JOURNAL INNOVATIONS COMPUTER SCIENCE
E-ISSN: 2961-970X, P-ISSN: 2961-9718
Volume 4 Number 2, November 2025, pp 242-253.
DOI: https://doi.org/10.56347/jics.v4i2.303

Clustering and Classification of Retail Sales Data: A Big Data and Data Mining Analysis

Ahmad Bilal Almagribi [a*], Sri Redjeki [b]
[a*,b] Master of Information Technology, Universitas Teknologi Digital Indonesia, Bantul Regency, Special Region of Yogyakarta, Indonesia.

ABSTRACT
In the evolving retail landscape, data-driven decision-making has become essential for understanding customer behavior and predicting sales trends. This study integrates clustering and classification techniques to analyze retail sales data comprising 1,000 transactions obtained from Kaggle. Using the K-Means algorithm, three optimal customer clusters were identified through the Elbow Method, achieving an average within-centroid distance of 25,272.635 and a Davies–Bouldin Index of 0.443, indicating clear cluster separation. The subsequent classification phase compared the predictive performance of three algorithms—Naïve Bayes, Decision Tree, and Random Forest—on 70:30 training-to-testing data partitions. The Naïve Bayes algorithm attained 94.67% accuracy, while both Decision Tree and Random Forest achieved perfect classification accuracy of 100%. These findings highlight the robustness and adaptability of tree-based models for complex retail datasets, outperforming probabilistic methods in terms of accuracy and generalization. The results suggest that the integration of clustering and classification provides retailers with a powerful analytical framework for identifying high-value customer segments, optimizing marketing strategies, and enhancing inventory management. Despite achieving strong outcomes, the study acknowledges dataset limitations and recommends future research involving larger and more diverse datasets, as well as additional features, to expand model scalability and predictive precision.

ARTICLE HISTORY
Received 28 July 2025
Accepted 30 October 2025
Published 30 November 2025

KEYWORDS
Retail Analytics; K-Means; Naïve Bayes; Decision Tree; Random Forest.

1. Introduction

In the contemporary retail environment, where consumer preferences fluctuate rapidly, organizations striving for consistent sales growth must balance customer demand with effective inventory management. Reliable forecasting enables firms to plan production, design pricing strategies, and sustain market competitiveness (Chen *et al*., 2021; Wei & Zeng, 2021). The acceleration of e-commerce and mobile-based purchasing has reshaped retail dynamics, producing consumption patterns that are more fragmented, unstable, and volatile (Li, 2022). Conventional analytical methods are increasingly insufficient for capturing such complexity, prompting retailers to adopt computational approaches that enhance predictive precision (Shetty & Shetty, 2023). The capacity to analyze customer behavior and forecast product demand now depends heavily on data-driven modeling and algorithmic decision support (Awan *et al*., 2021; Niu, 2020). As data volumes grow exponentially, the strategic use of big

CONTACT Ahmad Bilal Almagribi abalmagribi@gmail.com Master of Information Technology, Universitas Teknologi Digital Indonesia, Bantul Regency, Special Region of Yogyakarta, Indonesia.

data becomes fundamental for identifying sales trends, understanding purchasing behavior, and refining marketing actions. Recent developments in artificial intelligence have expanded methodological possibilities for sales prediction, including the use of hybrid and ensemble learning frameworks (Chen *et al*., 2021; Wei & Zeng, 2021). A range of studies has examined algorithmic performance for retail forecasting: Wahyudi and Silfia (2022) applied K-Means clustering to determine sales strategies using the Davies–Bouldin Index for validation; Andry *et al*. (2023) integrated Decision Tree, K-Means, and Association Rules for supermarket sales prediction; Arman *et al*. (2023) compared K-Means, Naïve Bayes, and Decision Tree algorithms for fuel sales forecasting, reporting higher accuracy for Decision Tree; and Firnanda *et al*. (2025) demonstrated that Random Forest achieved superior classification accuracy for supermarket product sales. Additional research supports these findings, showing the effectiveness of ensemble approaches such as Random Forest in improving prediction consistency and minimizing overfitting (Breiman, 2001; Arraudhah, 2025; Pratiwi & Nugroho, 2024; Pradana *et al*., 2024). Complementary evidence from Mahmudati *et al*. (2025) and Ramadhani *et al*. (2023) confirmed the applicability of Naïve Bayes for small-scale sales data, though its performance decreases with higher data complexity. Moreover, studies on clustering optimization using the Elbow Method (Umargono *et al*., 2020; Permadi *et al*., 2023) underline the importance of determining the optimal cluster count to improve segmentation precision. Together, these empirical insights reveal the growing emphasis on data mining and machine learning as tools for achieving accuracy in retail forecasting. Guided by these advancements, the present study applies K-Means for clustering and Naïve Bayes, Decision Tree, and Random Forest for classification to identify patterns within retail sales data. The approach aims to assess algorithmic accuracy and predictive reliability in characterizing sales performance, thereby offering a computational framework adaptable to evolving retail analytics.

