Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2247~2253
e-ISSN: 2548-6861

Factor Analysis Influencing Review Scores on E-Commerce Platforms Using Machine Learning

Meylani Utari [1]*, Kanda Januar Miraswan [2], Anna Dwi Marjusalinah [3]*****
[*]Computerized Accounting, Faculty of Computer Science, Universitas Sriwijaya
[**]Informatics Engineering, Faculty of Computer Science, Universitas Sriwijaya
[***]Informatics Management, Faculty of Computer Science, Universitas Sriwijaya
meylani@ilkom.unsri.ac.id [1], kandajm@ilkom.unsri.ac.id [2], annadwimarjusalinah@unsri.ac.id [3]

Article Info

Article history:

Received 2026-03-05
Revised 2026-04-28
Accepted 2026-05-05

Keyword:

Customer Satisfaction,
E-Commerce,
Machine Learning,
Random Forest,
Review Score.

ABSTRACT

In recent years, the rapid growth of e-commerce has made customer reviews an important indicator of product quality, service performance, and customer satisfaction. Review scores play a crucial role in influencing purchasing decisions and evaluating overall shopping experiences on e-commerce platforms. This research aims to analyze the main factors influencing customer review scores by integrating logistics, transaction, and product-related variables using a machine learning approach. The dataset consists of various e-commerce transaction attributes, including delivery information, payment details, and product characteristics. A Random Forest classifier is employed to predict customer review scores and to identify dominant influencing factors through feature importance analysis. The results show that logistics-related factors, particularly delivery time, are the most influential variables affecting review scores, followed by payment value, freight value, and product price. This study also emphasizes the significance of understanding how models work and their real-world applications, offering useful guidance on enhancing logistics efficiency, ensuring clear transaction records, and maintaining high standards of product information. Product attributes such as description length, weight, and physical dimensions also contribute significantly to customer satisfaction. By combining predictive capability and interpretability, this research provides valuable insights for sellers and e-commerce platform managers to improve service quality, optimize logistics performance, and enhance customer satisfaction.

I. INTRODUCTION

The growth of e-commerce in Indonesia is increasing quickly as more people use the internet and digital tools. The fast growth of online shopping has led to more use of data analysis to improve competition and connect better with customers [1] . Online shopping has become a regular part of everyday life, including buying everyday items and lifestyle products. In the e-commerce world, customer reviews are very important because they show how good the service is, how efficient the shipping is, and how happy customers are with their experience. A review score or rating is a number that shows how customers feel about a product after they get it. This score plays a big role in helping other customers decide what to buy. User-created content like online reviews is commonly seen as a helpful tool for learning about what customers want and helping them make buying choices [2].

Kotler et.al. in 2019 say that customer satisfaction comes from comparing what customers expect with what they actually get [3]. If the shopping experience is good or better than expected, customers tend to leave positive reviews. But if the product doesn't meet their expectations, the reviews are usually negative. In e-commerce, customer satisfaction is greatly affected by logistics related factors such as delivery speed, shipping costs, delivery accuracy, and product packing quality. Recent studies show that delivery time is one of the most dominant factors influencing customer review scores on e-commerce platforms [4] [5]. Service reliability and how

quickly a service responds are still key factors when judging how well a digital service works [6].

In addition to delivery issues, other factors related to the transaction, such as product price, the value of the payment, and the payment methods used, also shape how customers feel. Features of the product like its size, weight, how accurate the description is, and how many photos are available also influence what customers expect. Rachmiani et. al. in 2024 highlight that complete and accurate product information significantly affects customer satisfaction and online review behavior. If the product information is missing or wrong, it can make customers unhappy and lead to bad reviews [7]. In addition, shipping cost and transaction value have been shown to significantly influence customer perceptions of fairness and service quality in e-commerce transactions [8] [9] . The value a customer feels from the information shared during a transaction greatly influences how they see the product or service and their actions after buying it [10].

