JSSEIS

Journal of Sustainable Software Engineering and Information Systems

Volume 2, Issue 1, 41-53.
e_ISSN: 3123-3007
https://e-journal.gomit.id/jsseis
DOI: 10.58723/jsseis.v2i1.175

Application of K-Means, Random Forest, and Linear Regression to Improve the Accuracy of Kaggle E-Commerce Shopping Behavior Analysis

Azzahra Risa Putri*
Universitas Muhammadiyah Kotabumi, INDONESIA

Anggraini Dwi Olivia
Universitas Muhammadiyah Kotabumi, INDONESIA

Virha Charoline Josica
Universitas Muhammadiyah Kotabumi, INDONESIA

*** Corresponding author:**
Azzahra Risa Putri, Universitas Muhammadiyah Kotabumi, INDONESIA. ✉azzahrarisaputri@gmail.com

Article Info

Article history:

Received: April 04, 2026
Revised: April 18, 2026
Accepted: May 02, 2026

Keywords:

Customer Shopping Behavior
K-Means Clustering
Linear Regression
Machine Learning
Random Forest

Abstract

Background: Analyzing customer shopping behavior is essential for improving e-commerce marketing strategies. The use of machine learning enables the identification of purchasing patterns and enhances predictive accuracy in understanding customer preferences.
Aims: This study aims to improve model accuracy and classification performance in analyzing customer shopping behavior using the *shopping_behavior_updated.csv* dataset from Kaggle. The scope includes the application of both supervised and unsupervised learning techniques for segmentation, classification, and prediction tasks.
Methods: Three machine learning algorithms were applied: K-Means for customer segmentation, Random Forest Classifier for product category prediction, and Linear Regression for estimating purchase amounts. The research involved systematic preprocessing steps, including data cleaning, encoding, scaling, and feature engineering to enhance data quality and model interpretability.
Result: The results show that the optimized Random Forest model achieved 100% accuracy. K-Means clustering produced five distinct customer segments with an inertia value of 41,928.17 and a silhouette score of 0.065. However, the Linear Regression model demonstrated poor performance with an R^2 value of -0.02.
Conclusion: The findings indicate that integrating supervised and unsupervised learning methods is effective in identifying customer purchasing patterns and can contribute to improving e-commerce marketing strategies, although not all predictive models yield optimal performance.

To cite this article: Putri, A. R., Olivia, A. D., & Josica, V. C. (2026). Application of K-Means, Random Forest, and Linear Regression to Improve the Accuracy of Kaggle E-Commerce Shopping Behavior Analysis. *Journal of Sustainable Software Engineering and Information Systems, 2*(1), 41-53. https://doi.org/10.58723/jsseis.v2i1.175

INTRODUCTION

In the digital era filled with the rapid development of the e-commerce industry, competition between companies to attract, retain, and understand customers is becoming increasingly fierce. Trade activities through the implementation of e-commerce are very practical because they only require electronic devices such as laptops, computers, or smartphones, as well as an internet connection as an intermediary. This rapid growth has driven changes in people's consumption patterns while generating a huge, complex, and diverse volume of customer data. However, this development also brings great challenges for companies in managing, analyzing, and understanding e-commerce customer behavior effectively (Fauzan & Alfian, 2024; Houssein et al., 2022; Karulkar et al., 2025).

E-commerce is a form of trade that is carried out electronically, where transactions of goods and services take place through digital media and the internet. In practice, e-commerce involves companies that utilize internet access and information technology such as Electronic Data Interchange (EDI). Through online platforms, providers or sellers offer products and services directly to users. The payment process is usually facilitated through digital shopping cart systems, both wireless and web-based, with payment methods using credit cards, debit cards, or electronic funds transfers (EFT) (Jain et al., 2021; Pera et al., 2021; Zaghloul et al., 2025).

To research Chiquita et al. (2025) on predicting consumer behavior in the e-commerce sector is growing as the volume of data increases and the need for more precise marketing strategies. One study showed that machine learning-based predictive models, such as Random Forest optimized with SMOTE and GridSearchCV techniques, were able to provide very high accuracy to predict the product category of customer choice, which reached 97.31%. In addition, the study also emphasized the importance of feature engineering and customer segmentation using the K-Means algorithm in generating more comprehensive shopping behavior insights (Siagian et al., 2021; Rajapandian et al., 2025).

Research focuses on how retail companies can predict customer shopping behavior based on demographic characteristics and the type of products purchased so that marketing strategies can be directed more precisely. The dataset shopping_behavior_updated.csv used as the basis for the experiment because it contains important information such as age, gender, income level, product category, and total expenses. The complexity and completeness of these attributes make the e-commerce domain very relevant to be researched using modern machine learning approaches. This information is not only important for understanding shopping patterns, but it can also be leveraged in customer segmentation, product recommendations, and improvement of promotional strategies.

