# Consumer Behavior Analysis in Retail Transactions Using Machine Learning Techniques

Andreas Kanavos*, Gerasimos Vonitsanos† and Phivos Mylonas‡
*Department of Informatics
Ionian University, Corfu, Greece
akanavos@ionio.gr
†Computer Engineering and Informatics Department
University of Patras, Patras, Greece
mvonitsanos@ceid.upatras.gr
‡Department of Informatics and Computer Engineering
University of West Attica, Athens, Greece
mylonasf@uniwa.gr

***Abstract*—Retail analytics has become essential for supermarkets, which generate vast volumes of transactional data on a daily basis. Harnessing this data to model consumer behavior enables informed decision-making in marketing, inventory management, and customer engagement. This paper proposes an integrated framework that combines association rule mining, k-means clustering, and Random Forest classification to extract and exploit insights from retail transactions. Association rules reveal frequent product co-occurrences, supporting interpretable strategies for cross-selling and shelf organization. Clustering uncovers distinct consumer segments, highlighting heterogeneity in purchasing patterns across shopper groups. Random Forest achieves superior predictive performance, demonstrating robustness in modeling transaction-level attributes and forecasting product-line preferences. Experiments on publicly available supermarket datasets validate the effectiveness of the framework, showing that specialized models outperform global baselines and that descriptive insights complement predictive results. By integrating descriptive, unsupervised, and supervised learning, the framework provides both actionable knowledge and interpretable forecasts, offering a comprehensive decision-support tool for modern retail environments.**

***Index Terms*—Consumer Behavior Analysis, Retail Analytics, Transaction Data, Machine Learning, Association Rule Mining, Clustering**

## I. INTRODUCTION

The analysis of consumer behavior is a cornerstone of modern retail strategy, providing insights into customer needs, preferences, and purchasing patterns. Among retail sectors, supermarkets generate particularly large volumes of transactional data that, when systematically analyzed, can deliver substantial competitive advantages [4], [5]. Advances in digital technologies and business intelligence have further emphasized the importance of transforming raw transactions into actionable knowledge, enabling improved personalization, optimized product placement, and more efficient inventory management [21].

979-8-3315-0450-2/24/$31.00 © 2024 IEEE

Machine learning and data mining techniques have become central to this transformation. Association rule mining captures frequent co-purchasing patterns [1], clustering supports the identification of consumer groups with distinct purchasing behaviors [12], and classification algorithms such as Random Forest provide predictive insights into future purchasing decisions [3]. Despite their effectiveness, most previous studies have examined these methods in isolation, which restricts their ability to deliver a comprehensive understanding of consumer behavior. This limitation highlights the need for integrated analytical frameworks that combine descriptive and predictive perspectives.

In response to this need, the present work introduces a unified machine learning framework for retail transaction analysis. The framework integrates association analysis, clustering, and classification within a single pipeline, enabling complementary insights across descriptive, unsupervised, and supervised learning tasks [16]. Through this integration, it becomes possible to identify meaningful co-purchasing patterns, uncover heterogeneity across consumer segments, and generate accurate forecasts of product-line preferences. Beyond methodological novelty, the framework demonstrates practical value by producing interpretable outputs that can inform shelf organization, targeted marketing campaigns, and inventory decisions.

The main contributions of this study are twofold. First, it provides an integrated analytical framework that combines multiple machine learning techniques to bridge the gap between descriptive and predictive modeling in retail analytics. Second, it validates the effectiveness of the framework on publicly available supermarket datasets, showing that integrated models outperform traditional baselines and that descriptive insights complement predictive outcomes. Taken together, these contributions establish a practical and interpretable approach to data-driven decision making in retail environments.

The remainder of the paper is organized as follows. Section II reviews prior research on consumer behavior analysis and data mining techniques in retail. Section III presents the proposed framework and details the integration of association rule mining, clustering, and classification. Section IV describes the datasets, preprocessing steps, experimental setup, evaluation metrics, and implementation details. Section V reports the findings, including identified purchasing patterns, consumer segments, and predictive results. Finally, Section VI summarizes the contributions, discusses managerial implications, and outlines directions for future research.

