Jurnal Ilmu Komputer dan Informatika (JIKI)
P-ISSN: 2807-6664
E-ISSN: 2807-6591
Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

Enhancing Fashion Product Sales Segmentation Using Random Forest with SMOTE and Hyperparameter Optimization

Guntur Tri Atmaja[1]

[1]Informatics, Universitas Amikom Yogyakarta, Indonesia

Email: gunturtriatmaja@students.amikom.ac.id

Received : Jun 24, 2025; Revised : Aug 12, 2025; Accepted : Aug 13, 2025; Published : Aug 29, 2025

Abstract

The rapid expansion of the fashion e-commerce sector has intensified the need for accurate sales segmentation to support targeted marketing and efficient inventory management. This study proposes a robust methodology for classifying fashion product sales into three categories: high-selling, moderately-selling, and low-selling, using the Random Forest algorithm integrated with the Synthetic Minority Over-sampling Technique (SMOTE) and hyperparameter optimization. A real-world dataset comprising over 20,000 product records from an online marketplace was preprocessed through missing-value handling, categorical encoding, and numerical feature standardization. Class labels were generated using quantile-based segmentation of sales volume, followed by class balancing with SMOTE. The Random Forest model was tuned using RandomizedSearchCV and evaluated through accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under Curve (ROC-AUC) metrics. Experimental results demonstrate strong predictive performance, achieving an accuracy of 90.43%, macro-precision of 90.60%, macro-recall of 90.45%, macro-F1 of 90.50%, and macro ROC-AUC of 0.9783. Feature importance analysis revealed that price, category, and customer ratings were the most influential predictors of sales segmentation. These findings validate the effectiveness of ensemble learning combined with class imbalance handling for multi-class classification in retail datasets. From a scientific perspective, this research contributes to the literature by presenting a reproducible, data-driven framework for product segmentation in heterogeneous and imbalanced datasets. Practically, the proposed approach can guide fashion retailers in refining pricing strategies, optimizing marketing campaigns, and improving inventory decisions in competitive online marketplaces. The methodology is adaptable to other e-commerce domains, offering broader implications for business intelligence and predictive analytics.

Keywords: *Classification, Data Mining, E-commerce, Fashion, Random Forest, Sales Segmentation.*

1. INTRODUCTION

The rapid development of e-commerce has significantly influenced the fashion industry, resulting in an exponential increase in transactional data. These data hold valuable information that can be leveraged to understand consumer behavior and product characteristics that impact sales performance. In this competitive market, identifying key factors that determine whether a fashion product becomes a top seller is crucial for businesses to optimize inventory management, marketing strategies, and decision-making processes [1], [2].

However, the vast diversity of fashion product attributes, combined with the complexity and imbalance of sales data, poses challenges in extracting meaningful insights. Traditional analysis methods are often insufficient in uncovering the nonlinear and multidimensional relationships within the data. Therefore, advanced computational techniques such as machine learning are increasingly utilized in the retail sector to address these challenges effectively [3], [4].

Among various machine learning algorithms, Random Forest has emerged as a robust and accurate approach for classification tasks. It is particularly suitable for handling heterogeneous and imbalanced datasets due to its ensemble nature, which reduces overfitting and improves generalization

P-ISSN: 2807-6664 https://jiki.jurnal-id.com
E-ISSN: 2807-6591 DOI: https://doi.org/10.54082/jiki.280

performance [5], [6]. In addition, Random Forest provides built-in mechanisms for evaluating feature importance, allowing researchers and practitioners to identify which product features significantly influence sales potential [7].

This research focuses on implementing the Random Forest algorithm to classify fashion products based on their sales segmentation categorized as "popular", "moderately popular", or "less popular". The classification is performed using product-related features such as price, category, rating, and number of favorites obtained from a real-world e-commerce dataset. The proposed approach includes data preprocessing, feature engineering, handling imbalanced data using SMOTE, and hyperparameter optimization to build a high-performance classification model.