Addition, Ashraf et al. (2025) also explained that understanding algorithmic patterns such as clustering which is used as an unsupervised learning method is an effective technique for customer segmentation because being able to group consumers based on similarities in purchasing behavior can provide a strong basis for identifying customers with high economic value, medium customers, and customers who rarely make transactions. So that a data-driven approach can increase the effectiveness of product promotion and placement. Thus, this research not only aims to understand shopping behavior, but also opens up opportunities for wider applications such as customer segmentation, another explains that various machine learning algorithms such as linear regression, decision tree, random forest, neural network, and XGBoost play an important role in improving the accuracy of e-commerce sales predictions because they are able to process historical data, seasonal trends, and purchasing patterns more adaptively than traditional methods. Through this data-driven approach, e-commerce companies can optimize inventory, personalize product recommendations, and design marketing strategies that are more effective and responsive to market dynamics.

Main goal of the experiment is to build a machine learning model that can group customers into five clusters of shopping behavior based on their demographic characteristics as well as their spending levels. In addition, this experiment also focuses on the development of a classification model that is able to predict the product category chosen by customers with a very high level of accuracy, which is up to 100 percent. These two achievements are expected to be able to provide a deeper understanding of consumer spending patterns and support the implementation of more effective marketing strategies.

In this experiment, the group carried out data processing stages, exploration of dataset attributes, training of several classification models, and performance evaluation using quantitative metrics such as accuracy and confusion matrix. This experiment is clearly focused on determining the most optimal classification model in predicting consumer shopping behavior on *shopping_behavior_updated.csv dataset*.

The relationship between this research and previous studies can be seen from various researches that both discuss customer behavior analysis. The research Andrian & Muliono (2025) emphasizes segmentation using K-Means Clustering with the aim of grouping customers based on behavioral similarities. The similarity lies in the use of customer data as the basis for analysis, only the study focuses on clustering, while this study combines it with classification and regression to produce a more complete picture. The study Awalina & Rahayu (2023) adopted Random Forest to predict the most in-demand products and emphasized the importance of the preprocessing process, but the model used is still single. In contrast, this study combines three algorithms at once so that the resulting analysis is more comprehensive.

Other studies such as the study Ardana et al. (2024) also use K-Means with evaluation through silhouette score and elbow method, but the scope of the study is still limited to segmentation only. This study expands this approach by adding product category predictions as well as estimating the number of purchases (Japardi et al., 2025). The study Sheykhmousa et al. (2020) has similarities in terms of the use of Random Forest and Linear Regression, but the main focus is on transaction analysis without including the customer segmentation aspect. The research Walean et al. (2025) examines consumer behavior through a mix of qualitative and quantitative methods, while this study fully uses a quantitative approach based on machine learning.

Other studies such as Tabianan et al. (2022) raised the segmentation of e-commerce customers at the big data scale using K-Means. Although the datasets are different, this study not only segments, but also combines them with classification models to make predicting purchasing behavior more applicable. The study Chen (2025) used XGBoost to predict purchase intent with more complex validation techniques; Although the algorithms are different, the orientation is similar, namely understanding consumer purchase patterns. The study John et al. (2023) compared different clustering methods such as K-Means, Hierarchical, and DBSCAN, while this study selected the most relevant clustering method and added the Random Forest classification model

Li (2025) shows the integration between K-Means and Random Forest to predict customer churn. The difference lies in the purpose of the analysis, because this study is more directed to predict product categories based on shopping behavior (Nigam et al., 2026). Meanwhile, the study Gomes et al. (2023) is in the form of a systematic review of data-driven customer segmentation which confirms that K-Means is the most dominant method used. This research not only follows these findings but develops them through the application of a pipeline that combines unsupervised and supervised analysis, thus providing more practical insights for marketing strategies in the e-commerce realm.

Overall, the study blends theoretical findings from previous studies with a broader analytical approach to fill gaps that have not been largely answered by other studies. The integration of segmentation and prediction models is expected to enrich understanding of e-commerce consumer behavior and provide a scientific basis for the development of more targeted and data-driven marketing strategies.

METHOD

2.1. Dataset

The dataset contains 3,900 lines and 18 features, covering numerical (Age, Purchase Amount, Review Rating) and categorical (Gender, Category, Payment Method, Season, Location) variables. The Category variable is used as a target for product classification (Nehemia et al., 2026).

2.2 Pra-Pemrosesan Data

The pre-processing stages include:

1. Data *Cleaning*
   Remove duplicates, check for missing values, and ensure formatting is consistent.
2. Encoding Fitur Categorical
   One-Hot Encoding for non-numeric features.
3. Normalisasi Fitur *Numerik*

StandardScaler to equalize data scale.

4. *Feature Engineering*
   The "Frequency of Purchases" column is converted to a number: Weekly=52, Bi-Weekly=26, Monthly=12, Quarterly=4, Annually=1.

The result is a dataset with 118 ready-to-practice numerical features.