## II. Related Work

In recent years, machine learning and data mining have become central to the analysis of consumer behavior in retail environments. Research has advanced along several directions, including association rule mining for product recommendation, clustering for customer segmentation, and supervised learning for predictive analytics.

Apriori remains a cornerstone of market basket analysis [1]. Variants of the algorithm have been widely applied to supermarket datasets, successfully identifying frequent product co-occurrences and informing cross-selling strategies. However, most of these works focus exclusively on descriptive rule generation without integrating predictive components, limiting their contribution to decision-making processes.

Clustering has also been recognized as a valuable tool in consumer analytics. A systematic review of data mining applications in customer relationship management emphasized clustering as an effective method for segmentation [12]. Building on this, k-means has been applied to form behavioral groups of consumers [17], offering practical value for targeted marketing. Other studies in retail contexts have applied segmentation techniques [15], but often without linking clusters to predictive modeling tasks or evaluating ensemble-based methods, leaving segmentation insights underutilized.

Supervised learning approaches have likewise gained prominence in retail analytics. Early studies demonstrated the applicability of classification algorithms for structured retail data [22], while later work highlighted the strong performance of ensemble methods such as Random Forests, which combine predictive accuracy with robustness to noisy, high-dimensional data [3]. Additional contributions have shown that ensemble methods enhance interpretability through feature importance analysis, strengthening their value for managerial decision support [14]. More recently, applications of machine learning to supermarket sales data have demonstrated the utility of decision trees, Random Forests, and Gradient Boosting for tasks such as predicting product demand, forecasting customer purchasing behavior, and classifying transactions by product line [8]. Despite their predictive strength, these studies generally assess classifiers in isolation rather than embedding them into integrated pipelines.

Overall, prior research has produced valuable insights but remains methodologically fragmented—often restricted to descriptive association rules, unsupervised segmentation, or standalone classifiers. To address these limitations, the present work introduces an integrated framework that unifies association rule mining, clustering, and supervised classification. This combination enables both descriptive and predictive insights: uncovering product relationships, identifying customer segments, and forecasting product-line preferences. Such a holistic approach provides a more comprehensive and actionable foundation for data-driven decision-making in the retail sector.

## III. Proposed Analytical Framework

This study employs three machine learning techniques—Association Rule Mining, k-means clustering, and Random Forest Classification—for extracting knowledge from supermarket transaction data. These methods were selected for their complementary strengths: Association Rule Mining uncovers frequent co-purchasing relationships, k-Means Clustering identifies consumer segments with distinct purchasing profiles, and Random Forest provides predictive insights into purchasing behaviors. Together, they form an integrated analytical pipeline for consumer behavior analysis in retail.

### A. Association Rule Mining

Association Rule Mining is a widely used technique for discovering relationships between items in transactional datasets. It aims to identify rules of the form A $\rightarrow$ B, where the purchase of itemset A implies a likelihood of purchasing itemset B. The Apriori algorithm [1] generates these rules by iteratively expanding frequent itemsets that satisfy minimum thresholds of support and confidence. The evaluation of such rules relies on three key metrics [17]:

- **Support**: measures how frequently items $A$ and $B$ occur together in the dataset.

$$Support(A \rightarrow B) = \frac{\text{Transactions containing } \{A, B\}}{\text{Total transactions}}$$

- **Confidence**: quantifies the probability of purchasing $B$ given that $A$ has been purchased.

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$$

- **Lift**: evaluates the significance of the rule by comparing the observed co-occurrence of $A$ and $B$ to what would be expected if they were independent.

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)}$$

In this study, Apriori is used to identify frequent product combinations in shopping baskets, providing insights for product placement, cross-selling, and promotional strategies [9].

### B. k-Means Clustering

Clustering is an unsupervised learning technique that partitions data into groups based on similarity [10], [11]. The k-Means algorithm [17] is used to segment customers based on their purchasing behavior. The objective is to minimize the within-cluster variance, expressed as:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where $C_i$ is the set of points assigned to cluster $i$, $\mu_i$ is the centroid of cluster $i$, and $\|x - \mu_i\|^2$ is the squared Euclidean distance between data point $x$ and the centroid $\mu_i$. The algorithm iteratively updates cluster assignments and centroids until convergence, yielding groups of customers with similar purchasing patterns.