As technology improves, machine learning is increasingly used to predict review scores, identify patterns in customer behavior, and determine the key factors influencing customer satisfaction. Machine learning approaches are particularly effective in handling large datasets and capturing complex, non-linear relationships among multiple variables [11] [12] [13]. Tools such as the Random Forest algorithm help model complex interactions between variables and determine which factors have the greatest impact on customer ratings [14].

However, most previous studies tend to analyze logistics, transaction, and product factors separately or rely on conventional statistical approaches, limiting their ability to capture combined and non-linear effects. This study's contribution is in combining logistics, transaction, and product factors into a single machine learning model, with a focus on making the results clear and easy to understand. Although Random Forest is commonly applied, this study emphasizes the integration of multiple variables and provides a thorough analysis of their effects. Future research should consider comparing this method with other techniques like Gradient Boosting or XGBoost in order to achieve a more comprehensive assessment [13]. Therefore, the novelty of this research lies in the integration multiple dimensions of logistics, product, and transaction-related variables within a single Random Forest-based machine learning framework to predict customer review scores. In addition to prediction, this research employs feature importance analysis to identify dominant factors influencing customer satisfaction, providing interpretable insights for e-commerce [15] [16]. More attention is being given to explainable artificial intelligence, which shows how important it is to have machine learning models that are easy to understand when making important decisions [17] [18].

Based on this, this research aims to find out the main factors that influence review scores using Machine Learning. It also wants to give suggestions to sellers and e-commerce platforms to help them improve their services.

II. METHOD

The research uses a quantitative approach that combines data mining and machine learning. The analysis is about creating models to predict customer review scores on online shopping sites and finding out which factors are most important for customer satisfaction. Machine learning was chosen because it works well with big datasets and can find complicated, non-linear relationships.

This research uses the CRISP-DM (Cross Industry Standard Process for Data Mining) method as the research approach because it can systematically explain the data analysis process, which consists of several stages as shown in Figure 1.

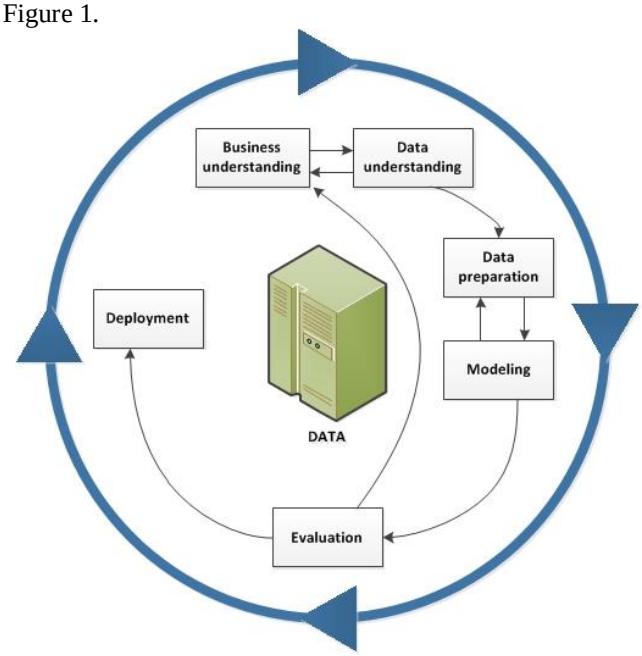

Figure 1. CRISP-DM Method [19]

A. Business Understanding

In this stage, the problem being addressed is how to predict customer review scores on e-commerce platforms and identify the key factors influencing customer satisfaction levels. The results of this stage include an explanation of the research objectives, the selection of relevant variables, and the determination of a machine learning approach as the method of analysis.

B. Data Understanding

The data used in this research is secondary data derived from the Brazilian E-Commerce Public Dataset by Olist from Kaggle. The dataset includes over 100,000 transaction records, with 41 variables that provide details on logistics, transaction data, product information, and customer reviews. The review score is evaluated on an ordinal scale ranging from 1 to 5, but in this study, it is considered as a multi-class classification task, which is a standard method in customer satisfaction analysis [20][21].

The chosen variables capture important elements of the customer experience, such as logistics performance, transaction value, and the quality of product information. These aspects are consistent with well-known theories related to customer satisfaction and perceived value [3][10].