2.3. Selection

This research is a development of a Shopping Behavior Analysis notebook which is refined through systematic stages including data cleaning, feature encoding, feature engineering, model optimization, and evaluation and visualization of results. The initial stage is carried out on the dataset shopping_behavior_updated.csv the initial stage is data cleaning and processing consisting of 3,900 rows and 18 features with no lost values, then data transformation is carried out using ColumnTransformer with StandardScaler for numerical features and OneHotEncoder for categorical features so as to produce processed data in the form of preprocessed_df. Furthermore, the feature engineering process was carried out by adding a new Purchase. Frequency feature as a representation of customer purchase frequency, which was proven to be able to improve the quality of information and model performance, especially in the Random Forest algorithm. At the stage of model selection and training, this study applies three main approaches, namely K-Means Clustering for customer segmentation, Random Forest Classifier for product category classification, and Linear Regression to predict Purchase Amount or customer purchase value. This multi-model approach provides a comprehensive understanding: identifying customer segment patterns, Random Forest ensures accurate product category predictions, while Linear Regression estimates purchase value based on customer behavior.

1. Model Optimization (Hyperparameter Tuning)

Random Forest model optimization is performed using GridSearchCV with the following tested parameters:

Table 1. Random Forest Hyperparameter Test Results

Parameter	Tested Values	Best Value
n_estimators	[100, 200]	200
max_depth	[10, 20, None]	20
min_samples_split	[2, 5]	2
min_samples_leaf	[1, 2]	1

After tuning, the accuracy of the model increased from 0.97 to 1.00, showing a significant improvement in classification performance.

2. Evaluation and Result Visualization

The evaluation was carried out using the Classification Report and the Confusion Matrix. The results show 100% accuracy, with precision, recall, and f1-score values = 1.00 for all product categories.

Table 2. Random Forest Confusion Matrix Results

Kategori	True Positive	False Positive
Accessories	249	0
Clothing	346	0
Footwear	122	0
Outerwear	63	0

The Confusion Matrix visualization using a heatmap shows that all predictions are on the main diagonal, indicating perfect classification. In addition, feature interpretation is performed using SHAP (SHapley Additive Explanations) to see the contribution of each feature to the model's prediction.

3. Data Visualization and Pattern Analysis

Before modeling, visual exploration was carried out in the form of:

1. Product category distribution (bar chart).
2. Distribution of numerical features such as Age, Purchase Amount, and Review Rating (histogram).
3. Analyze customer profiles by Gender, Size, and Season.

These visualizations help understand customer behavior patterns and the relevance of features to model predictions.