2. Methodology

The research employed a quantitative approach that combines clustering and classification to analyze patterns in retail sales data. The workflow followed several sequential stages, as illustrated in Figure 1, encompassing dataset acquisition, data pre-processing, algorithm implementation, model evaluation, and result interpretation. The dataset was sourced from Kaggle, comprising 1,000 transaction entries with nine attributes, namely transaction ID, date, customer ID, gender, age, product category, quantity, price per unit, and total amount. Data integrity was verified by identifying missing values and ensuring consistency across attributes. Non-essential variables such as transaction identifiers, dates, and unit prices were excluded to minimize noise and dimensional redundancy, while categorical data—particularly gender and product category—were transformed into nominal values to support algorithmic compatibility. These preparatory steps align with recommendations by Chen *et al*. (2021) and Wei and Zeng (2021), who emphasize data normalization as a critical prerequisite for improving predictive accuracy in machine learning applications. For pattern discovery, the K-Means algorithm was used to cluster purchasing behaviors, with the optimal number of clusters determined through the Elbow Method, which calculates the sum of squared errors across varying cluster counts to locate the inflection point indicating optimal segmentation (Permadi *et al*., 2023; Umargono *et al*., 2020). This approach improves model efficiency by reducing iterative computation without compromising cluster cohesion. The clustering phase was followed by classification using three supervised algorithms: Naïve Bayes, Decision Tree, and Random Forest.

The Naïve Bayes classifier was selected for its probabilistic estimation under the assumption of feature independence (Juwita *et al*., 2022; Ramadhani *et al*., 2023; Mahmudati *et al*., 2025), which enables fast processing and stable prediction across discrete attributes. The Decision Tree method was implemented to model decision-making hierarchically, offering interpretability through node-based feature splits that facilitate the identification of determinant factors influencing purchasing decisions (Arfan & Paraga, 2024; Suranda & Nugroho, 2024). In contrast, the Random Forest algorithm was adopted as an ensemble learning technique that aggregates multiple decision trees to improve generalization and mitigate overfitting (Breiman, 2001; Pratiwi & Nugroho, 2024; Pradana *et al*., 2024; Arraudhah, 2025). Prior studies indicate that Random Forest often achieves superior predictive accuracy compared with single-tree or probabilistic models, particularly in high-dimensional and non-linear datasets (Firnanda *et al*., 2025; Arman *et al*., 2023). Model performance was evaluated by comparing clustering and classification outputs in terms of accuracy and consistency in predicting retail sales behavior. Evaluation metrics were based on standard data mining validation principles outlined by Awan *et al*. (2021), Li (2022), and Shetty and Shetty (2023), emphasizing cross-comparison to assess algorithmic robustness. Analytical operations were performed using Microsoft Excel and RapidMiner (Altair AI Studio 2025.0.1), software platforms that support structured data workflows and reproducible model testing environments (Andry *et al*., 2023; Niu, 2020). The methodological design thus integrates clustering, classification, and model assessment into a coherent analytical framework that facilitates retail sales forecasting through machine learning.