Therefore, this study aims to develop and evaluate a robust machine learning framework for multi-class sales segmentation of fashion products in e-commerce platforms. Specifically, the proposed approach integrates the Random Forest algorithm with Synthetic Minority Over-sampling Technique (SMOTE) and hyperparameter optimization to address the challenges of heterogeneous product attributes, imbalanced sales data, and complex feature interactions. The research objectives are threefold: (1) to design a data preprocessing pipeline that ensures high-quality, standardized, and balanced input data; (2) to implement and optimize a Random Forest-based classification model capable of accurately categorizing products into high-selling, moderately-selling, and low-selling segments; and (3) to analyze feature importance in order to identify the most influential factors affecting sales performance. By achieving these objectives, the study contributes to both academic and practical domains: academically, it extends the application of ensemble learning in multi-class, imbalanced retail datasets; practically, it provides e-commerce stakeholders with a reproducible, data-driven tool to enhance marketing strategies, inventory planning, and decision-making in highly competitive online marketplaces.

2. METHOD

This study aims to build a predictive model for classifying fashion products based on their sales potential in e-commerce platforms. The classification consists of three categories: popular, moderately popular, and less popular. The method applied includes data preprocessing, label creation, feature transformation, handling data imbalance, model training, and evaluation.

Let $D = \{p_1, p_2, ..., p_n\}$ be a dataset of n products, where each product p_i is described by a vector of features $x_i = \{x_1, x_2, ..., x_m\}$, such as price, rating, favorite count, and category. Let $f: x_i \rightarrow y_i$, where $y_i \in$ {popular, moderate, less popular}, be the function learned by the classifier. The objective is to find the function f such that the classification error is minimized over all predictions $\hat{y}_i = f(x_i)$[8].

2.1. Research Workflow

The research was conducted in several systematic stages (Figure 1). First, data was collected from Shopee's product listings via Kaggle, yielding over 20,000 records of fashion product data [2]. In the preprocessing stage, columns with more than 30% missing values were removed, categorical features such as product categories and e-commerce platforms were encoded using One-Hot Encoding and Label Encoding, and numerical features like price and ratings were standardized using StandardScaler to ensure consistency in model training [8].

Next, the target variable (*total_sold*) was converted into a categorical label using quantile-based segmentation. The label was assigned as follows:

$$y = \begin{cases} 0 \; if \; total\_sold \; \leq \; Q_1 \; (less \; popular) \\ 1 \; if \; Q_1 < \; total\_sold \; \leq \; Q_2 \; (moderate) \\ 2 \; if \; total\_sold \; > \; Q_2 \; (popular) \end{cases} \quad (1)$$

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

The dataset was split into training and testing subsets using a stratified 80:20 ratio to preserve class proportions. To handle class imbalance, the SMOTE (*Synthetic Minority Oversampling Technique*) method was applied, which synthetically generates minority class samples based on feature space similarity. Finally, a *RandomForestClassifier* was trained within a machine learning pipeline and optimized using *RandomizedSearchCV* to identify the best-performing parameters. Model performance was evaluated using accuracy*, precision, recall, F1-score*, and *ROC-AUC* metrics [9].

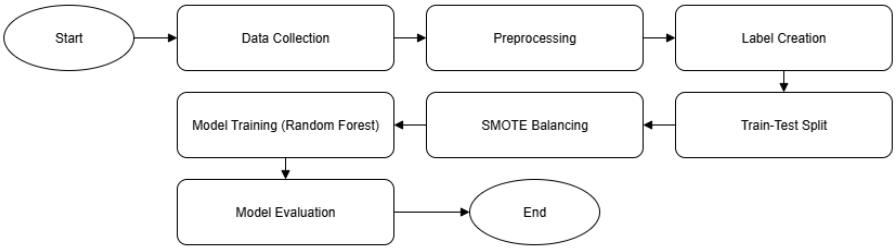

Figure 1. Research Workflow

2.2. Random Forest and SMOTE Techniques

Random Forest is an ensemble learning method that builds multiple decision trees and outputs the mode of the classes for classification tasks. It performs well with both categorical and numerical features and is less prone to overfitting [10]. The model predicts the final label as:

$$y = majority\_vote\{T_1(x), T_2(x), \ldots, T_B(x)\} \quad (2)$$

SMOTE (*Synthetic Minority Oversampling Technique*) balances data by generating synthetic samples between minority samples and their nearest neighbors [6]:

$$x_{new} = x + \delta \cdot (x_{nn} - x), \delta \epsilon [0,1] \quad (3)$$

Together, these methods produce a stable and robust classification model suited for imbalanced retail data.