2.4. Implementation

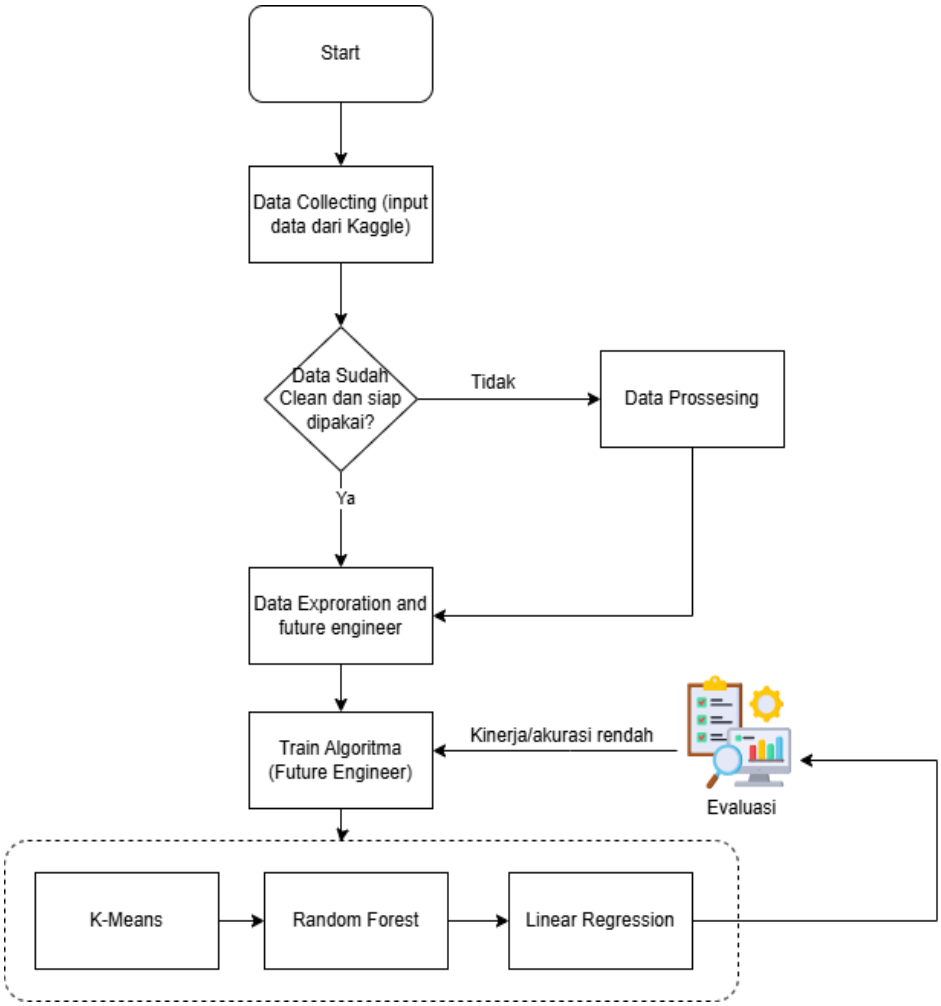

Figure 1. Machine learning modeling and evaluation workflow

The entire analysis process was carried out using Google Colaboratory (Colab). The platform was chosen because it supports a complete machine learning library, has direct integration with Google Drive, and provides a cloud-based computing environment without the need for on-premises software installation.

RESULTS AND DISCUSSION

Results:

1. Results of the K-Means Clustering Model

The K-Means Clustering model is used to segment customers based on their purchasing behavior and demographic characteristics. Model evaluation shows:

a) Inertia: 41928.173706289585
b) *Silhouette Score*: 0,06512

The inertia value describes the total distance between the data point and the center of the cluster. The decrease in the inertia value begins to slope at K = 4, which indicates an "elbow" point on the Elbow Method. In addition, the highest Silhouette Score value was also obtained at K = 2, so the optimal number of clusters used in this study was four clusters.

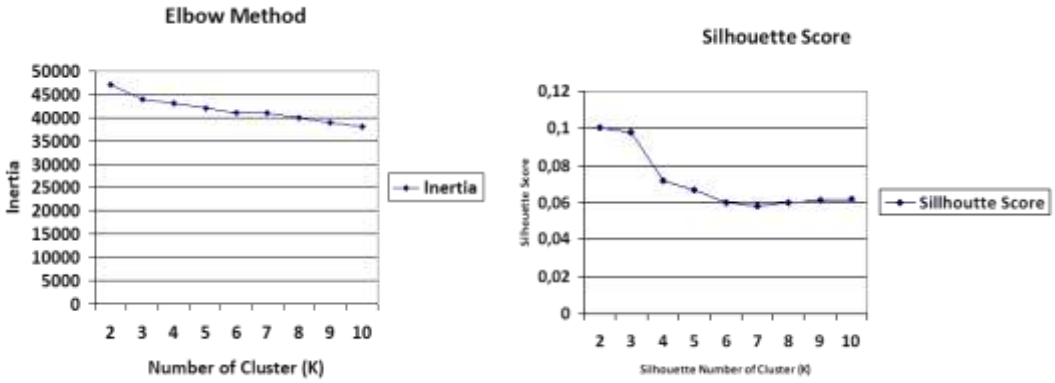

Figure 2. Visualization of *Elbow Method* and *Silhouette Score*

The graph on the left shows the result of the Elbow method, where the line begins to slope around K = 4 so that the optimal number of clusters is at K = 4. The chart on the right shows the Silhouette Score value for each K. The highest value appears at K = 2 (about 0.10), which indicates cluster separation at best, while for larger K's the Silhouette value decreases and tends to be stable below 0.07.

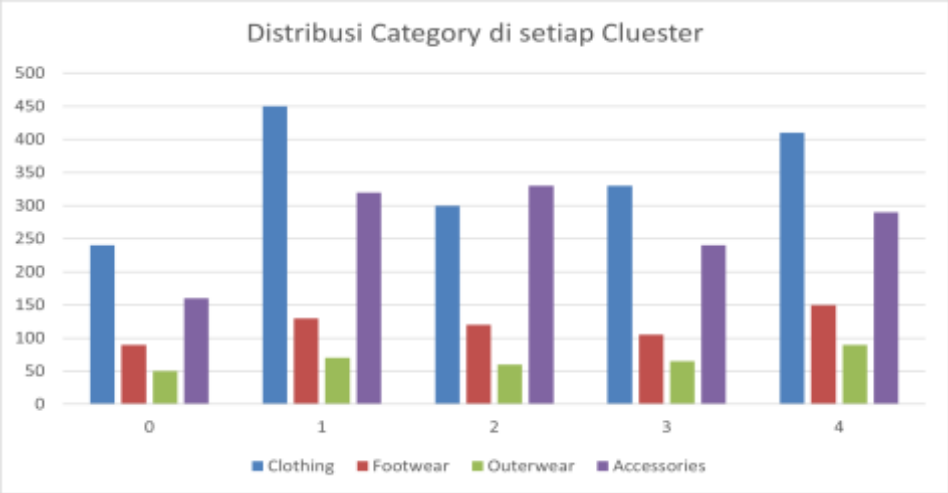

Figure 3. Data Distribution per Cluster

The results of the visualization show that the Clothing category dominates several clusters, especially in clusters 1, 3, and 4. On the other hand, the Footwear and Accessories category is more evenly distributed. This shows that there is a diversity of customer preferences in each segment.