Customer features such as purchase frequency, total spending, and product category preferences are used as inputs. This segmentation enables the identification of distinct consumer groups (e.g., frequent buyers, occasional shoppers, category-specific customers) and supports the design of targeted marketing campaigns [20].

### C. Random Forest

Random Forest is a widely used ensemble learning technique that combines the predictions of multiple decision trees to achieve robust classification performance [3], [19]. Each tree is trained on a bootstrapped sample of the dataset, and at every split, a random subset of features is considered, which increases diversity among trees and enhances generalization. The final output of the ensemble is obtained through majority voting across all trees:

$$\hat{y} = \mathrm{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{1}$$

where $h_t(x)$ denotes the prediction of the $t$-th tree, and $T$ represents the total number of trees in the forest. This ensemble approach reduces variance, limits overfitting, and operates effectively with minimal data preprocessing [18]. Moreover, Random Forests provide estimates of feature importance, which are valuable for understanding the relative contribution of different input variables [13], [7].

In this study, Random Forest is applied to supermarket transaction data to predict purchasing behaviors. The model highlights key factors—such as purchase frequency, spending levels, and dominant product categories—that shape consumer decisions. These findings enable actionable insights for targeted marketing strategies, personalized recommendations, and more efficient inventory management, underscoring the method's value in retail analytics.

## IV. Experimental Evaluation

This section presents the experimental evaluation of the proposed framework for analyzing supermarket transactions. The aim is to examine the effectiveness of Association Rule Mining, k-Means Clustering, and Random Forest Classification in uncovering purchasing patterns, segmenting consumers, and predicting behavioral outcomes. Each technique was assessed using task-appropriate metrics: Support, Confidence, and Lift were used to evaluate the strength of association rules, cluster distributions, and visual inspection were employed to assess the interpretability of k-Means clustering results. Finally, accuracy, precision, recall, F1-score, and C-index were applied to measure the predictive performance of the Random Forest model.

### A. Dataset

The experiments utilized two publicly available datasets:

- **Groceries Dataset (Kaggle) [6]** – consisting of 38,765 supermarket transactions, where each record contains a list of purchased items. This dataset is particularly suited for Association Rule Mining and market basket analysis. This dataset is widely used in market basket analysis and is particularly suitable for Association Rule Mining, as it enables the discovery of frequently co-purchased items and product associations.

- **Supermarket Sales Dataset (Kaggle) [2]** – containing 1,000 sales records across three branches, annotated with attributes such as customer type, gender, product line, total spending, tax, payment method, and purchase date. This dataset supports both clustering and classification tasks. This richer structure makes it appropriate for clustering tasks, where customers can be segmented into behavioral groups, and for classification tasks, where purchasing preferences can be predicted.

Before applying machine learning techniques, both datasets were preprocessed to ensure quality and consistency. Duplicate entries were removed, missing values were handled where necessary, and numerical attributes (e.g., unit price, total amount, and tax) were normalized to avoid scale imbalances. Categorical variables, including the product line method, were encoded into numerical form to ensure compatibility with the algorithms. For the Groceries dataset, transactions were further transformed into a binary matrix representation, where each column corresponds to an item and each row to a transaction, enabling efficient application of the Apriori algorithm.

### B. Experimental Setup and Preprocessing

The experiments were implemented in Python 3.9, using a set of well-established libraries for data analysis and machine learning. Pandas and NumPy were employed for data preprocessing, including cleaning, normalization, and transformation of raw transaction data into formats suitable for analysis. Matplotlib was used to produce visualizations that illustrate consumer segments and purchasing patterns.

For model implementation, Mlxtend was applied to run the Apriori algorithm for association rule mining, while Scikit-learn was used for clustering with k-Means and classification with Random Forest. Feature scaling was performed using StandardScaler, and the datasets were partitioned into training and test subsets using the train–test split method. Performance was evaluated through accuracy, precision, recall, F1-score, and C-index.