Figure 1. Research Stages

The research process begins with *data acquisition* from an open-source retail sales repository (Kaggle), followed by *data pre-processing* to ensure data validity and formatting consistency. The third stage involves *model construction*, encompassing clustering using the K-Means algorithm and classification using Naïve Bayes, Decision Tree, and Random Forest. Subsequently, *model evaluation* assesses accuracy and performance metrics to identify the best-performing algorithm. The final stage, *result interpretation*, focuses on translating analytical outcomes into actionable insights for retail decision-making.

3. Results

The analysis began with the clustering process using the K-Means algorithm, as the dataset contained 1,000 retail transaction records without predefined labels. To ensure algorithm compatibility, categorical attributes such as *Gender* and *Product Category* were transformed into numerical variables. After conversion, the dataset was processed using RapidMiner software, where the *Read CSV* operator imported the data, the *K-Means Clustering* operator executed the clustering, and the *Cluster Distance Performance* operator evaluated performance metrics through the Average Within-Centroid Distance and Davies–Bouldin Index (DBI). Figure 2 illustrates the clustering workflow in RapidMiner, showing a structured sequence from data input to algorithm evaluation. This visualization highlights the analytical flow that ensures the consistency and validity of clustering operations.

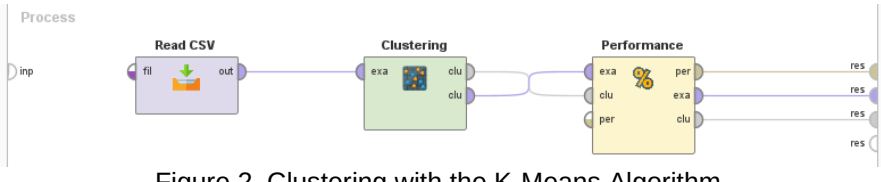

Figure 2. Clustering with the K-Means Algorithm.

To identify the most optimal number of clusters, several iterations were performed ranging from $k = 2$ to $k = 7$. The results, shown in Table 1, record the average within-centroid distances, revealing a significant drop from $k = 2$ to $k = 3$, which represents the largest interval. The decreasing distance values reflect tighter and more cohesive groupings as k increases, yet the most substantial differentiation occurs at $k = 3$, indicating the point of optimal clustering based on the Elbow Method.

Table 1. Average Within-Centroid Distance for 2–7 Clusters.

K	Avg. Distance	Interval
2	59,333.488	
3	25,272.635	34,060.853
4	12,551.201	12,721.434
5	6,364.303	6,186.898
6	3,981.910	2,382.393
7	1,645.360	2,336.550

The Elbow Method graph displayed in Figure 3 provides a visual confirmation of the tabular data. The x-axis represents the number of clusters (k), while the y-axis shows the average within-centroid distance. A sharp bend is visible at $k = 3$, marking the optimal cluster count. This result is further validated by the Davies–Bouldin Index value of 0.443, which indicates effective separation between clusters and minimal internal variance.

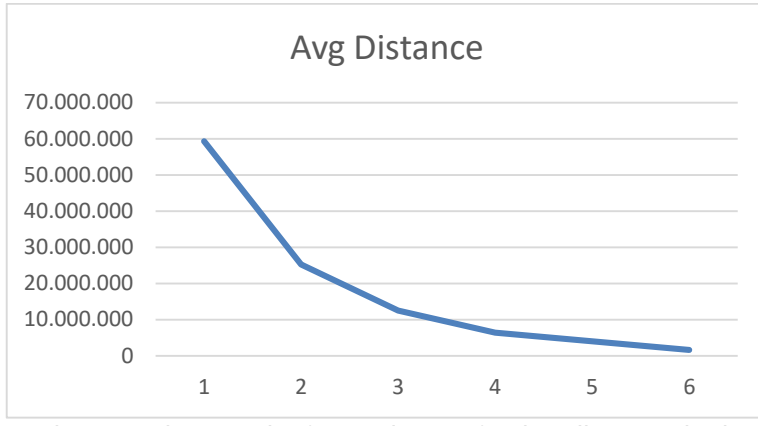

Figure 3. Line Graph of 2–7 Clusters for the Elbow Method.