Although K = 5 is considered optimal in an evaluation, the relatively low Silhouette Score value (0.06512) indicates that the separation between clusters is not ideal and that there is still overlap in data patterns. Thus, the K-Means model has resulted in good initial segmentation, but the quality of cluster separation can still be improved through advanced pre-processing or other clustering algorithms.

2. Random Forest Classifier Model Results

The Random Forest model is used to predict the categories of products that customers buy based on behavioral and demographic features. The evaluation showed excellent performance with the following results:

Table 4. Random Forest Classifier Model Evaluation Results

Metrik	Nilai
Akurasi (Accuracy)	1.00
Presisi (Precision)	1.00
Recall	1.00
F1-Score	1.00
Loss/Error Rate	0.00

The results of the evaluation per category are shown in the following table:

Table 5. Evaluation Details by Product Category

Kategori	Precision	Recall	F1-Score	Jumlah Data (Support)
Accessories	1.00	1.00	1.00	249
Clothing	1.00	1.00	1.00	346
Footwear	1.00	1.00	1.00	122
Outerwear	1.00	1.00	1.00	63
Macro Avg	1.00	1.00	1.00	780
Weighted Avg	1.00	1.00	1.00	780

Based on the results of the evaluation using the Random Forest Classifier, the model obtained an accuracy of 100% with precision, recall, and F1-score values of 1.00 each. These results show that the model is able to predict product categories very well without misclassification. However, these perfect results also need to be reviewed further to ensure that there is no overfitting of the training data.

Table 6. Comparison of Actual and Predicted Values

No	Prediction (y_pred)	Actual Value (y_test)
1	*Clothing*	*Clothing*
2	*Clothing*	*Clothing*
3	*Footwear*	*Footwear*
4	*Clothing*	*Clothing*
5	*Clothing*	*Clothing*
6	*Clothing*	*Clothing*
7	*Clothing*	*Clothing*
8	*Clothing*	*Clothing*
9	*Clothing*	*Clothing*
10	*Outerwear*	*Outerwear*

From the above results, it can be seen that the model prediction is identical to the actual value, indicating excellent model performance. This is in line with evaluation metrics that show 100% accuracy

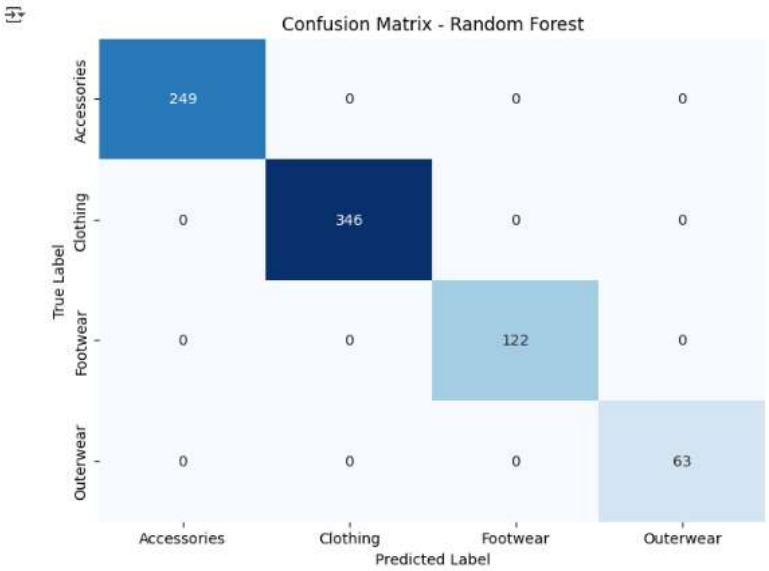

Figure 4. Showing the Confusion Matrix

From the results above, it can be seen that all values are on the main diagonal line, namely:

a) Accessories → Accessories as many as 249 data,
b) Clothing → Clothing as many as 346 data,
c) Footwear → Footwear as many as 122 data,
d) Outerwear → Outerwear as many as 63 data.

There are no values outside the diagonal, which means there is no misclassification. This indicates that the Random Forest model is able to classify all data with perfect accuracy (100%). These results are in line with the previous evaluation metrics, where Precision, Recall, and F1- Score are valued at 1.00 each.