### C. Evaluation Metrics

The performance of the proposed framework was evaluated using metrics specific to each analytical task. For Association Rule Mining, the quality of the extracted rules was measured using Support, Confidence, and Lift. Support quantifies the relative frequency of itemsets in the transaction database. Confidence expresses the conditional probability of purchasing an item given the presence of another. Lift evaluates the strength of the association compared to random co-occurrence. These metrics collectively determine the relevance and usefulness of discovered product associations.

For clustering with k-Means, evaluation was conducted through the interpretability of the resulting consumer segments. In the absence of ground-truth labels, cluster quality was assessed through visual inspection of cluster distributions and analysis of characteristic features, including purchase frequency, total spending, and product line preferences. This approach highlights the practical utility of segmentation for marketing and customer profiling.

For classification with Random Forest, model performance was evaluated using accuracy, defined as the proportion of correctly classified instances relative to the total number of instances. Accuracy was selected as the primary metric due to the balanced nature of the dataset and its suitability for assessing overall classification reliability. To complement this, precision, which measures the proportion of correctly predicted positive cases among all predicted positives, and recall, which quantifies the proportion of true positives correctly identified among all actual positives, were employed to capture class-specific performance. The F1-score, calculated as the harmonic mean of precision and recall, was further included to balance the trade-off between false positives and false negatives. In addition, the Concordance Index (C-Index) was computed as a measure of discriminative ability, assessing how effectively the model ranks predictions across pairs of instances.

### *D. Hyperparameter Tuning*

Hyperparameter tuning was conducted to enhance model generalization and ensure reliable convergence of the applied algorithms. For Random Forest, parameters including the number of trees, maximum depth, and minimum samples per split were optimized using grid search with cross-validation. For k-Means, the optimal number of clusters was identified through the elbow method and silhouette analysis, enabling meaningful customer segmentation. In the case of Apriori, minimum support and confidence thresholds were systematically varied to achieve a balance between the number of rules generated and their interpretability. This process reduced overfitting, improved model robustness, and increased the practical relevance of the framework for analyzing supermarket transactions.

## V. Results and Discussion

This section presents the performance evaluation of the proposed framework, which integrates Association Rule Mining, k-Means Clustering, and Random Forest Classification for analyzing supermarket transactions. Each method was evaluated using appropriate metrics: Support, Confidence, and Lift for association rules, cluster distributions and interpretability for customer segmentation, and accuracy, precision, recall, F1-score, and C-index for classification tasks. The results are reported through figures and tables, followed by a detailed analysis and discussion of the insights derived from each component.

Figure 1 presents the top ten association rules extracted with the Apriori algorithm, evaluated using Support, Confidence, and Lift. Support values range from 2.4% to 4.2%, which is typical in large-scale retail datasets where specific product combinations appear in only a small fraction of transactions. Despite these modest values, several rules exhibit substantial confidence levels (29%–40%), indicating that the identified associations recur consistently across customers.

All rules achieve lift values above 1.2, confirming that the observed co-occurrences are not random but reflect genuine purchasing behavior. For instance, the rule whole milk, yogurt → other vegetables, rolls/buns combines relatively high confidence with a lift of 1.6, underscoring its potential relevance for cross-selling and product placement strategies. Similarly, rules involving staple items such as whole milk and sausage highlight systematic co-purchasing patterns that can inform data-driven decisions in supermarket management.