TABLE I
RESEARCH VARIABLE TABLE

No	Variable Type	Variable Name	Code	Scale
1	Dependent	Review Score	Y	Ordinal (1–5)
2	Independent (Logistics)	Delivery Time	X1	Ratio
3	Independent (Logistics)	Freight Value	X2	Ratio
4	Independent (Logistics)	Estimated Delivery	X3	Ratio
5	Independent (Product)	Product Weight	X4	Ratio
6	Independent (Product)	Product Length	X5	Ratio
7	Independent (Product)	Product Height	X6	Ratio
8	Independent (Product)	Product Width	X7	Ratio
9	Independent (Product)	Product Description Length	X8	Ratio
10	Independent (Product)	Product Photos Quantity	X9	Ratio
11	Independent (Payment)	Price	X10	Ratio
12	Independent (Payment)	Payment Value	X11	Ratio
13	Independent (Payment)	Payment Installments	X12	Ratio
14	Independent (Categorical)	Product Category Name	X13	Nominal
15	Independent (Categorical)	Payment Type	X14	Nominal

C. Data Preparation

This stage begins with a comprehensive data preprocessing process to ensure data quality and reliability of the model. Missing values in numerical variables were handled using median imputation, while categorical variables were treated using mode imputation. Outliers were identified using the Interquartile Range (IQR) method and handled through capping to reduce the influence of extreme values without removing important observations.

Furthermore, categorical variables such as product category and payment type were transformed into numerical representations using one-hot encoding to make them suitable for machine learning algorithms.

This stage includes cleaning the data, handling missing data, creating new features, converting categorical variables into a processable form, and splitting the data into training (80%) and testing (20%) sets.

To prevent data leakage, special care was taken in handling temporal variables like delivery time. These variables were created solely based on information that was either available prior to or at the time the review was submitted, making sure that no data from after the review was used in training the model [11][13].

The dataset has a mix of review scores that isn't balanced, which is something that often happens in classification tasks and can make the predictions favor the most common classes [20] [21]. The research didn't use any methods to balance the data; however, the impact of this imbalance was taken into account during evaluation and mentioned as a limitation, following common approaches in recent machine learning research that focuses on making models easier to understand [21].

D. Modelling

The modelling stage is performed using the Random Forest Classifier algorithm to predict customer review scores. The selection of Random Forest is based on its capability to handle datasets with a large number of variables and capture complex, non-linear relationships among features, which are commonly found in e-commerce transaction data [7]. In addition, ensemble-based models such as Random Forest have been shown to provide robust predictive performance and lower overfitting risk compared to single learning models [4].

Furthermore, Random Forest provides feature importance analysis, enabling the identification of the most influential factors affecting customer review scores and supporting model interpretability in customer satisfaction analysis [8]. Interpretability is now a key need in today's machine learning systems to make sure decisions are clear and trustworthy [17]. Recent studies highlight that being able to understand how machine learning models make decisions is very important, especially when these models are used to help make decisions in e-commerce [15] [16]. The model is trained using training data with a carefully determined number of trees to ensure stable and optimal prediction performance.

No extensive hyperparameter tuning was performed in this research. This method is often used in research where the focus is on making sure the models are both stable and easy to understand, rather than just trying to make them as efficient as possible [13]. The Random Forest model was trained with a set number of trees to keep the model stable and easy to understand. The number of trees was carefully chosen to make sure the model works well and gives reliable predictions.

E. Evaluation

The model is evaluated using several metrics such as accuracy, confusion matrix, precision, recall, and F1-score. In this study, F1-score is emphasized as a primary evaluation metric because it provides a balance between precision and recall, making it more suitable for imbalanced classification problems where minority classes need special attention [20][21]. These evaluation metrics are commonly used to check how well a classification model is working, especially when the classes are not [20] [21]. Feature importance analysis is used to identify the main factors influencing review scores. Feature importance helps explain how a model makes decisions and makes it easier to understand and choose the best options in machine learning [15] [16].