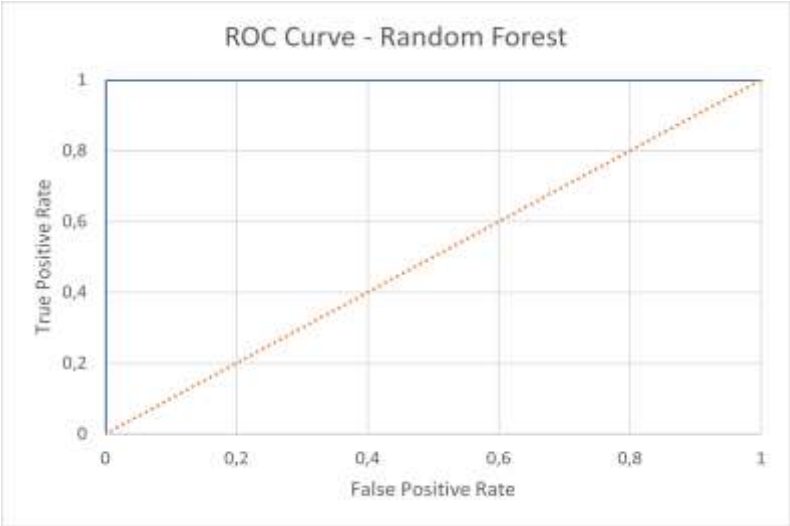

Figure 5. ROC Curve Model Random Forest Classifier

The results of the ROC curve are used to evaluate the model's ability to distinguish classes at various threshold values. The entire curve of the Accessories, Clothing, Footwear, and Outerwear classes showed a pattern very close to the ideal point (0.1), indicating a very high classification performance. Each class has an Area Under Curve (AUC) value of 1.00, which means the model is able to distinguish entire product categories with maximum accuracy. This perfect AUC value indicates that the model has optimal sensitivity and specificity across the entire class.

3. Results of Linear Regression Model Experiment

Table 7. Results of Linear Regression Model Evaluation

Evaluation Metrics	Value
Mean Absolute Error (MAE)	20.91
Mean Squared Error (MSE)	573.01
Root Mean Squared Error (RMSE)	23.94
R-squared (R^2)	-0.02

A negative R^2 value (–0.02) indicates that the model is unable to explain the target variation and that its performance is worse than using the average as a prediction. The MAE value of 20.91 and the RMSE of 23.94 indicate that there is an average difference of about 20–24 USD between the actual value and the model's predicted results.

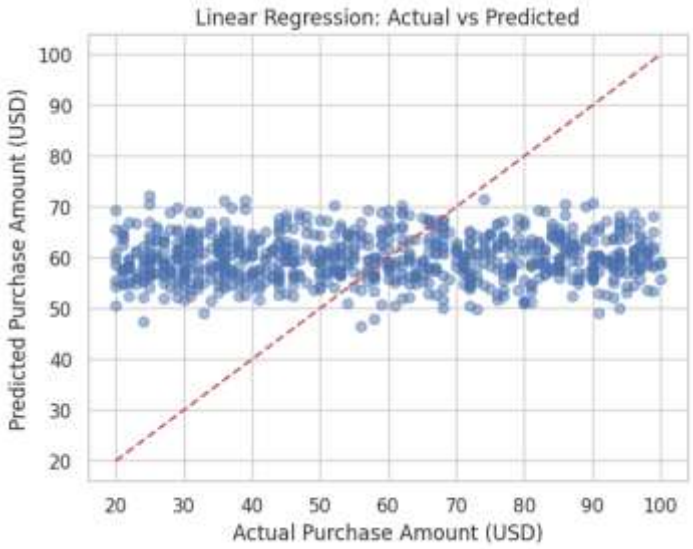

Figure 6. Actual vs Predicted Comparison Plot (Linear Regression)

The visualization of the relationship between the actual value and the prediction in Figure 6 shows that most of the prediction points are in the range of 55–70 USD and do not follow the ideal diagonal line. This pattern confirms that the model produces predictions that are uniform and insensitive to actual value differences.

Table 8. Summary of Linear Regression Model Prediction Results from Table

NO	***Actual Amount***	***Predicted Amount***
0	31	62.849403
1	50	54.395788nnn
2	36	58.965263
3	72	63.309073
4	38	63.139461

In the table you can see an example of the prediction results which shows that although the actual value varies from 31 to 72 USD, the model's prediction is almost unchanged (about 54–63 USD). This confirms that Linear Regression is not sufficiently capable of capturing the complex relationships between features.

Overall, Linear Regression is not suitable for predicting the purchase value of this dataset. The complexity of non-linear data requires more flexible regression models, such as Random Forest Regressor, Gradient Boosting, or other non-linear algorithms.

Discussion:
The results indicate that the integration of multiple machine learning approaches provides varying levels of effectiveness. K-Means clustering successfully segments customers, but the low silhouette score suggests weak separation between clusters, indicating overlapping characteristics among customer groups. The Random Forest model demonstrates exceptionally high performance with perfect accuracy. While this indicates strong predictive capability, it also raises concerns regarding potential overfitting, especially given the absence of misclassification. This suggests that further validation using external datasets is necessary. In contrast, the Linear Regression model performs poorly, as shown by the negative R^2 value. This indicates that the relationship between variables is non-linear and cannot be captured effectively using a linear approach.