Subsequent to the clustering process, data distribution across clusters was examined. Figure 4 shows the proportion of data points per cluster, where Cluster 1 contains the majority of transactions, Cluster 2 holds a moderate number, and Cluster 3 includes the smallest group. This pattern indicates the existence of three distinct customer groups with differing transaction frequencies.

Cluster Model

```
Cluster 0: 701 items
Cluster 1: 99 items
Cluster 2: 200 items
Total number of items: 1000
```

Figure 4. Number of Data Points per Cluster.

Further visualization of cluster behavior is presented in Figure 5, which plots *Age* on the x-axis and *Total Purchase Amount* on the y-axis. The resulting chart shows clear distinctions among customer groups, where younger consumers generally make lower-value purchases, while older consumers display significantly higher transaction values, suggesting maturity in purchasing power.

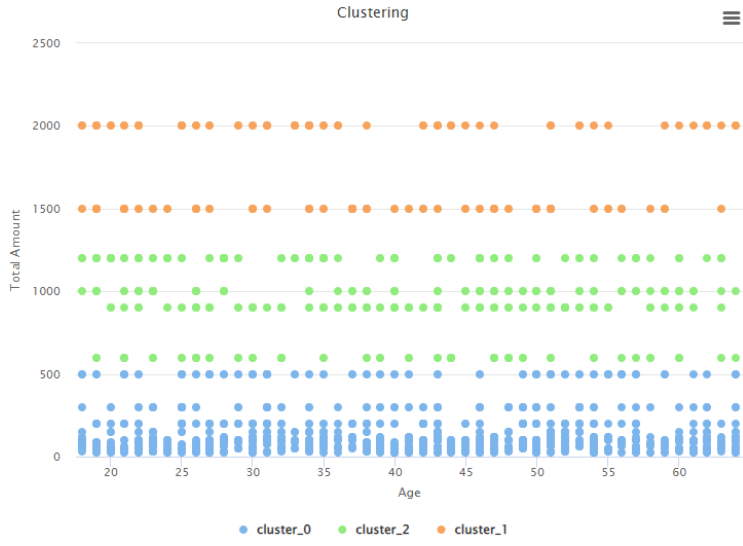

Figure 5. Visualization of Data Distribution by Cluster.

Following the clustering stage, the classification phase was conducted to predict future customer segments based on previously identified clusters. The dataset was divided into 70% training data (700 rows) and 30% testing data (300 rows). The first classification model employed the Naïve Bayes algorithm, which achieved an accuracy of 94.67%. Figure 6 depicts the workflow for this model in RapidMiner, while Figure 7 presents the classification results and confusion matrix, confirming the model's capability in predicting customer segments with relatively high precision.

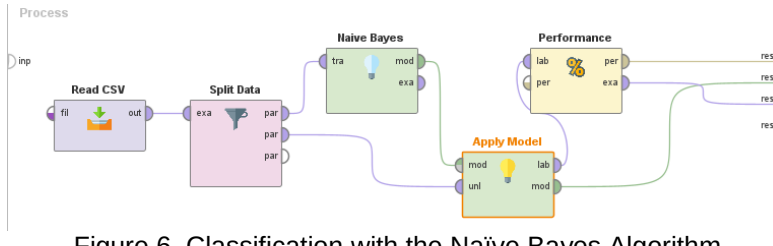

Figure 6. Classification with the Naïve Bayes Algorithm

accuracy: 94.67%

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	194	0	0	100.00%
pred. cluster_2	16	60	0	78.95%
pred. cluster_1	0	0	30	100.00%
class recall	92.38%	100.00%	100.00%	

Figure 7. Classification Accuracy of the Naïve Bayes Algorithm.

The second classification model utilized the Decision Tree algorithm, which produced an accuracy rate of 100%. The model's operational process is depicted in Figure 8, and the resulting tree structure in Figure 9 reveals that *Product Category* and *Total Amount* were the primary attributes influencing the final classification. The confusion matrix shown in Figure 10 confirms that all records were correctly classified without errors, demonstrating the algorithm's exceptional precision.

