amount or spending for each customer.
  This feature represents the monetary value of the customer's transactions, indicating their profitability or contribution to the business.
  Example: *monetary = sum_of_purchase_amounts*

TABLE I. RFM Quartile Table.

	R	F	M
0.25	22.0	1.0	298.11
0.50	58.0	2.0	646.30
0.75	151.0	4.0	1567.20

CustomerID	R	F	M	R_quartile	F_quartile	M_quartile	RFM
12347.0	40	6	3598.21	2	1	1	211
12348.0	76	3	784.44	3	2	2	322
12349.0	19	1	1457.55	1	4	2	142
12350.0	311	1	294.40	4	4	4	444
12352.0	73	6	1265.41	3	1	2	312
...	...	...	...	...	...	...	...
18280.0	160	1	180.60	4	4	4	444
18281.0	4	1	80.82	1	4	4	144
18282.0	216	2	176.60	4	3	4	434
18283.0	10	16	2039.58	1	1	1	111
18287.0	0	3	1837.28	1	2	1	121

4205 rows × 7 columns

FIGURE 7. RFM Segmentation Table.

Customer segmentation is a crucial strategy for businesses to effectively target their marketing efforts and maximize profitability. One popular method for segmenting customers is RFM analysis, which evaluates customers based on Recency, Frequency, and Monetary value. By dividing customers into segments based on their RFM scores, businesses can identify their most valuable customers and implement targeted marketing campaigns. In this dataset, we have segmented customers into six categories: Best Customers, Loyal Customers, Big Spenders, Almost Lost, Lost Customers, and Lost Cheap Customers. In Figure 7 each segment represents a different level of engagement and potential value to the business, allowing for tailored marketing strategies to be implemented.

- Best Customers: These are customers who have the best scores in all three RFM categories (Recency, Frequency, and Monetary). Specifically, they have the best Recency score (1), the best Frequency score (1), and the best Monetary score (1). In your dataset, there are 440 customers in this segment.
- Loyal Customers: Loyal customers may not have the absolute best scores in all three categories, but they still demonstrate strong loyalty to the business. In your dataset, there are 1012 customers classified as Loyal Customers.
- Big Spenders: These customers may not purchase as frequently as the Best Customers or Loyal Customers, but they spend a significant amount when they do make a purchase. They have the best Monetary score (1) but may not necessarily have the best scores in the other categories. There are 1051 customers in this segment in your dataset.
- Almost Lost: These customers may have shown signs of decreased activity or engagement with the business. They may have previously been more active but have become less so recently. In your dataset, there are 105 customers classified as Almost Lost.
- Lost Customers: These customers have not made purchases in a long time and are considered lost to the business. They have the worst scores in the Recency category. In your dataset, there are 11 customers classified as Lost Customers.
- Lost Cheap Customers: These customers have not made purchases in a long time, like Lost Customers, but when they did make purchases, they typically spent less than other segments. They have low scores in both the Recency and Monetary categories. In your dataset, there are 431 customers classified as Lost Cheap Customers.

Feature correlation analysis is essential to understand the relationships between different features in a dataset. In the context of RFM analysis for customer segmentation, we can perform correlation analysis to examine how the Recency, Frequency, and Monetary features are related to each other. By assessing the correlation in Figure 8 between these three key dimensions of customer behaviour, we can gain insights into patterns such as whether customers who make purchases more frequently tend to spend more money per transaction or if customers who have made recent purchases are more likely to make repeat purchases. Moreover, understanding feature correlations helps in identifying redundant or highly correlated features, which can be beneficial for feature selection and model simplification.

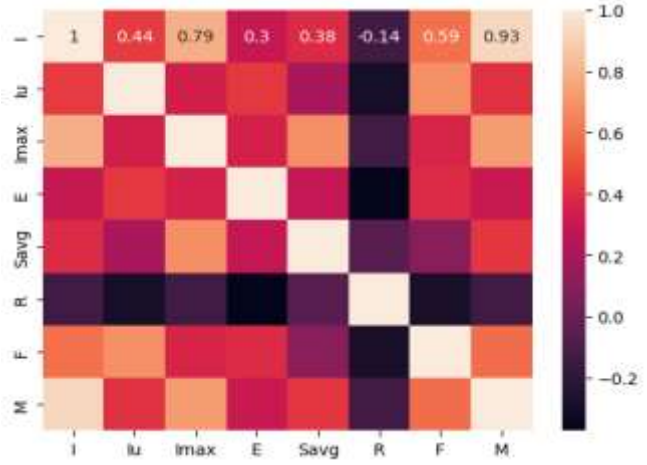

FIGURE 8. Feature Correlation Heatmap.