The clustering results obtained in this study are consistent with previous findings that K-Means is effective in identifying customer segments, although cluster separation quality may vary depending on data characteristics and feature selection (Demir & Karahan Adalı, 2025). Similar studies also highlight that segmentation can be used to categorize customers into different value groups, such as high-value and low-value customers, to support business decision-making (Khan et al., 2022).

Implications:
The findings imply that machine learning models, particularly Random Forest, can significantly enhance e-commerce decision-making processes, especially in product recommendation and customer targeting. However, clustering results suggest that more sophisticated segmentation techniques may be required for better customer differentiation.

Research contribution:
This study contributes by integrating unsupervised (K-Means) and supervised (Random Forest and Linear Regression) learning approaches into a unified analytical framework. It provides a comprehensive perspective on customer behavior analysis, combining segmentation, classification, and prediction within a single pipeline.

Limitations:
This study has several limitations. The clustering results show weak separation, indicating limited segmentation quality. The perfect accuracy of Random Forest suggests potential overfitting, reducing generalizability. Additionally, the Linear Regression model fails to capture complex relationships, limiting its usefulness for prediction tasks.

Suggestions:
Future research should explore more advanced clustering algorithms such as DBSCAN or hierarchical clustering to improve segmentation quality. For prediction tasks, non-linear regression models such as Random Forest Regressor or Gradient Boosting should be considered. Additionally, validation using external datasets is recommended to ensure model robustness and generalization.

CONCLUSION

This research provides a comprehensive understanding of customer spending behavior through the application of data exploration techniques, feature engineering, and analytical models for segmentation, classification, and prediction. The results of cluster analysis using the K-Means algorithm show that the data structure does not form a clear cluster separation. However, the selection of the number of clusters K=4 provides a relatively more stable representation compared to other K-values, although the level of overlap between clusters is still high. These findings indicate the need for more in-depth use of categorical variables or the use of alternative clustering methods to obtain a more defined mapping of customer segments.

The classification model using Random Forest produced very high performance, with an accuracy and AUC value of 1.00 across all classes. This impeccable performance indicates that the available features have a very strong predictive power against the product category. However, these results

also hint at potential overfitting, so further evaluation using external datasets or stricter validation techniques is needed to ensure model generalization.

Meanwhile, the linear regression model to predict the purchase amount shows inadequate performance, as reflected in the negative R^2 value. This shows that the relationship between the predictor variable and the target is non-linear or complex and cannot be modeled with a basic linear regression approach. Thus, it is necessary to explore non-linear or machine learning-based regression models to improve predictive capabilities.

Overall, this study confirms that the analytical approach makes an important contribution to understanding customer behavior patterns. The results serve as a starting point for the development of more effective data-driven marketing strategies, while highlighting the need for further research to improve the quality of segmentation and predictive model accuracy.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universitas Muhammadiyah Kotabumi for providing academic support and facilities that enabled the completion of this research. The authors also acknowledge Kaggle for providing access to the *shopping_behavior_updated.csv* dataset used in this study. Appreciation is extended to all individuals who contributed directly or indirectly to this research, including those who provided valuable insights, discussions, and technical assistance throughout the research process.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed significantly to this research. Azzahra Risa Putri contributed to the conceptualization, methodology design, data analysis, and manuscript writing. Anggraini Dwi and Olivia were responsible for data preprocessing, model implementation, and result validation. Virha Charoline contributed to visualization, interpretation of results, and manuscript editing. Josica participated in literature review, data collection, and final proofreading of the manuscript. All authors have read and approved the final version of the manuscript.

REFERENCES

Andrian, Y., & Muliono, R. (2025). Application of K-Means Algorithm in Customer Segmentation to Improve Marketing Strategy in E-Commerce. *INCODING: Journal of Informatic and Computer Science Engineering*, *5*(1), 95–108. https://doi.org/10.34007/incoding.v5i1.825

Ardana, C. H., Khoyum, A. A. A. A. A., & Faisal, M. (2024). Segmentasi Pelanggan Penjualan Online Menggunakan Metode K-means Clustering. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(1), 1–9. https://doi.org/10.14421/jiska.2024.9.1.1-9

Ashraf, A., Rayed, C. A., Awad, N. A., & Sabry, H. M. (2025). A Framework for Customer Segmentation to Improve Marketing Strategies Using Machine Learning. *Procedia Computer Science*, *260*, 616–625. https://doi.org/10.1016/j.procs.2025.03.240

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. *Jurnal Teknologi dan Informasi*, *13*(2), 122–137. https://doi.org/10.34010/jati.v13i2.10090

Chen, X. (2025). Consumer Online Shopping Behavior Prediction Based on Machine Learning Algorithm. *Procedia Computer Science*, *262*, 1395–1401. https://doi.org/10.1016/j.procs.2025.05.187