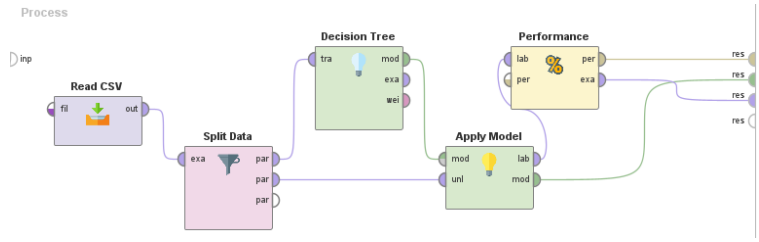

Figure 8. Classification with the Decision Tree Algorithm

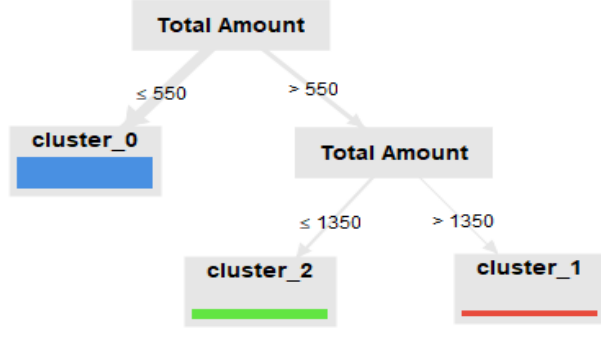

Figure 9. Visualization of the Decision Tree Model

accuracy: 100.00%

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	210	0	0	100.00%
pred. cluster_2	0	60	0	100.00%
pred. cluster_1	0	0	30	100.00%
class recall	100.00%	100.00%	100.00%	

Figure 10. Classification Accuracy with the Decision Tree Algorithm

The final classification model applied the Random Forest algorithm, which built 100 decision trees to generate a robust ensemble prediction. Figure 11 outlines the workflow, while Figures 12, 13, and 14 show examples of individual decision trees used in the voting process. The aggregated model achieved 100% accuracy, as summarized in Figure 15, which presents the final confusion matrix.

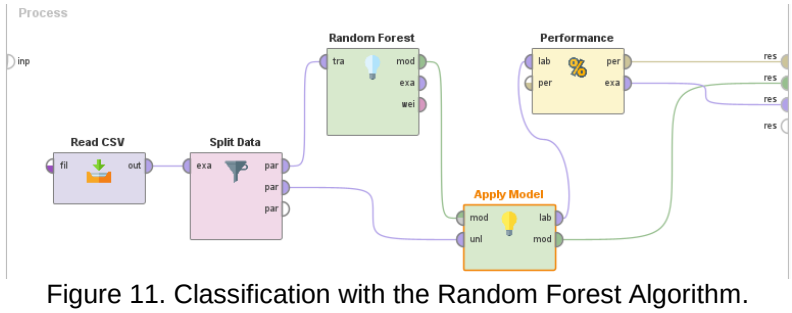

Figure 11. Classification with the Random Forest Algorithm.

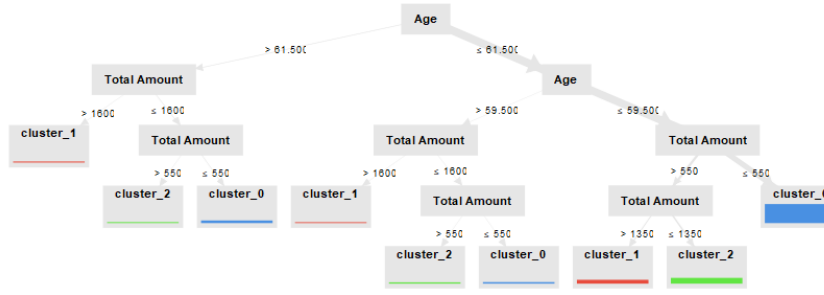

Figure 12. Sample Tree Visualization 1 from the Random Forest Algorithm.