V. SYSTEM ARCHITECTURE

Implementation Description for Implementation Diagram Mapping according to Figure 10:

A. DATA UNDERSTANDING AND PREPARATION

- **Data Semantics Analysis**: Initially, explore the dataset using pandas to understand its structure, features, and data types.
- **Distribution and Statistics Analysis**: Utilize descriptive statistics to understand the distribution of variables such as sales, customer IDs, product IDs, etc.

- **Data Quality Assessment**: Use techniques such as checking for missing values, outliers, and inconsistencies in the dataset to ensure data quality.
- **Variables Transformation and Generation**: Perform necessary transformations on variables, such as converting data types, encoding categorical variables, and generating new features as required by the task.
- **Pairwise Correlations Analysis**: Calculate pairwise correlations between variables to identify relationships and potential redundancies. Eliminate redundant variables if necessary.

B. CLUSTERING ANALYSIS

- **K-means Clustering:** Identify the optimal value of k using techniques such as the elbow method or silhouette score as depicted in Figure 9.
- **Density-based Clustering:** Study clustering parameters such as minimum samples and epsilon for DBSCAN.
  Characterize and interpret clusters obtained from DBSCAN.
- **Hierarchical Clustering:** Compare different hierarchical clustering results using different linkage methods (e.g., single, complete, average).
  Visualize and analyze dendrograms to understand cluster hierarchy and structure.
- **Alternative Clustering Techniques***:* Explore additional clustering techniques provided by the clustering library, such as agglomerative clustering or G-means clustering.

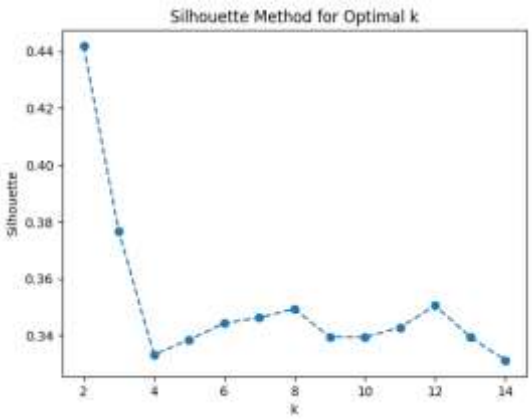

FIGURE 9. Silhouette Method for Optimal k.

C. PREDICTIVE ANALYSIS

- **Customer Profile Definition***:* Define a customer profile based on indicators computed during data preparation, considering attributes like the total number of items purchased, distinct items bought, maximum items purchased per session, and Shannon entropy of purchasing behaviour.
- **Label Generation***:* Compute labels for customers indicating high-spending, medium-spending, or low-spending categories based on defined thresholds.
- **Model Selection and Evaluation***:* Evaluate the performance of different predictive models (e.g., logistic regression, decision trees, random forests) on training and test sets. Discuss preprocessing techniques applied to manage potential challenges in prediction.

D. SEQUENTIAL PATTERN MINING

- **Modelling Customer Sequences**: Model customers as sequences of baskets, where each basket contains purchased items.
- **Sequential Pattern Mining Algorithm**: Implement the Generalized Sequential Pattern (GSP) algorithm to mine frequent sequential patterns in customer basket sequences as shown in Figure 11.
- **Pattern Analysis and Interpretation**: Discuss resulting sequential patterns, considering their support, length, and interpretation in the context of customer purchasing behaviour.

These implementation steps outline the process flow for analyzing customer behaviour, performing clustering analysis, predictive modelling, and sequential pattern mining based on the provided project description and sequential pattern mining code. Each step contributes to understanding customer preferences, segmenting customers, and making predictions to enhance supermarket retail strategies [7].

FIGURE 10. Architectural Diagram of the Proposed Process.