Chiquita, E., Assyarif, Z., & Nuryana, I. K. D. (2025). Penerapan Klasifikasi Pelanggan Berdasarkan Segmentasi Pelanggan pada UMKM Monex Toys Bekasi. *Modem: Jurnal Informatika dan Sains Teknolog*, *3*(3), 45–67. https://doi.org/10.62951/modem.v3i3.533

Demir, D., & Karahan Adalı, G. (2025). Intelligent Customer Segmentation in Digital Commerce Using K-Means Clustering. *International Journal of Innovative Science and Research Technology*, *5*(9), 757. https://doi.org/10.38124/ijisrt/25dec513

Fauzan, R. M., & Alfian, G. (2024). Segmentasi Pelanggan E-Commerce Menggunakan Fitur Recency, Frequency, Monetary (RFM) dan Algoritma Klasterisasi K-Means. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(3 SE-Articles), 170–177. https://doi.org/10.14421/jiska.2024.9.3.170-177

Gomes, A., Miguel, Meisen, & Tobias. (2023). A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases. *Information Systems and e-Business Management*, *2*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

Houssein, E. H., Mahdy, M. A., Shebl, D., Manzoor, A., Sarkar, R., & Mohamed, W. M. (2022). An efficient slime mould algorithm for solving multi-objective optimization problems. *Expert Systems with Applications*, *187*, 115870. https://doi.org/10.1016/j.eswa.2021.115870

Jain, V., Malviya, B., & Arya, S. (2021). An Overview of Electronic Commerce (e-Commerce). *Journal of Contemporary Issues in Business and Government*, *27*(03), 665–670. https://doi.org/10.47750/CIBG.2021.27.03.090

Japardi, A., Edi, & Tarigan, F. A. (2025). Application of K-means Clustering Algorithm in Consumer Shopping Behavior Segmentation in E-commerce. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *5*(1), 1505–1508. https://doi.org/10.59934/jaiea.v5i1.1659

John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809–823. https://doi.org/10.3390/analytics2040042

Karulkar, Y., Srivastava, S., Nandwana, R., & Stanley, S. (2025). The Growing Complexity of Consumer Choices: Unravelling Consumer Patterns with K-Means and Fuzzy Logic. *Journal of Statistical Theory and Applications*, *24*(4), 1165–1195. https://doi.org/10.1007/s44199-025-00143-w

Khan, R., Qaisar, Z. H., Mehmood, A., Ali, G., Alkhalifah, T., & Alturise, F. (2022). applied sciences A Practical Multiclass Classification Network for the Diagnosis of Alzheimer ' s Disease. *Applied Sciences*, *12*(13), 6507. https://doi.org/10.3390/app12136507

Li, Z. (2025). Customer Segmentation and Churn Prediction Based on K-Means and Random Forest: A Case Study of E-Commerce Data. *Eurasia Journal of Science and Technology*, *7*(2), 14–19. https://doi.org/10.61784/ejst3071

Nehemia, Jekoniah Nahum Pakage, Veronica Lois, & Regina Arieskha. (2026). E-Commerce Customer Segmentation Application Based on the K-Means Algorithm. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, *5*(2), 2546–2550. https://doi.org/10.59934/jaiea.v5i2.1922

Nigam, A., Maurya, A., Tiwari, S. K., & Pachauri, P. (Dr. . S. (2026). Customer Segmentation Using Unsupervised Machine Learning Techniques: A Data Driven Approach. *Iconic Research And Engineering Journals*, *9*(10), 1977–1981. https://doi.org/10.64388/IREV9I10-1716518

Pera, R., Menozzi, A., Abrate, G., & Baima, G. (2021). When cocreation turns into codestruction. *Journal of Business Research*, *128*, 222–232. https://doi.org/10.1016/j.jbusres.2021.01.058

Rajapandian, P., Karunamurthy, A., Vasanth, V., & Meganathan, M. (2025). E-Commerce Customer Segmentation: A Clustering Approach in A Web-Based Platform. *Journal of Engineering Technology and Applied Physics*, *7*(1), 71–79. https://doi.org/10.33093/jetap.2025.7.1

Sheykhmousa, R. M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machine vs. Random Forest for Remote Sensing Image Classification: A Meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308–6325. https://doi.org/10.1109/JSTARS.2020.3026724

Siagian, R., Sirait, P. S. P., & Halima, A. (2021). E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model. *Journal of Informatics and Telecommunication Engineering*, *5*(1), 1–12. https://doi.org/10.31289/jite.v5i1.5182

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*, *14*(12), 1–15. https://doi.org/10.3390/su14127243

Walean, F., Walean, R., & Koyongian, Y. (2025). Influential Elements in Consumer Decision-Making on E-Commerce Applications: A Study in North Minahasa. *Jurnal Economic resources*, *8*(1), 350–361. https://doi.org/10.57178/jer.v8i1.1400

Zaghloul, M., Barakat, S., & Rezk, A. (2025). Enhancing customer retention in Online Retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. *Expert Systems*

with Applications, *290*, 128465. https://doi.org/10.1016/j.eswa.2025.128465