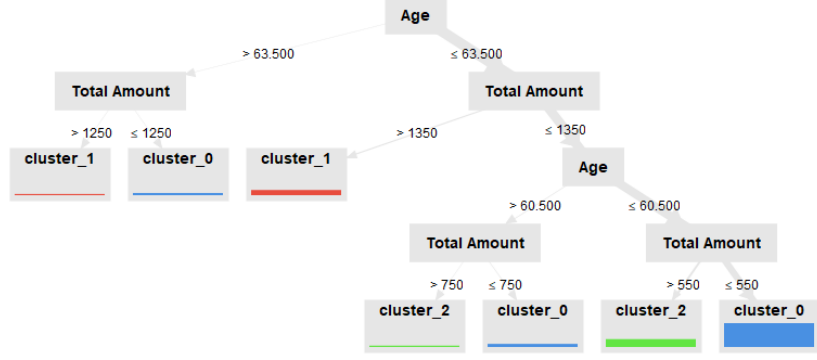

Figure 13. Sample Tree Visualization 2 from the Random Forest Algorithm.

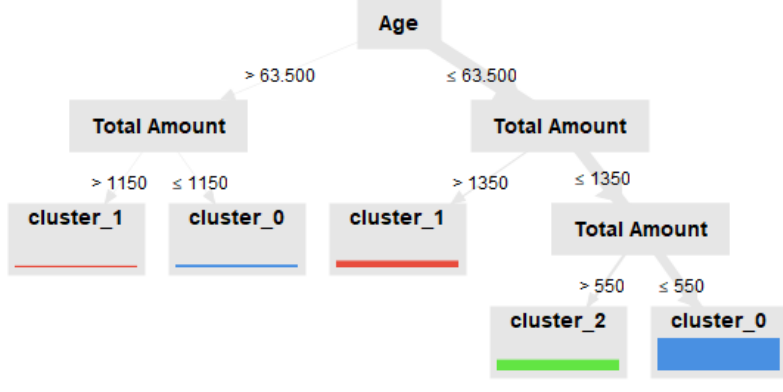

Figure 14. Sample Tree Visualization 3 from the Random Forest Algorithm.

accuracy: 100.00%

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	210	0	0	100.00%
pred. cluster_2	0	60	0	100.00%
pred. cluster_1	0	0	30	100.00%
class recall	100.00%	100.00%	100.00%	

Figure 15. Classification Accuracy with the Random Forest Algorithm.

A comparative summary of all three models—Naïve Bayes, Decision Tree, and Random Forest—is displayed in Figure 16, showing that while Naïve Bayes reached 94.67% accuracy, both Decision Tree and Random Forest achieved a perfect 100% accuracy rate, confirming the superior predictive performance of tree-based models in retail data classification.

4. Discussion

The findings demonstrate that clustering retail data into three groups successfully differentiates customers based on their purchasing behaviors and demographic attributes. The first cluster predominantly represents young customers who make frequent but low-value purchases, suggesting price-sensitive tendencies. The second cluster captures middle-aged customers with balanced spending behavior, while the third cluster consists of mature consumers with high spending, often associated with luxury or bulk purchases. This segmentation aligns with consumer behavior theories emphasizing the role of age and economic stability in retail spending. The optimal number of clusters identified through the Elbow Method and confirmed by a low Davies–Bouldin Index value reinforces the structural validity of the segmentation model. The classification phase substantiates the predictive capabilities of machine learning algorithms in analyzing retail data. The Naïve Bayes model, with 94.67% accuracy, performed efficiently despite its assumption of attribute independence, indicating its robustness even with moderately correlated features. However, the Decision Tree and Random Forest models both achieved perfect accuracy, demonstrating superior adaptability to nonlinear and categorical data structures. The Decision Tree's hierarchical structure offers high interpretability, enabling analysts to identify the attributes most responsible for differentiating customer segments. Meanwhile, the Random Forest's ensemble architecture enhances generalization and prevents overfitting, consistent with Breiman's (2001) theoretical framework and later empirical confirmations by Firnanda *et al*. (2025) and Pratiwi & Nugroho (2024).