F. Deployment

The final stage is understanding the results obtained, determining strategies to improve customer satisfaction, and drawing relevant conclusions for sellers and e-commerce platforms. Such data-driven insights are commonly used to

help make better decisions and improve services in e-commerce settings [5] [22].

III. Results And Discussion

A. Dataset Description and Initial Analysis

The dataset used in this research consists of 41 variables, including information about orders, logistics, product characteristics, payment methods, and customer review scores as the dependent variable. The review scores range from 1 to 5, with 4 and 5 being the most frequent values, indicating that most customers provide positive ratings shown in Figure 2.

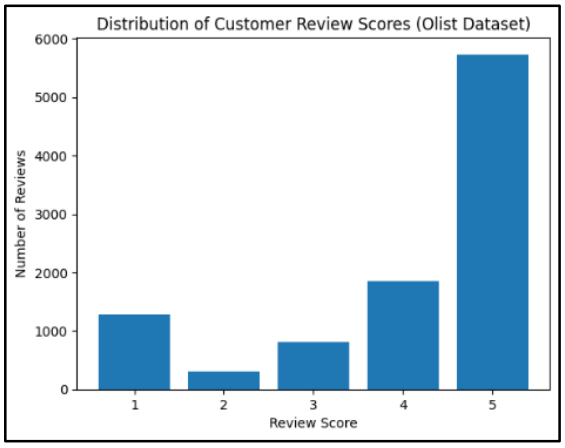

Figure 2. Distribution of Customer Review Scores in the Brazilian E-Commerce Public Dataset by Olist

The following figure shows how customer review scores are spread out in the Brazilian E-Commerce Public Dataset from Olist that higher scores, like 4 and 5, make up most of the data. This shows there's an uneven balance between the different score categories, which can make it harder for the model to predict the lower scores accurately.

From the initial analysis, it appears that logistic variables such as delivery time and shipping cost value have considerable variation across transactions. Additionally, transaction variables and product characteristics also show significant differences across product categories. This condition indicates that a machine learning-based modelling approach with a non-linear approach is necessary.

B. Correlation Analysis

Before starting the modeling process, a correlation assessment was performed utilizing Pearson correlation coefficients to investigate the connections among numerical variables. The findings reveal that the majority of variables exhibit low to moderate correlation levels, suggesting that there are no significant multicollinearity concerns. This strengthens the reliability of the Random Forest model, which demonstrates greater resilience to multicollinearity in comparison to linear models [11].

C. Random Forest Modelling Results

Modelling was performed using the Random Forest Classifier algorithm, where the data was divided into 80% for training and 20% for testing. The model was trained with 200 decision trees to make the prediction results more stable and reduce the likelihood of overfitting to the training data. The modelling results show that Random Forest is able to understand the relationships between variables well, particularly for review classes with high scores (4-5).

D. Model Performance Evaluation

The performance evaluation of the model is conducted using several metrics such as accuracy, confusion matrix, precision, recall, and F1-score. Overall, the model shows quite good prediction results, especially for classes with a larger amount of data.

TABLE II
Confusion Matrix For Review Score Prediction

Actual \ Predicted	1	2	3	4	5
1	69	2	2	3	130
2	11	9	1	3	37
3	8	0	27	6	117
4	10	0	1	71	293
5	18	1	10	30	1066

Based on the confusion matrix results, the model performs well in predicting review of majority classes (especially review score 5 with a recall of 0.947). However, it does not do as well with the less common scores, like 1 and 2. This condition is primarily influenced by the imbalanced distribution of review score classes in the dataset, where higher review scores dominate the data. The imbalance causes the model to be biased toward majority classes, leading to a higher number of misclassifications in minority classes. Even though the overall accuracy and F1-score look good, they are mostly because of the common scores. So, when looking at how well the model works, we need to be careful, especially when checking how it handles the less frequent review scores. Future research could look into using data balancing methods to enhance the accuracy of predictions for less common review score categories. Confusion matrix for the prediction shown in Figure 3.