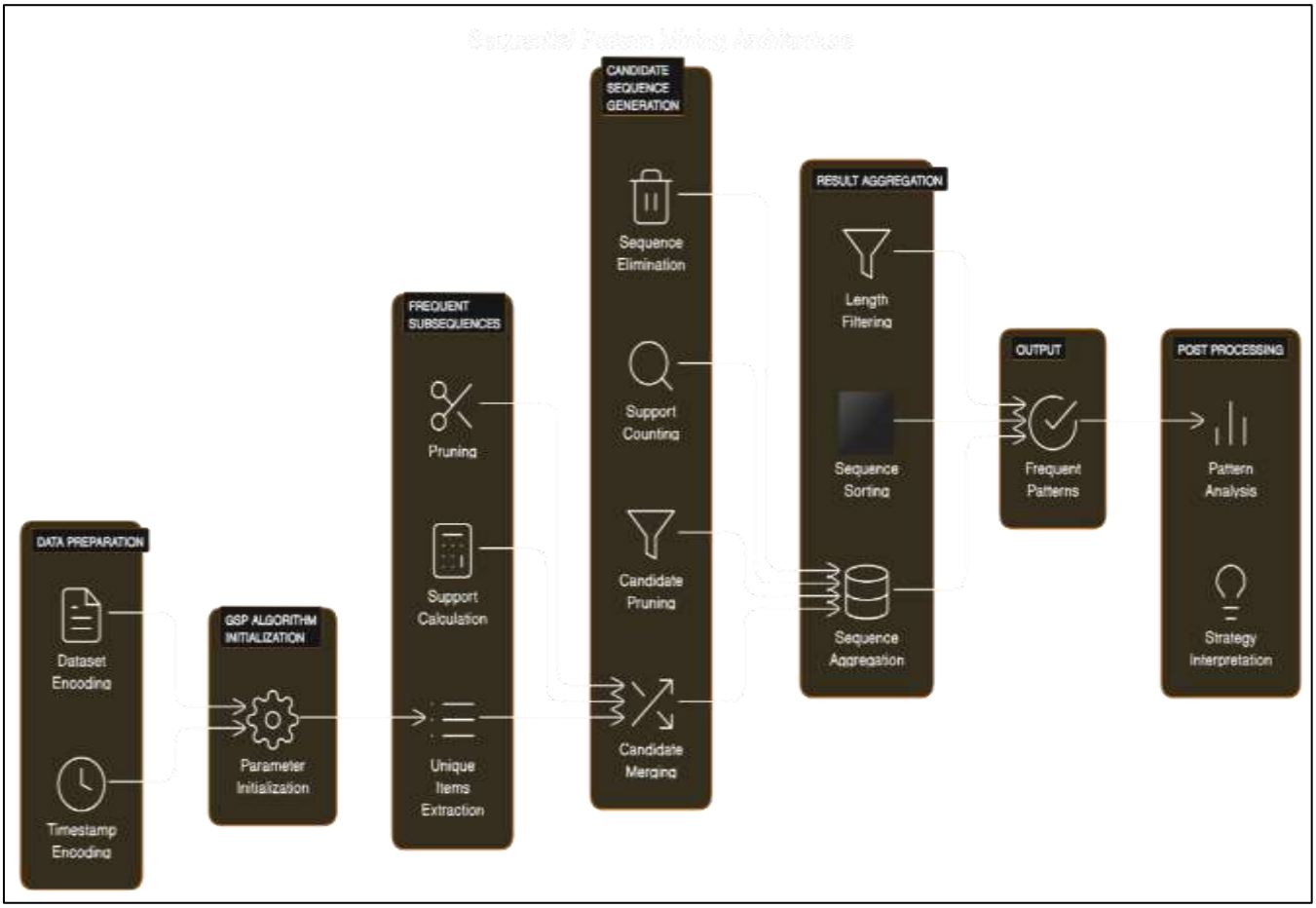

FIGURE 11. **Sequential Diagram of the Proposed Process.**

VI. MODULAR DECOMPOSITION

Given below in Figure 12 is the modular decomposition for the proposed technology

A. OBJECTIVES

- Analyze customer behaviour in supermarket retail.
- Develop predictive models for customer behaviour.
- Utilize data mining techniques for comprehensive analysis.

B. COMPONENTS

- Data Collection
- Data Preprocessing
- Customer Segmentation
- Predictive Modeling
- Algorithm Selection
- Model Training
- Model Evaluation
- Feature Importance Analysis
- Result Interpretation
- Conclusion

C. TASKS

- Gather supermarket retail data including sales, transactions, and customer demographics.
- Clean, preprocess, and integrate the collected data for analysis.
- Segment customers based on their shopping behaviour, preferences, and demographics.
- Develop predictive models to forecast customer behaviour, such as purchase propensity or churn.
- Select appropriate data mining algorithms for predictive modelling.
- Train the predictive models using the prepared data.
- Evaluate the performance of the trained models using relevant metrics.
- Analyze the importance of features in predicting customer behaviour.
- Interpret the results of the analysis and conclude.
- Provide recommendations based on the findings to improve supermarket retail strategies.