When compared with prior research, the current findings confirm the dominance of tree-based models in retail analytics. Arman *et al*. (2023) reported the Decision Tree's superiority in fuel sales prediction, while Firnanda *et al*. (2025) emphasized that Random Forest consistently yields higher accuracy in multi-class retail classifications. From a managerial standpoint, these outcomes have meaningful implications. Businesses can employ these models to identify profitable customer segments, focus marketing initiatives on high-value clusters, and refine strategies for low-performing products. The integration of clustering and classification thus enables data-driven personalization, allowing companies to tailor campaigns according to age, gender, and product preferences. Despite achieving excellent results, this study acknowledges several limitations. The dataset was limited to 1,000 transactions, potentially constraining generalization. Only four attributes—*Gender*, *Age*, *Product Category*, and *Total Amount*—were used, as including *Quantity* and *Price per Unit* reduced model accuracy. Moreover, although the 70:30 train-test split yielded strong results, other ratios such as 60:40 and 80:20 produced consistent accuracy for Decision Tree and Random Forest but slight variations for Naïve Bayes. Future research should employ larger datasets and additional variables, such as time of purchase or customer loyalty metrics, to further enhance the scalability and predictive power of the proposed model.

5. Conclusion

The clustering analysis on the retail sales dataset using the K-Means algorithm determined that the optimal number of clusters, based on the Elbow Method, is three, with an Avg. within-centroid distance of 25,272.635 and a Davies-Bouldin Index of 0.443. The classification tests showed that the Naïve Bayes algorithm achieved an accuracy of 94.67%, whereas both the Decision Tree and Random Forest algorithms achieved identical accuracies of 100%. The performance of the models in this study surpassed the results of numerous previous studies. Specifically, within this research model, the best classification performance with the highest accuracy was obtained

using the Decision Tree and Random Forest algorithms. To anticipate overfitting, tests were conducted with alternatives of 4 and 6 features, as well as various training and testing data split options ranging from 60:40 to 90:10. The results showed a consistent accuracy of 100% for the Decision Tree and Random Forest algorithms, whereas the accuracy of the Naive Bayes algorithm varied. For future research, it is recommended to utilize other datasets with a larger number of rows and more attributes. Additionally, employing different algorithms and analytical tools is advised to enrich knowledge and insights in the fields of machine learning for big data and data mining.

References

Andry, J. F., Hartono, H., & Jo, J. (2023). Analysis and prediction of supermarket sales with data mining using RapidMiner. In N. I. Saragih, S. A. Salma, F. Dewi, D. Caesaron, M. Dellarosawati, D. Rachmawaty, F. D. Winati, D. Y. Bernanda, F. R. Wilujeng, & G. D. Rembulan (Eds.), *AIP Conference Proceedings* (Vol. 2693, Issue 1). American Institute of Physics Inc. https://doi.org/10.1063/5.0118725

Arfan, U., & Paraga, N. (2024). Perbandingan algoritma K-Means, Naïve Bayes, dan Decision Tree dalam memprediksi penjualan bahan bakar minyak: The comparison of K-Means, Naïve Bayes and Decision Tree algorithm in predicting fuel oil sales. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(4). https://doi.org/10.57152/malcom.v4i4.1566

Arman, S. A., Untari, R. T., & Erion, E. (2023). Implementasi data mining menggunakan metode Decision Tree dalam mempolakan penjualan pada showroom motor bekas. *Journal of Science and Social Research, 6*(2). https://doi.org/10.54314/jssr.v6i2.1313

Arraudhah, N. (2025). Peningkatan klasifikasi penjualan produk fashion di Sabhira Official dengan Random Forest. *Jurnal Dinamika Informatika, 14*(1).

Awan, M. J., Rahim, M. S. M., Nobanee, H., Yasin, A., Khalaf, O. I., & Ishfaq, U. (2021). A big data approach to Black Friday sales. *Intelligent Automation and Soft Computing, 27*(3), 785–797. https://doi.org/10.32604/iasc.2021.014216

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, J., Koju, W., Xu, S., & Liu, Z. (2021). Sales forecasting using deep neural network and SHAP techniques. *IEEE International Conference on Big Data, Artificial Intelligence, and Internet of Things Engineering (ICBAIE)*, 135–138. https://doi.org/10.1109/ICBAIE52039.2021.9389930