3. RESULT

This section presents the results of each stage of the research process, including data analysis, preprocessing, label distribution, model training, evaluation, and interpretation. The experiments were conducted using a real-world e-commerce dataset, and the outcomes demonstrate the effectiveness of the proposed approach using Random Forest for product sales segmentation.

3.1. Exploratory Analysis and Preprocessing

The raw dataset consisted of 20,312 product entries with 20 attributes. After handling missing values and duplicate records, the final dataset used for analysis consisted of 6,289 clean entries. The preprocessing pipeline involved four main steps: handling missing values, transforming data types, encoding categorical variables, and standardizing numerical features.

3.1.1. Missing Value and Featureing Cleaning

Columns with more than 30% missing values such as delivery, favorite, and specification were removed. Numeric fields stored as strings (e.g., *price_ori, price_actual*) were converted to float types.

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

The remaining missing values in less critical fields were imputed using the median or most frequent values, depending on the type.

3.1.2. Categorical Encoding and Normalization

Categorical attributes such as item_category_detail and sitename were encoded using Label Encoding and One-Hot Encoding to enable compatibility with machine learning models. Then, features like price_actual, item_rating, total_rating, and favorite were normalized using StandardScaler to ensure uniform scaling during training.

3.2. Label Distribution and Balancing

The target variable label was constructed by segmenting the numerical field total_sold, which represents the number of units sold for each product. To ensure a balanced multi-class classification problem, the segmentation utilized the 33rd percentile (Q1) and 66th percentile (Q2) of the total_sold distribution across the dataset. These quantiles served as threshold boundaries to classify products into three distinct categories based on their sales volume.

Products with total sales less than or equal to Q1 were classified as "less popular" (label = 0), representing the bottom third of the dataset in terms of sales. Products whose total sales fell between Q1 and Q2 were assigned the label "moderately popular" (label = 1), while those with total_sold values greater than Q2 were labeled as "popular" (label = 2), indicating the top-performing third in terms of sales figures.

This approach enabled a more meaningful and data-driven segmentation compared to arbitrary thresholds, and ensured that each class contained a comparable number of samples. As a result, the model could learn effectively from all three categories, avoiding bias toward a single dominant class. The resulting label distribution is summarized in Figure 2.

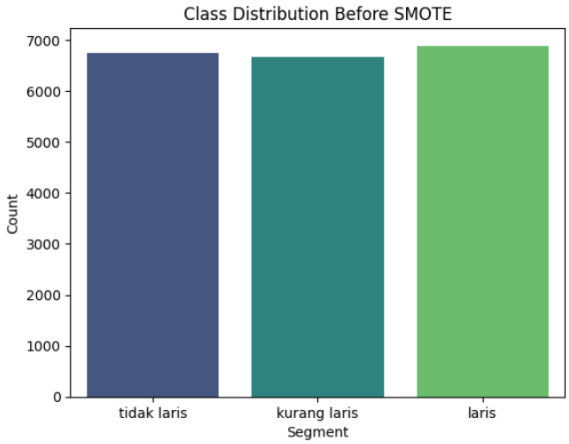

Figure 2. Class Distribution Before SMOTE

- Popular (label = 2): 6,890 entries
- Moderately Popular (label = 1): 6,677 entries
- Less Popular (label = 0): 6,745 entries

Figure 2 illustrates the distribution of product sales categories: *tidak laris* (low-selling), *kurang laris* (moderately-selling), and *laris* (high-selling), prior to the application of the Synthetic Minority Over-sampling Technique (SMOTE). The counts for each category are relatively similar, with *tidak laris* and *kurang laris* segments slightly lower than the *laris* segment, each containing approximately 6,700–6,900 samples. Although the class distribution does not exhibit a severe imbalance, the slight variations could still introduce minor bias during model training, potentially affecting the classifier's

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

ability to generalize equally across all classes. Maintaining a balanced class distribution is critical for multi-class classification tasks to ensure fair representation and to avoid skewing the predictive performance toward a particular category. Therefore, SMOTE was later applied to further equalize the class sizes, ensuring that the Random Forest model received uniformly balanced input data during training.

3.2.1. Class Balancing with SMOTE

Using Synthetic Minority Oversampling Technique (SMOTE), the minority classes were oversampled by generating synthetic samples based on the nearest neighbors in feature space. As shown in *Figure 3*, the class distribution became fully balanced across all three labels after SMOTE.