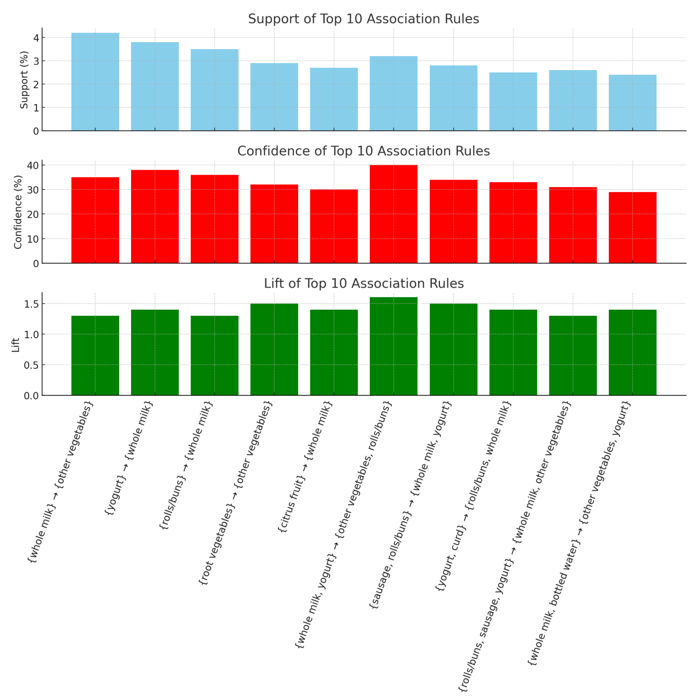


Fig. 1. Empirical Assessment of Apriori-Derived Association Rules Using Support, Confidence, and Lift

Figure 2 presents the segmentation of supermarket customers obtained with the k-Means algorithm, which identified three distinct clusters. Cluster 1 (35%) comprises high-spending customers characterized by frequent purchases across a wide range of product categories. Cluster 2 (40%) includes medium-spending consumers with balanced purchasing patterns in both frequency and diversity. Cluster 3 (25%)

represents low-frequency shoppers whose transactions are concentrated in specific categories such as beverages and snacks.

These results confirm that consumer behavior in supermarket transactions is heterogeneous and can be effectively summarized through well-defined profiles. The derived segments provide actionable insights for retailers, supporting targeted marketing campaigns, loyalty program design, and personalized product recommendations. Moreover, identifying high-value groups enables the development of differentiated promotional strategies and more efficient resource allocation in retail management.

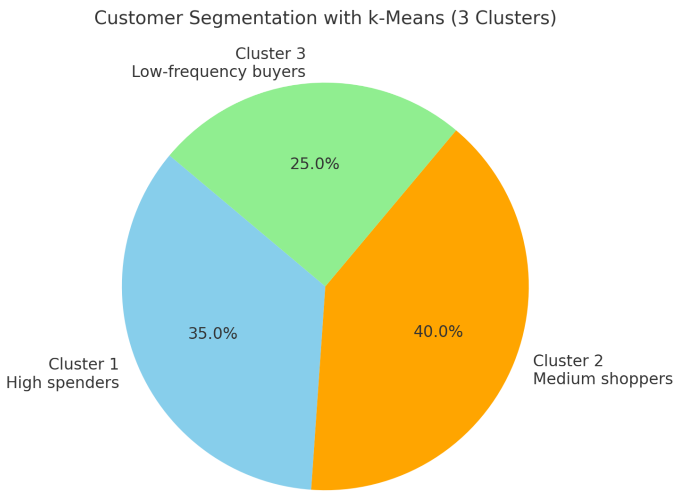


Fig. 2. Distribution of Consumer Groups Identified by k-Means Algorithm

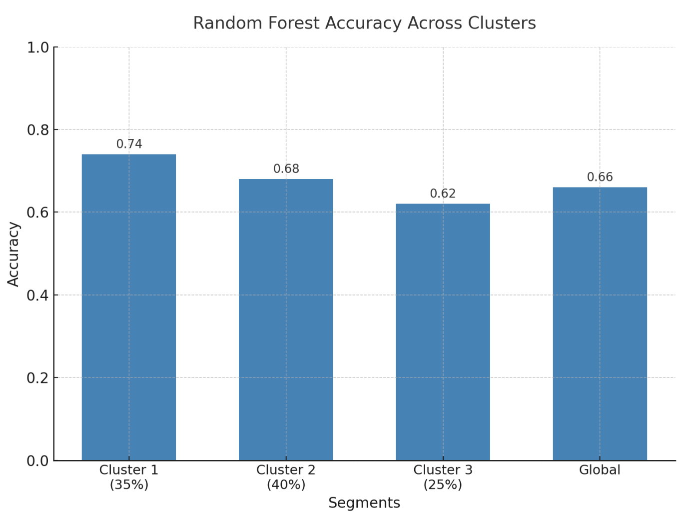


Fig. 3. Impact of Customer Segmentation on Random Forest Accuracy

Figure 3 presents the classification accuracy of Random Forest models trained independently for each customer cluster and compares them to a global model without segmentation. Results indicate that Cluster 1 (0.74) and Cluster 2 (0.68) outperform the global baseline (0.66), demonstrating that conditioning prediction on customer segments enhances model performance. By contrast, Cluster 3 (0.62) yields lower accuracy, which can be attributed to the greater heterogeneity and reduced predictability of shoppers with low frequency.