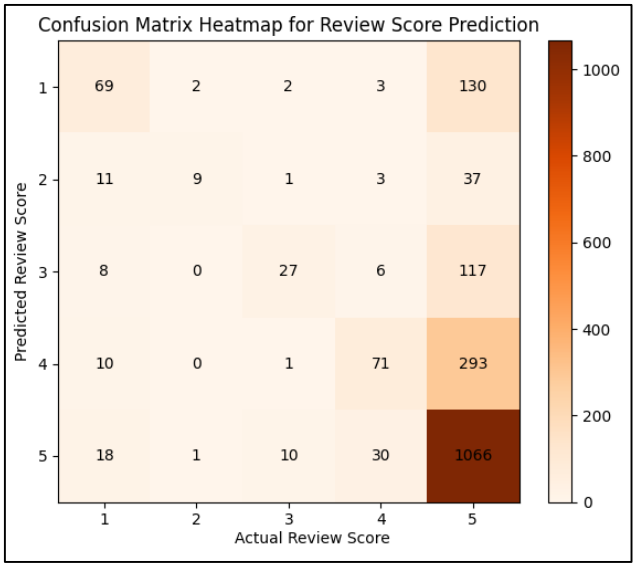

Figure 3. Confusion Matrix Heatmap for Review Score Prediction

Figure 3 shows a heatmap of the confusion matrix using an orange color scheme to make it easier to read. The darker areas show where predictions are more common, and it clearly shows that the correct predictions are mostly in the higher review score categories.

E. Additional Analysis: Review Score Grouping

To better understand customer satisfaction patterns, we did another analysis by grouping review scores into three categories: low (scores 1–2), medium (score 3), and high (scores 4–5). This helps make the analysis easier and deals with the imbalance in the original five-class system. The confusion matrix shows that the model makes more mistakes in predicting the minority classes, especially scores 1 and 2, but performs better for the more common classes like scores 4 and 5.

By grouping the scores into broader categories, the model can more clearly tell the difference between dissatisfied and satisfied customers, which supports earlier findings. This approach also backs up the feature importance results, showing that logistics factors, especially delivery time, are the biggest influences on customer satisfaction. This grouping method is an extra way to better understand how the model works without changing the main model structure.

F. Analysis of Feature Importance

An analysis of feature importance was conducted to determine the variables that most influence review scores. Feature importance analysis helps understand machine learning models by showing which factors have the most impact on the results [15][16]. Table III shows that delivery time is the main factor, followed by payment value, freight value, and price, which play a role in determining customer satisfaction and influencing it.

In terms of quantity, the significance of delivery time stands out as the most crucial factor, with a value of 0.125, closely followed by payment value at 0.112 and freight value at 0.101. This suggests a distinct disparity in their roles among the most influential variables. Moreover, the variables related to logistics as a whole play a major part in the overall feature importance, further emphasizing their key influence on customer satisfaction. Nonetheless, it is essential to recognize that the importance of features in Random Forest can be skewed towards those that exhibit greater variance or possess a higher number of categories. As a result, understanding these findings is best achieved when combined with specialized knowledge and additional analysis to prevent incorrect interpretations [15][16].

These results match up with recent research that shows how important delivery performance and the value of the transaction are in affecting customer satisfaction on e-commerce websites. As growth occurs, there is an increase in full purchases. Logistics and transaction factors play an important role in shaping customer satisfaction. Service quality assessment frameworks keep changing in digital commerce studies, highlighting how important it is to have reliable performance and quick. Studies show that good logistics and clear buying processes help customers have a better experience and are more likely to buy again.

TABLE III
RANKING VARIABLES BASED ON FEATURE IMPORTANCE

Rank	Important Factors	Importance
1	**Delivery Time**	**0.125**
2	**Payment Value**	**0.112**
3	**Freight Value**	**0.101**
4	**Price**	**0.091**
5	Product Description Length	0.089
6	Product Weight	0.078
7	Product Name Length	0.071
8	Product Height	0.063
9	Product Length	0.061
10	Product Width	0.060
11	Product Category	0.057
12	Payment Installments	0.050
13	Product Photos Qty	0.033
14	Payment Sequential	0.007