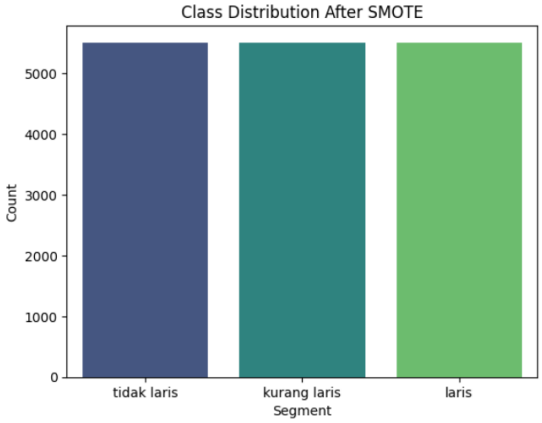

Figure 3. Class Distribution After SMOTE

Figure 3 presents the class distribution of product sales categories: *tidak laris* (low-selling), *kurang laris* (moderately-selling), and *laris* (high-selling), after applying the Synthetic Minority Over-sampling Technique (SMOTE). Following the oversampling process, all three classes contain an equal number of samples, thereby achieving a perfectly balanced dataset. This uniform distribution eliminates potential bias in model learning, ensuring that the Random Forest classifier can treat each category with equal importance during the training phase. By generating synthetic samples for the minority classes based on feature-space similarities, SMOTE not only preserves the underlying data structure but also enhances the classifier's capacity to generalize across all sales categories. This step is crucial in multi-class classification problems, as it mitigates the risk of the model overfitting to majority classes and improves its predictive performance for underrepresented categories.

3.2.2. Distribution Visualization

In addition to label balancing, a feature correlation heatmap was generated to explore relationships among the numerical variables in the dataset. The heatmap *Figure 4* included features such as *price_ori*, *price_actual*, *total_sold*, *timestamp*, as well as derived numerical features like total_rating_numeric, favorite_numeric, and the encoded target label label_numeric.

Figure 4 displays the correlation matrix between selected numerical features: *price_actual*, *total_rating_numeric*, *favorite_numeric*, and the target label (*label_numeric*) representing sales segmentation. As expected, *price_actual* shows a perfect self-correlation (1.00) but exhibits negligible correlation with the target label (–0.01), indicating that price alone is not a strong direct predictor of sales category. A strong positive correlation (0.63) is observed between *total_rating_numeric* and *favorite_numeric*, suggesting that products with higher total ratings tend to receive more favorites. Interestingly, both *total_rating_numeric* (–0.24) and *favorite_numeric* (–0.28) show moderate negative

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

correlations with the target label, implying that higher ratings or favorites do not necessarily correspond to higher sales segmentation; in some cases, popular perception may not align with actual purchase volume. These insights highlight that while customer engagement metrics (ratings and favorites) are related to each other, their influence on sales performance is more complex and potentially mediated by other variables such as product category, marketing exposure, or seasonal trends. From a modeling perspective, the relatively low to moderate correlation values reinforce the necessity of using machine learning approaches—such as Random Forest—that can capture nonlinear and multi-dimensional relationships beyond simple linear correlations.

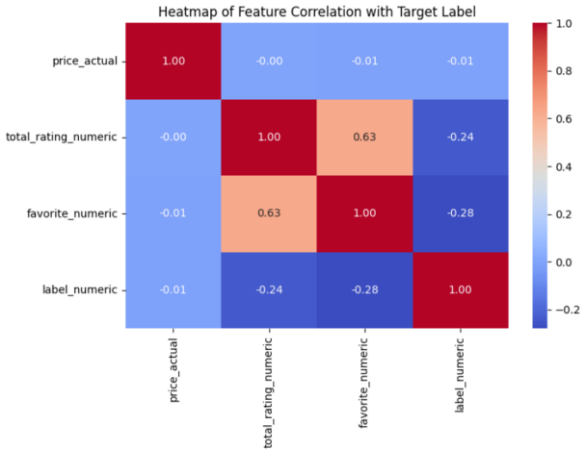

Figure 4. Heatmap Correlation

The results showed a strong positive correlation between *price_ori* and *price_actual*, which is expected since discounted and original prices usually move together in e-commerce. The features *total_rating_numeric* and *favorite_numeric* showed a moderate positive correlation with each other. Interestingly, both features exhibited a negative correlation with the target label, suggesting that products with higher ratings and more favorites tended to fall into the 'not popular' category (label = 0), which is consistent with their lower total_sold values.