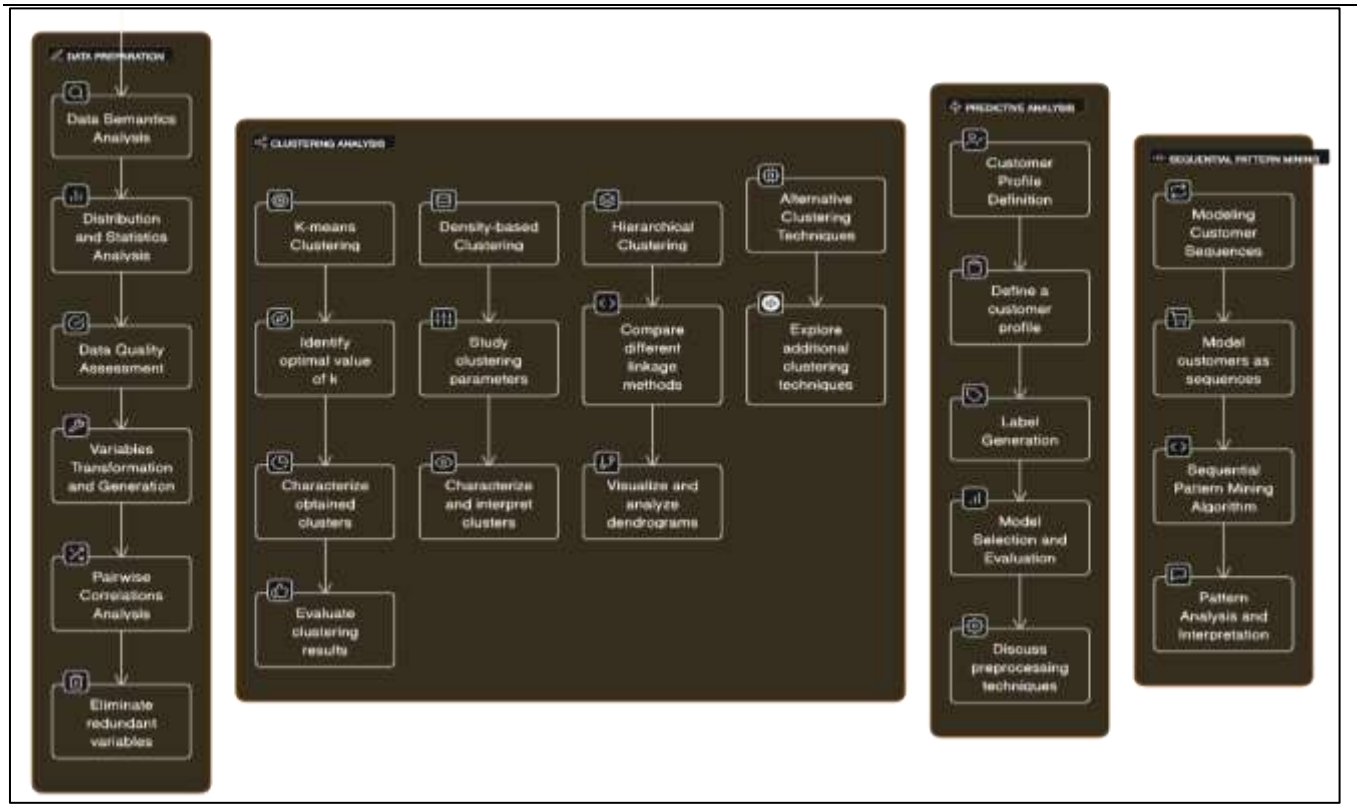

FIGURE 12. Flowchart Diagram of the Proposed Architecture.

Each component serves a specific purpose, allowing for a structured and organized workflow. For instance, data collection involves gathering diverse datasets encompassing sales, transactions, and customer demographics, while preprocessing focuses on cleaning and integrating these datasets for analysis. Customer segmentation aims to categorize customers based on various parameters, facilitating targeted marketing strategies. Predictive modelling utilizes advanced algorithms to forecast customer behaviour, such as purchase propensity or churn. Algorithm selection ensures the use of appropriate techniques tailored to the specific objectives.

Model training and evaluation assess the performance of predictive models using relevant metrics, while feature importance analysis highlights key factors influencing customer behaviour predictions. Finally, the result interpretation and conclusion provide insights derived from the analysis, offering recommendations to enhance supermarket retail strategies.