Table I reports the predictive performance of Random Forest in forecasting the product line purchased within each customer cluster. The model achieved its best results in Cluster 1 (high spenders), with a precision of 0.76, recall of 0.72, F1-score of 0.74, and a C-Index of 0.79. These values indicate that product-line preferences among high-value customers are relatively consistent and can be predicted with strong reliability. Cluster 2 (medium shoppers) exhibited moderate performance, with precision of 0.70, recall of 0.66, F1-score of 0.68, and a C-Index of 0.75, reflecting more heterogeneous purchasing behavior. Prediction was most challenging in Cluster 3 (low-frequency buyers), where the Random Forest achieved lower precision (0.64), recall (0.60), F1-score (0.62), and a C-Index of 0.71, highlighting the difficulty of modeling the irregular and less consistent purchasing behavior of this segment.

The global model trained without clustering reached a precision of 0.68, recall of 0.64, F1-score of 0.66, and a C-Index of 0.74, which is lower than the performance obtained in Cluster 1 and Cluster 2 individually. This confirms that conditioning the model on customer segments enhances its discriminative ability, particularly for high-value and medium-value shoppers. Overall, these results highlight the benefit of combining unsupervised segmentation with supervised prediction, as it enables the framework to capture both consumer heterogeneity and product-line tendencies more effectively.

TABLE I
RANDOM FOREST FOR PRODUCT LINE PREDICTION ACROSS CLUSTERS

| Cluster | Accuracy | Precision | Recall | F1-Score | C-Index |
|---|---|---|---|---|---|
| 1 | 0.74 | 0.76 | 0.72 | 0.74 | 0.79 |
| 2 | 0.68 | 0.70 | 0.66 | 0.68 | 0.75 |
| 3 | 0.62 | 0.64 | 0.60 | 0.62 | 0.71 |
| Global | 0.66 | 0.68 | 0.64 | 0.66 | 0.74 |

## A. *Discussion*

The comparative evaluation highlights three main findings. First, association rule mining complemented these methods by identifying frequent product co-occurrences, providing interpretable insights that contextualize and validate the predictive result. Second, k-Means clustering effectively captured customer heterogeneity, as segment-level models consistently outperformed the global baseline, particularly for high-value and medium-value shoppers. Third, Random Forest proved to be the most effective predictive component of the framework, offering strong discriminative ability and balanced precision–recall performance across clusters.

From a retail analytics perspective, these findings emphasize the importance of combining predictive accuracy with interpretability. While Random Forest generates reliable product-line forecasts, its decision boundaries are often opaque to business users. Association rules bridge this gap by uncovering explicit product-to-product relationships that not only support the plausibility of predictions but also provide managers with actionable knowledge for cross-selling and shelf-placement strategies. In this way, association rules strengthen the trustworthiness of the predictive outputs by aligning them with observable consumer behavior, while simultaneously increasing their business relevance. Similarly, clustering offers actionable customer profiles that link predictive outcomes to well-defined behavioral segments, enabling targeted marketing initiatives and loyalty programs tailored to distinct groups.

## VI. Conclusions and Future Work

This study has demonstrated the effectiveness of machine learning methods in supporting retail decision-making by analyzing supermarket transactions. By leveraging transactional attributes such as product categories, spending levels, purchase frequency, and customer segments, the proposed framework generates actionable insights that enhance customer understanding and support data-driven marketing, personalization, and inventory management.