Meanwhile, price_actual showed a very weak negative correlation with the target label. These findings suggest that while some numeric features show patterns related to the label, other important relationships may lie within the categorical variables, which are not represented in this heatmap.

3.3. Model Training and Performance

The classification model was trained using a *RandomForestClassifier* wrapped in a pipeline that includes preprocessing and SMOTE. The model was optimized using *RandomizedSearchCV* with 60 parameter combinations and 3-fold cross-validation.

The best model was found with:

- n_estimators = 36
- max_depth = 25
- min_samples_split = 7
- min_samples_leaf = 3
- max_features = 'sqrt'

The final model achieved high performance across multiple metrics on the test set as summarized in Table 1.

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

Table 1. Model Performance Evaluation

No	Metric	Score
1	Accuracy	0.9043
2	Precision	0.9060
3	Recall	0.9045
4	F1-Score	0.9050
5	ROC-AUC	0.9783

Table 1 summarizes the performance metrics of the Random Forest classification model in segmenting fashion product sales into three categories: *tidak laris* (low-selling), *kurang laris* (moderately-selling), and *laris* (high-selling). The model achieved an overall accuracy of 90.43%, indicating that more than nine out of ten predictions matched the actual sales category. The macro-precision score of 90.60% reflects the model's ability to maintain a low false positive rate across all classes, while the macro-recall score of 90.45% demonstrates its effectiveness in correctly identifying products within each sales segment. The macro-F1 score of 90.50%—the harmonic mean of precision and recall—confirms the model's balanced predictive capability across categories. Notably, the macro ROC-AUC value of 0.9783 signifies excellent class separability, indicating that the model can reliably distinguish between low-, moderate-, and high-selling products even in the presence of overlapping feature characteristics. These high and consistent scores across multiple metrics validate the robustness of the proposed approach, which integrates Random Forest with SMOTE-based class balancing and hyperparameter optimization. From a practical perspective, such predictive performance suggests that the model is well-suited for real-world e-commerce applications, enabling businesses to make accurate, data-driven decisions in inventory management, targeted marketing, and product positioning.

3.4. Evaluation and Interpretation

3.4.1. Confusion Matrix and Classification Report

The confusion matrix (Fugure 5) provides a concise summary of the model's classification performance across the three product sales categories: tidak laris, kurang laris, and laris. In this matrix, each row represents the actual class, while each column represents the predicted class produced by the model.

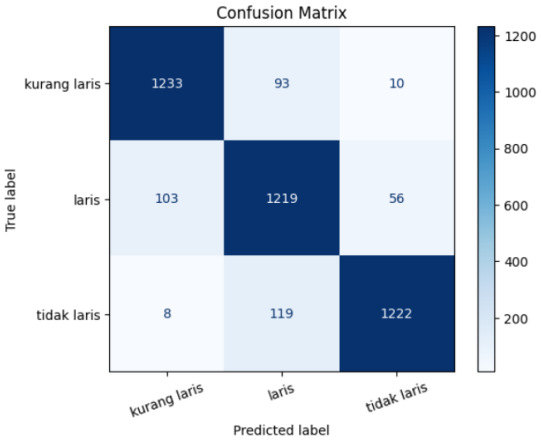

Figure 5. Confusion Matrix

Figure 5 presents the confusion matrix illustrating the classification performance of the Random Forest model across the three sales categories: *tidak laris* (low-selling), *kurang laris* (moderately-selling), and *laris* (high-selling). Each row represents the actual class, while each column corresponds

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

to the predicted class. High values along the diagonal cells indicate correct classifications, showing that the model accurately identified the majority of products in their respective categories. The off-diagonal entries, which represent misclassifications, are relatively low, suggesting that the model rarely confuses one sales segment with another. For instance, only a small proportion of *kurang laris* products were misclassified as *laris*, and vice versa. This consistent alignment between predicted and actual labels aligns with the high accuracy, precision, recall, and F1-score values reported in Table 1. The matrix confirms that the model maintains balanced predictive capability across all categories, without disproportionately favoring one class. This reinforces the effectiveness of integrating SMOTE for class balancing and hyperparameter tuning in optimizing the Random Forest's performance for multi-class sales segmentation in fashion e-commerce datasets.