The structured workflow for analyzing supermarket retail strategies begins with data collection, where diverse datasets encompassing sales transactions, customer demographics, and product information are gathered. This comprehensive dataset ensures a holistic view of the retail environment, facilitating more accurate analysis and insights. Subsequently, data preprocessing focuses on cleaning and integrating the collected datasets to ensure data quality and consistency. Tasks such as handling missing values, removing duplicates, and standardizing formats prepare the data for analysis and modelling. Customer segmentation follows, aiming to categorize customers based on parameters such as purchase history, demographics, and behaviour patterns. By segmenting customers into distinct groups, retailers can tailor marketing strategies, promotions, and product offerings to meet the specific needs and preferences of each segment. Predictive modelling utilizes advanced algorithms to forecast customer behaviour, such as purchase propensity, product preferences, or churn likelihood. Model training and evaluation assess the performance of predictive models using relevant metrics, ensuring the reliability and effectiveness of the algorithms in making accurate predictions and capturing underlying patterns in the data. Feature importance analysis examines the significance of different factors or variables in influencing customer behaviour predictions, allowing retailers to prioritize resources and efforts on the most impactful strategies. Finally, the result interpretation and conclusion synthesize the findings of the analysis, providing actionable insights for improving supermarket retail strategies. By following this structured workflow, retailers can leverage data-driven insights to optimize operations, enhance customer experiences, and drive business growth in the competitive retail landscape.

The structured workflow for analyzing supermarket retail strategies is a systematic approach designed to extract actionable insights from data to enhance business decision-making. It begins with meticulous data collection, encompassing a wide range of datasets including sales transactions, customer demographics, product information, and other relevant sources. This comprehensive dataset serves as the foundation for subsequent analysis, providing a holistic view of the retail landscape.

VII. PREDICTIVE ANALYSIS

For this, we will be using Neural Networks and SVC

- **Data Preparation**: Prepare the dataset with RFM features (Recency, Frequency, Monetary) as input variables and customer segments as the target variable.
  Encode categorical target variables (customer segments) if necessary.
- **Splitting Data**: Split the dataset into training and testing sets for model training and evaluation, typically using a 70-30 or 80-20 split [9][10].
- **Model Training**: Train an SVC classifier using the training data.
  Train a Neural Network classifier using the training data.
  Experiment with different hyperparameters and architectures to optimize model performance.
  To train an SVC classifier and a Neural Network classifier using the training data, we'll use Python's scikit-learn library for SVC and TensorFlow/Keras for the Neural Network. We will also experiment with different hyperparameters and architectures to optimize model performance.
  This approach allows us to experiment with different hyperparameters and architectures to optimize the performance of both classifiers. Adjustments can be made to the hyperparameter and architecture search spaces based on the specific requirements of the problem and the characteristics of the dataset.

- **Model Evaluation**:

  For SVC:

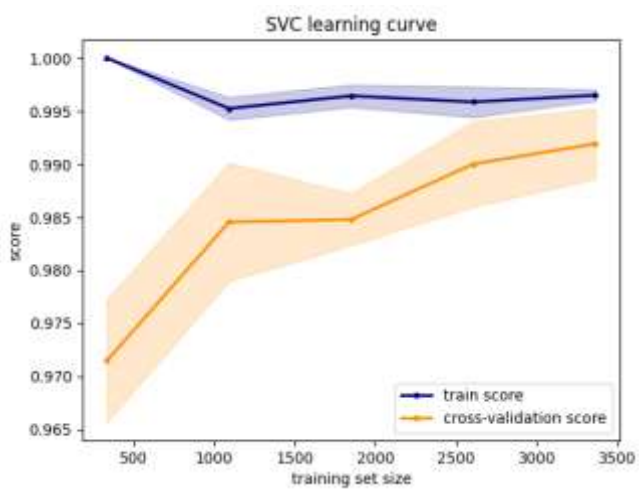

FIGURE 13. **SVC Learning Curve.**

The learning curve in Figure 13 depicts the relationship between the model's performance (on the y-axis) and the size of the training set (on the x-axis). In this specific case, the y-axis represents the average score, which could be accuracy, precision, recall, or F1 score.

Here are some observations based on the graph:

Overall Performance: The average score appears to be relatively high across the entire training set size range, indicating that the SVC model is performing well.
Training vs. Cross-Validation: The two curves in the graph represent the training score (solid line) and the cross-validation score (dashed line). The training score shows a slight upward trend as the size of the training set increases, which is expected as the model learns from more data. The cross-validation score seems to fluctuate slightly but generally stays around 0.97, suggesting that the model is not overfitting the training data.
Limited Data: The x-axis only goes up to 3500, which might be a relatively small dataset size for training complex models like SVMs.
Overall, the graph suggests that the SVC model is likely performing well on this specific dataset. However, it's important to note that the generalizability of this model to unseen data cannot be definitively determined from this graph alone.