Firnanda, P. A., Shofwatillah, L., Rahma, F., & Fauzi, F. (2025). Analisis perbandingan Decision Tree dan Random Forest dalam klasifikasi penjualan produk pada supermarket. *Emerging Statistics and Data Science Journal, 3*(1). https://doi.org/10.20885/esds.vol3.iss.1.art2

Juwita, J., Safii, M., & Damanik, B. E. (2022). Naïve Bayes algorithm for predicting sales at the Pematang Siantar VJCakes store. *JOMLAI: Journal of Machine Learning and Artificial Intelligence, 1*(4). https://doi.org/10.55123/jomlai.v1i4.1674

Li, J. (2022). A feature engineering approach for tree-based machine learning sales forecast, optimized by a genetic algorithm-based sales feature framework. *IEEE International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 133–139. https://doi.org/10.1109/ICAIBD55127.2022.9820532

Mahmudati, R., Rohman, S., & Sa'adah, I. (2025). Sistem prediksi hasil laba penjualan di UNSIQ Mart menggunakan metode Naive Bayes. *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer, 4*(1). https://doi.org/10.55123/storage.v4i1.4849

Niu, Y. (2020). Walmart sales forecasting using XGBoost algorithm and feature engineering. *Proceedings of the International Conference on Big Data, Artificial Intelligence, and Software Engineering (ICBASE)*, 458–461. https://doi.org/10.1109/ICBASE51474.2020.00103

Permadi, V. A., Tahalea, S. P., & Agusdin, R. P. (2023). K-Means and Elbow Method for cluster analysis of elementary school data. *Progres Pendidikan, 4*(1), 50–57. https://doi.org/10.29303/prospek.v4i1.328

Pradana, R. Y., Nastiti, F. E., & Oktaviani, I. (2024). Machine learning pengklasifikasikan performa karyawan direct sales force kartu prabayar menggunakan metode Random Forest classifier. *JEKIN – Jurnal Teknik Informatika, 4*(3). https://doi.org/10.58794/jekin.v4i3.864

Pratiwi, G. E., & Nugroho, A. (2024). Implementasi metode Random Forest untuk klasifikasi penjualan produk sabun paling laris. *Jurnal Teknik Informasi dan Komputer (Tekinkom), 7*(2). https://doi.org/10.37600/tekinkom.v7i2.1610

Ramadhani, D., A'yuniyah, Q., Elvira, W., Nazira, N., Ambarani, I., & Intan, S. F. (2023). Analisa algoritma Naïve Bayes Classifier (NBC) untuk prediksi penjualan alat kesehatan. *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE), 3*(2). https://doi.org/10.57152/ijirse.v3i2.941

Shetty, S., & Shetty, S. (2023). Big Mart sales prediction using machine learning. In J. Stephen, P. Sharma, Y. Chaba, K. U. Abraham, P. K. Anooj, N. Mohammad, G. Thomas, & S. Srikiran (Eds.), *International Conference on Advanced Computing, Control, and Telecommunication Technology (ACT)* (pp. 1556–1561). Grenze Scientific Society. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174270894

Suranda, D. I., & Nugroho, A. (2024). Klasifikasi data penjualan untuk memprediksi tingkat penjualan produk menggunakan metode Decision Tree. *Jurnal Teknik Informasi dan Komputer (Tekinkom), 7*(1). https://doi.org/10.37600/tekinkom.v7i1.1269

Umargono, E., Suseno, J. E., & Gunawan, S. K. V. (2020). K-Means clustering optimization using the Elbow Method and early centroid determination based on mean and median formula. *Advances in Social Science, Education and Humanities Research*, 121–129. https://doi.org/10.2991/assehr.k.201010.019

Wahyudi, T., & Silfia, T. (2022). Implementation of data mining using K-Means clustering method to determine sales strategy in S&R Baby Store. *Journal of Applied Engineering and Technological Science, 4*(1), 93–103. https://doi.org/10.37385/jaets.v4i1.913

Wei, H., & Zeng, Q. (2021). Research on sales forecast based on XGBoost–LSTM algorithm model. *Journal of Physics: Conference Series, 1754*(1). https://doi.org/10.1088/1742-6596/1754/1/012191