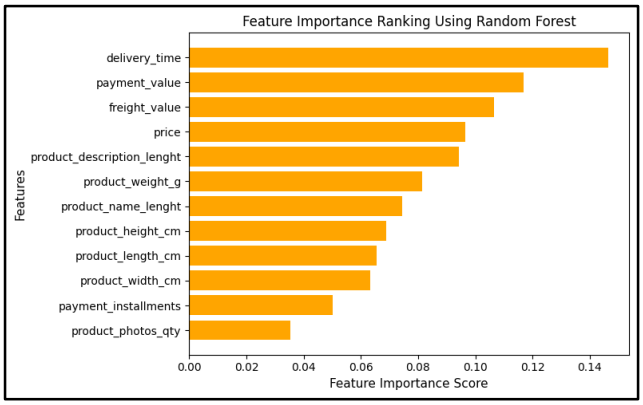

Figure 4. Feature Importance Ranking Using Random Forest

Although the significance of features offers a broad understanding of the impact of variables, it is not sufficient for clarifying intricate model behaviors. Techniques for enhanced interpretability like SHAP (SHapley Additive

exPlanations) and Partial Dependence Plots (PDP) can offer more profound insights into both overall and specific relationships among variables and model results [15][16][17]. These methods facilitate a more granular comprehension of how each feature affects prediction results, which is especially crucial in contexts of decision-making such as analytics for e-commerce. Consequently, it is advisable for future studies to integrate these techniques to improve the clarity and openness of machine learning models.

IV. CONCLUSION

This research aims to analyze the factors influencing customer review scores on e-commerce platforms using a machine learning approach, specifically the Random Forest algorithm. It is crucial to highlight that the outcomes of this research suggest correlations rather than definitive causal links. Machine learning algorithms uncover trends within data but do not determine cause-and-effect relationships. Thus, the findings should be evaluated carefully and complemented by a theoretical framework. In terms of practical application, e-commerce sites ought to focus on enhancing delivery efficiency, providing precise product details, and upholding clear pricing methods. By making these advancements, they can improve the customer experience and inspire more favorable feedback. This study is limited to a single dataset from one e-commerce platform. Future research is recommended to validate the model using datasets from different platforms or regions to ensure generalizability and robustness of the findings. Although this study focuses on the Random Forest model, the results indicate the need for further validation using alternative models. Therefore, future research is encouraged to conduct comparative experiments with models such as Gradient Boosting and XGBoost to ensure robustness and improve predictive performance.

Therefore, the results should be interpreted cautiously and supported by theoretical understanding. Based on the modelling results and feature importance analysis, it can be concluded that logistical factors are the main determinant in shaping customer satisfaction. This result matches what recent research has shown, highlighting how important timely delivery is for making online customers happy.

The delivery time variable proved to be the most dominant factor influencing review scores, indicating that delivery speed significantly impacts customer experience. Previous studies also show that how reliably and on time a delivery is made has a big effect on how people write online reviews. Additionally, factors such as transaction value, shipping value, and price also contributed significantly to customer ratings, particularly because transaction value is directly related to customer expectations regarding the quality of service received. This result matches studies showing that how clear transactions are and how much shipping costs are can change how customers feel about whether a platform is fair and how good its service is. Service quality assessment frameworks keep changing in digital commerce studies, highlighting how important it is to have reliable performance and quick responses.

Product characteristics, particularly the length of the description, weight, and dimensions of the product, also contribute to shaping customer satisfaction. Having accurate and detailed product information helps reduce uncertainty and makes customers more satisfied when they buy things online. Complete and accurate product information can reduce customer uncertainty, while large or heavy products carry a higher risk of dissatisfaction due to potential damage during shipping. Overall, the Random Forest algorithm proved effective in modelling complex relationships between variables and identifying the key factors influencing e-commerce customer review scores. Recent machine learning studies have shown that ensemble learning models like Random Forest are very good at dealing with complex and non-linear data in e-commerce.

REFERENCES

[1] K. Ariansyah, E.R.E. Sirait, B.A. Nugroho, M. Suryanegara, "Drivers of and barriers to e-commerce adoption in Indonesia: Individuals' perspectives and the implications," *Telecommunications Policy*, vol. 45, no. 8, pp. 102219, 2021, doi: 10.1016/j.telpol.2021.102219.