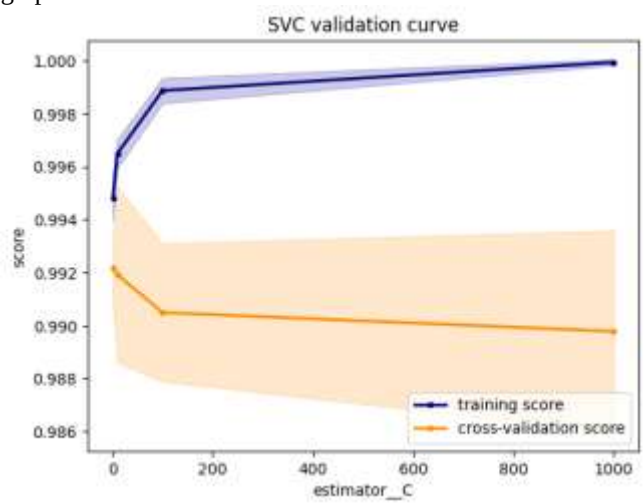

FIGURE 14. **SVC Validation Curve.**

Figure 14 features a detailed graph labelled "SVC validation curve," showcasing the performance of a model in terms of its training and cross-validation scores. The vertical axis of the graph is marked with scores ranging from 0.986 to 1.000, indicating the accuracy or effectiveness of the model under evaluation. Horizontally, the graph extends from 0 to 1000, possibly representing the number of estimators or a regularization parameter denoted as "C" that influences the model's complexity.

The graph in Figure 15 and Figure 16 displays two distinct lines, one representing the "training score" and the other the "cross-validation score." These lines seem to illustrate how the model's performance varies with changes in the parameter C, with both lines starting high and showing

slight variations as C increases. The presence of these lines suggests a careful analysis of the model's ability to generalize from training data to unseen data, a crucial aspect of machine learning model evaluation.

For MLP (Neural Networks):

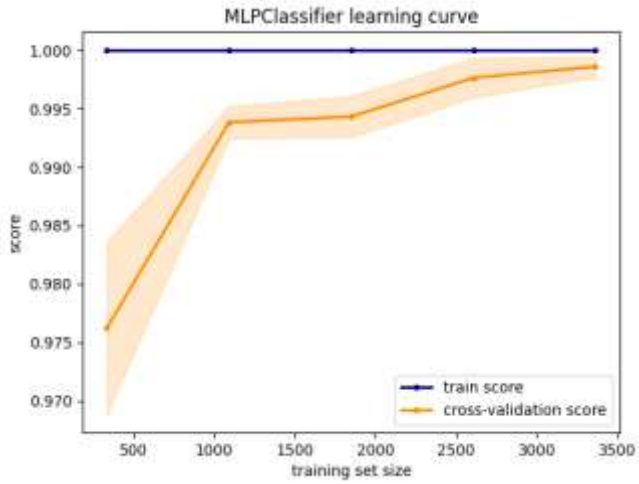

FIGURE 15. **MLP Learning Curve.**

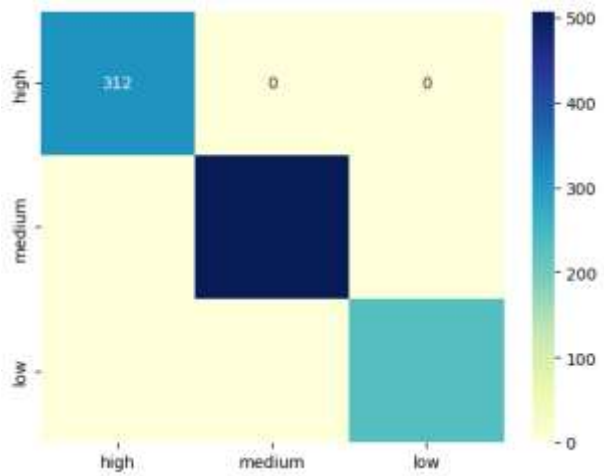

FIGURE 16. **Heat map of the accuracy scores**

VIII. SEQUENTIAL PATTERN MINING

The SPMF (Sequential Pattern Mining Framework) algorithm is a collection of algorithms designed for discovering sequential patterns in sequential data. This type of algorithm is commonly used in analyzing transactional datasets, such as those found in retail or e-commerce, to find patterns in the order in which events occur. In the context of the provided dataset, which contains transactional data, we can use SPMF algorithms to discover sequential patterns in customers' purchase behaviour.