Among the applied techniques, Random Forest emerged as the most effective predictive component, achieving balanced precision and recall across clusters while also yielding consistent F1-scores, which highlight its ability to minimize both false positives and false negatives. This robustness underscores its reliability in forecasting product line trends. Its predictive capabilities are complemented by association rule mining, which uncovers explicit product-to-product relationships, and by k-means clustering, which captures customer heterogeneity and defines distinct behavioral profiles. Together, these methods provide a unified framework that combines predictive power, interpretability, and segmentation, offering retailers a comprehensive tool for both operational and strategic decision-making.

Future work will advance this framework in several directions. First, automated hyperparameter optimization techniques will be explored to enhance predictive performance further. Second, additional machine learning models, including Gradient Boosting, Support Vector Machines, and Deep Neural Networks, will be compared to evaluate their suitability for retail analytics. Third, larger and more diverse datasets from multiple retail environments will be incorporated to enhance generalizability and capture variations in consumer behavior across contexts. Finally, integrating the temporal dynamics of transactions will allow the framework to account for seasonality and evolving purchasing patterns.

Looking forward, embedding this framework into intelligent retail decision-support systems and recommender platforms can provide real-time, personalized insights for both consumers and managers. Such tools can inform marketing strategies, optimize product placement, and enhance customer engagement, while also giving interpretable outputs that foster trust and adoption among business stakeholders. By striking a balance between accuracy and interpretability, the framework provides a practical pathway toward smarter, data-driven retail operations.

In conclusion, this work contributes to the growing field of retail analytics by demonstrating how ensemble learning, unsupervised segmentation, and rule-based discovery can be integrated into a coherent methodology. By bridging methodological advances with real-world retail challenges, it paves the way toward intelligent, context-aware systems that empower retailers, enhance customer satisfaction, and support sustainable business growth.

## References


[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I., et al.: Fast discovery of association rules. Advances in knowledge discovery and data mining **12**(1), 307–328 (1996)

[2] Ashraf, F.: Supermarket sales dataset (2025), https://www.kaggle.com/datasets/faresashraf1001/supermarket-sales, accessed: August 2025

[3] Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)

[4] Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information sciences **275**, 314–347 (2014)

[5] Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: From big data to big impact. MIS quarterly pp. 1165–1188 (2012)

[6] Dedhia, H.: Groceries dataset (2020), https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset, accessed: August 2025

[7] Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Pattern recognition letters **31**(14), 2225–2236 (2010)

[8] Granell Tormo, R.E.: Prediction, clustering and analysis of supermarket and retail energy demand data using machine learning techniques. Ph.D. thesis, Brunel University London (2024)

[9] Hahsler, M., Grün, B., Hornik, K.: arules-a computational environment for mining association rules and frequent item sets. Journal of statistical software **14**, 1–25 (2005)

[10] Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010)

[11] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press (1967)

[12] Ngai, E.W., Xiu, L., Chau, D.C.: Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications **36**(2), 2592–2602 (2009)

[13] Pal, M.: Random forest classifier for remote sensing classification. International journal of remote sensing **26**(1), 217–222 (2005)

[14] Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence **32**(3), 569–575 (2009)

[15] Singh, A., Rumantir, G., South, A., Bethwaite, B.: Clustering experiments on big transaction data for market segmentation. In: Proceedings of the 2014 International Conference on Big Data Science and Computing. pp. 1–7 (2014)

[16] Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics **8**(1), 25 (2007)

[17] Tan, P.N., Steinbach, M., Kumar, V.: Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining **487**(533), 58 (2013)

[18] Vonitsanos, G., Gounaridis, I., Kanavos, A., Mylonas, P.: Enhancing aviation efficiency through big data and machine learning for flight delay prediction. In: Novel & Intelligent Digital Systems Conferences. pp. 524–536. Springer (2024)

[19] Vonitsanos, G., Kanavos, A., Mylonas, P.: Evaluating machine learning techniques for enhanced prediction of building energy consumption. In: 2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM). pp. 50–57. IEEE (2024)

[20] Wedel, M., Kamakura, W.A.: Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media (2000)

[21] Wedel, M., Kannan, P.: Marketing analytics for data-rich environments. Journal of marketing **80**(6), 97–121 (2016)

[22] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record **31**(1), 76–77 (2002)