[2] A. Timoshenko, J. Hauser, "Identifying Customer Needs from User-Generated Content," *Marketing Science*, vol. 38, 2019, doi: 10.1287/mksc.2018.1123.

[3] P. Kotler, K. Keller, M. Brady, M. Goodman, T. Hansen, "Marketing Management," 2019.

[4] X. Lin, A.A. Mamun, Q. Yang, M. Masukujjaman, "Examining the effect of logistics service quality on customer satisfaction and re-use intention," *PLOS ONE*, vol. 18, no. 5, pp. e0286382, 2023, doi: 10.1371/journal.pone.0286382.

[5] M.S. Akturk, R.R. Mallipeddi, X. Jia, "Estimating impacts of logistics processes on online customer ratings: Consequences of providing technology-enabled order tracking data to customers," *Journal of Operations Management*, vol. 68, no. 6–7, pp. 775–811, 2022, doi: 10.1002/joom.1204.

[6] R.R. Mardhotillah, B.M. Wibawa, "E-Service Quality Factors and Customer Satisfaction in Shopee's E-Commerce Platform," *Journal of Applied Management and Business*, vol. 6, no. 1, pp. 22–34, 2025, doi: 10.37802/jamb.v6i1.1050.

[7] R. Rachmiani, N. Oktadinna, T. Fauzan, "The Impact of Online Reviews and Ratings on Consumer Purchasing Decisions on E-commerce Platforms," *International Journal of Management Science and Information Technology*, vol. 4, pp. 504–515, 2024, doi: 10.35870/ijmsit.v4i2.3373.

[8] M.Z. Fani, S. Hadi, H. Wirastomo, "Pengaruh Harga, Online Customer Rating, Biaya Ongkir dan Kecepatan Pengiriman Terhadap Keputusan Pembelian di Platform E-commerce (Studi Kasus Pengguna E-commerce di Kota Mataram)," *Journal of Applied Business and Banking (JABB)*, vol. 6, no. 2, 2025, doi: 10.31764/jabb.v6i2.36990.

[9] T. Zhu, Y. Lu, B. Wang, L. Zhao, "Pre-purchase and post-purchase sales promotions on ecommerce platforms: the effects of promotional benefits on customer-based brand equity," *Journal of Electronic Commerce Research*, vol. 24, no. 2, pp. 146–170, 2023.

[10] Y. Hu, "Linking Perceived Value, Customer Satisfaction, and Purchase Intention in E-Commerce Settings," *Advances in Intelligent and Soft Computing*, pp. 623–628, 2011, doi: 10.1007/978-3-642-23753-9_100.

[11] T. Hastie, J. Friedman, R. Tibshirani, "The elements of statistical learning: Data mining, inference, and prediction," 2001, doi: 10.1007/978-0-387-21606-5.

[12] G.S. Budhi, R. Chiong, I. Pranata, Z. Hu, "Using Machine Learning to Predict the Sentiment of Online Reviews: A New Framework for Comparative Analysis," 2021, doi: 10.1007/s11831-020-09464-8.

[13] D.A.S.S. Kumar, "Predicting Online Shopper Behavior: Machine Learning Approaches for Enhanced E-Commerce Insights," *IJSDR* -

International Journal of Scientific Development and Research, vol. 9, no. 12, pp. a166–a173, 2024.
[14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[15] C. Molnar, "Interpretable machine learning: A guide for making black box models explainable," 2022.
[16] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018, doi: 10.1145/3236009.
[18] N. Ullah, J.A. Khan, I. De Falco, G. Sannino, "Explainable Artificial Intelligence: Importance, Use Domains, Stages, Output Shapes, and Challenges," *ACM Comput. Surv.*, vol. 57, no. 4, 2024, doi: 10.1145/3705724.
[19] IBM, "IBM SPSS Modeler CRISP-DM Guide," 2021.
[20] H. He, E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
[21] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.
[22] K.C. Laudon, C.G. Traver, "E-commerce 2021: Business, Technology, and Society," 2021.