- **Data Preprocessing:** Prepare the dataset in a suitable format for sequential pattern mining. Each transaction should be represented as a sequence of items (products) purchased by a customer.
- **Algorithm Selection:** Prefix Span was chosen from the SPMF framework based on the specific requirements of the analysis.
- **Parameter Tuning**: The parameters such as minimum support threshold, maximum pattern length, and minimum pattern length should be configured based on the characteristics of the dataset and the desired patterns to be discovered.
- **Pattern Mining:** Apply the selected SPMF algorithm to the preprocessed dataset to mine sequential patterns [14].
  The output of the algorithm will be a set of sequential patterns representing common sequences of items purchased by customers.
- **Pattern Analysis**: The discovered sequential patterns should be analyzed to gain insights into customer behaviour and preferences.
  Identified frequent sequences of items that are often purchased together or in a particular order.
  Used visualization techniques to present the discovered patterns in an understandable format.
- **Interpretation and Action:** Interpret the discovered patterns to derive actionable insights for business decision-making.
  Utilize the insights to optimize product recommendations, marketing strategies, and inventory management processes.

Algorithm 1 Large-transaction generation algorithm using prefixspan for extracting large 1-sequential patterns from route database

Input:
RD (Route Database)
min_sup_count (Minimum support count)
Output:
r (Set of large 1-sequential patterns)
Method:

```
1: for each event (Bi, fi itemset;) in RD
2:    if itemset, is not empty then
3:       for each z itemset, // z is non-empty subset of
         itemset,
4:          if < B1, z > is not exist in I then
5:             add y = < B1, z > to I', and set its sup_count
         to 1;
6:          else
7:             increase sup_count of < Bj, z > by 1;
8:          end if
9:       end for
10:   end if
11: end for
12: for each y = < Bj z > in I
13:    if sup_count of y min_sup_count then
14:       give z a unique symbol and save to I';
15:    end if
16: end for
```

IX. RESULTS AND DISCUSSION

A. PREDICTIVE MODEL RESULTS

Before presenting the classification reports for the Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) models, it's essential to highlight the significance of predictive model evaluation in understanding the performance of these algorithms. Evaluation metrics such as precision, recall, and F1-score provide valuable insights into the models' ability to correctly classify instances across different classes [15]. These metrics, along with accuracy, macro-average, and weighted-average scores, offer a comprehensive view of the models' effectiveness in handling the given dataset.

The following are the results from the predictive models built:

TABLE II. Classification Report for SVM.

	precision	recall	f1-score	support
high	0.99	1	0.99	312
low	1	1	1	232
medium	1	0.99	0.99	508
accuracy			0.99	1052
macro Avg	0.99	0.99	0.99	1052
Weighted avg	0.99	0.99	0.99	1052

TABLE III. Classification Report for MLP.

	precision	recall	f1-score	support
high	1	1	0.99	312
low	1	1	1	232
medium	1	1	0.99	508
accuracy			0.99	1052
macro Avg	1	0.99	1	1052
Weighted avg	0.99	1	0.99	1052

Table II and Table III show both the Support Vector Classifier (SVC) and the Multi-Layer Perceptron (MLP) models demonstrated excellent performance across all classes with high precision, recall, and F1-scores. The accuracy of both models was 0.99, indicating that they were able to correctly classify instances into their respective classes with a high degree of accuracy. The MLP model showed slightly better performance in terms of the macro average F1-score. Overall, both models are suitable for the classification task and can be considered effective for predicting customer segments based on the provided dataset.

B. SEQUENTIAL DATA MINING RESULTS

The following are the results obtained from the prefixspan method:

>/home/Akila Bala/akila.bala24@gmail.com/OnlineRetail-Master/spmf.jar
===== PREFIXSPAN 0.99-2016 – STATISTICS =====
Total time ~ 756 ms
Frequent sequences count: 419
Max memory (mb): 257.60943603515625
minsup = 159 sequences.
Pattern count: 419

Post-processing to show results in terms of string values.
Post-processing completed.

Based on the results obtained from the PrefixSpan algorithm run:

- Total Time: The algorithm took approximately 756 milliseconds to complete.
- Frequent Sequences Count: There are 419 frequent sequences found in the dataset.
- Minimum Support (minsup): The minimum support threshold is 159 sequences.
- Pattern Count: A total of 419 patterns were discovered.
- Max Memory: The maximum memory used during the algorithm's execution was approximately 257.61 megabytes.

These results suggest that the PrefixSpan algorithm successfully mined frequent sequential patterns from the dataset with the specified minimum support threshold. The discovered patterns can provide valuable insights into the sequential behaviour of the data, which can be further analyzed and interpreted for various purposes such as market basket analysis, recommendation systems, or customer behaviour analysis.

X. CONCLUSION

In the analysis conducted, we delved into a transactional dataset to uncover valuable insights into customer behaviour and preferences. Through thorough exploratory data analysis (EDA), we identified trends such as popular products, customer demographics, and seasonal sales patterns. Data cleaning and preprocessing were crucial steps to ensure the dataset's quality, including handling missing values and removing irrelevant transactions like cancellations. Leveraging RFM (Recency, Frequency, Monetary) features, we employed predictive analysis techniques with Support Vector Machine (SVM) and Neural Network classifiers to accurately predict customer segments. Both models exhibited impressive performance metrics, highlighting their effectiveness in classifying instances into the correct segments. Additionally, sequential pattern mining using the PrefixSpan algorithm revealed frequent sequences of customer purchase behaviour, offering valuable insights for targeted marketing and personalized recommendations. By integrating these analyses, businesses can optimize strategies for inventory management, customer engagement, and overall operational efficiency, ultimately driving growth and enhancing customer satisfaction.

REFERENCES

[1] Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19, 197-208.

[2] Agarwal, P. (2014). Benefits and issues surrounding data mining and its application in the retail industry. International Journal of Scientific and Research Publications, 4(7), 1-5.

[3] Kumar, M. R., Venkatesh, J., & Rahman, A. M. Z. (2021). Data mining and machine learning in retail business: developing efficiencies for better customer retention. Journal of Ambient Intelligence and Humanized Computing, 1-13.

[4] Li, H. (2005, June). Applications of data warehousing and data mining in the retail industry. In Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005. (Vol. 2, pp. 1047-1050). IEEE

[5] Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. Machine Learning, 57, 83-113.

[6] Hormozi, A. M., & Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. Information systems management, 21(2), 62-71.

[7] Muley, P. A. (2022). Application of Data Mining Technique for Retail Industry. In Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021 (pp. 973-981). Springer Singapore.

[8] Zhang, X., Edwards, J., & Harding, J. (2007). Personalised online sales using web usage data mining. Computers in Industry, 58(8-9), 772-782.

[9] Ahmeda, R. A. E. D., Shehaba, M. E., Morsya, S., & Mekawiea, N. (2015, April). Performance study of classification algorithms for consumer online shopping attitudes and behaviour using data mining. In 2015 Fifth International Conference on Communication Systems and Network Technologies (pp. 1344-1349). IEEE.

[10] Srikant, R., & Agrawal, R. (1996, March). Mining sequential patterns: Generalizations and performance improvements. In International conference on extending database technology (pp. 1-17). Berlin, Heidelberg: Springer Berlin Heidelberg.

[11] Magnini, V. P., Honeycutt Jr, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use and limitations. Cornell Hotel and Restaurant Administration Quarterly, 44(2), 94-105.

[12] Ritbumroong, T. (2015). Analyzing customer behaviour using online analytical mining (OLAM). Integration of data mining in business intelligence systems, 98-118.

[13] Hemalatha, M. (2012). Market basket analysis–a data mining application in Indian retailing. International Journal of Business Information Systems, 10(1), 109-129.

[14] Huang, C. K., Chang, T. Y., & Narayanan, B. G. (2015). Mining the change of customer behaviour in dynamic markets. Information Technology and Management, 16, 117-138.

[15] Research Synthesis by Systematic Literature Review and Data Mining Techniques" Punpukdee, A., Wattana, C., Punpairoj, W., Srichuachom, U., Yaklai, P., & Trongtortam, S. (2021).

Dr. Kavitha Dhanushkodi is currently working as an Assistant Professor in the School of Computer Science and Engineering (SCOPE) at Vellore Institute of Technology, Chennai Campus, Chennai, Tamil Nadu, India. She has completed PhD and Master of Engineering in Computer Science and Engineering from Anna University, Chennai, Her research interests are mainly focused on Software Security, Internet of Things and Cyber Security. She has an overall teaching experience of 14 years in various academic institutions. She has published more than 40 Research papers to her credit in reputed international journals and conferences with high impact factors.

Akila Bala is currently a student Professor in the School of Computer Science and Engineering (SCOPE) at Vellore Institute of Technology, Chennai Campus, Chennai, Tamil Nadu, India. She is pursuing her Master's degree which specializes in Business Analytics, her research interests are mainly focused on Network Security, Artificial Intelligence and Data Science. She has appeared in many conferences and published one journal paper.

Nithin Kodipyaka is currently a student in the School of Computer Science and Engineering (SCOPE) at Vellore Institute of Technology, Chennai Campus, Chennai, Tamil Nadu, India. He is pursuing his Master's degree which specializes in Business Analytics, His research interests are mainly focused on Information Security, Deep Learning and Data Analytics. He has appeared in many conferences and published one journal paper.

Shreyas V is currently a student in the School of Computer Science and Engineering (SCOPE) at Vellore Institute of Technology, Chennai Campus, Chennai, Tamil Nadu, India. He is pursuing his Master's degree which specializes in Business Analytics, His research interests are mainly focused on Robotics Deep Learning and Data Mining.