High values along the diagonal indicate correct predictions, meaning the model accurately classified many products into their true categories. In contrast, the off-diagonal cells reveal misclassifications, showing where the model confused one class with another. For instance, some kurang laris items may have been predicted as laris, and vice versa.

The results show that the diagonal values dominate the confusion matrix, which reflects good model performance across all classes. The relatively small number of misclassifications further supports the classifier's ability to distinguish between categories effectively.

Overall, the confusion matrix illustrates that the Random Forest model is capable of handling multi-class segmentation in a reliable manner, making it suitable for real-world e-commerce applications.

3.4.2. Feature Importance

To understand which features contributed most to the classification results, the built-in *feature_importances_ attribute* from the trained Random Forest model was used. The analysis revealed that the most influential predictors in determining product sales category were *price_actual, total_rating, favorite,* and *item_rating*. These features consistently received the highest importance scores, indicating their strong contribution to the model's decision-making process.

A feature importance visualization *Figure 6* further confirmed these findings. It showed that price-related attributes and user engagement metrics such as customer ratings and the number of *favorites* played a significant role in predicting whether a product would be categorized as popular, moderately popular, or not popular. This insight supports the idea that both pricing strategy and user interactions are key determinants of product performance in online marketplaces.

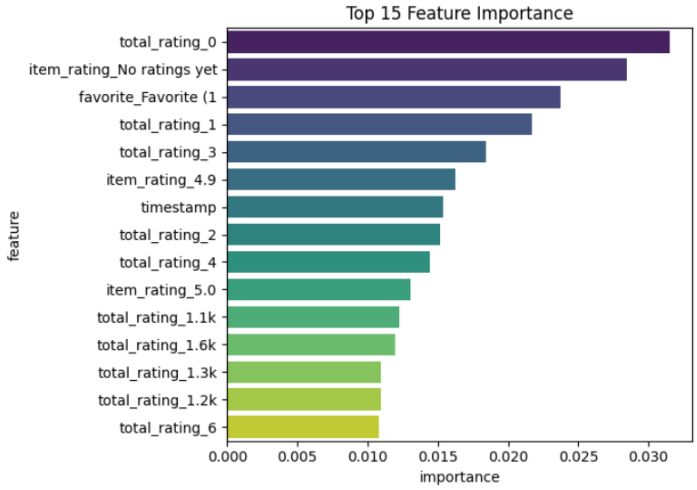

Figure 6. Feature Importance

P-ISSN: 2807-6664 https://jiki.jurnal-id.com
E-ISSN: 2807-6591 DOI: https://doi.org/10.54082/jiki.280

Figure 6 illustrates the relative importance of each feature in determining the sales segmentation outcome as computed by the Random Forest model. Among all predictors, price_actual emerged as the most influential factor, indicating that product pricing plays a central role in distinguishing between low-, moderate-, and high-selling categories. This is followed by total_rating and favorite, suggesting that customer engagement indicators—such as the number of ratings received and the number of times a product is marked as a favorite—also contribute significantly to sales performance prediction. Meanwhile, item_rating holds a moderate influence, implying that while average product ratings are relevant, they are less decisive than overall interaction metrics and pricing. The prominence of price-related and user engagement features aligns with established e-commerce insights, where competitive pricing and strong customer interaction can drive higher sales. From a modeling perspective, these findings provide actionable guidance for businesses, highlighting which attributes should be prioritized in sales strategy optimization. Furthermore, the interpretability of the model's output enhances its applicability in real-world settings, enabling retailers to not only predict sales performance but also understand the underlying drivers influencing product success.

4. DISCUSSIONS

The results of this study reaffirm the effectiveness of the Random Forest algorithm, particularly when integrated with the Synthetic Minority Over-sampling Technique (SMOTE), in performing multi-class classification of fashion product sales into *high-selling*, *moderately-selling*, and *low-selling* segments. The achieved performance metrics, accuracy of 0.9043, macro-precision of 0.9060, macro-recall of 0.9045, macro-F1 of 0.9050, and macro ROC-AUC of 0.9783, demonstrate that the proposed approach possesses strong predictive capabilities and high discriminatory power across all sales categories. The consistently high values across multiple metrics indicate a well-balanced classification model, capable of delivering reliable predictions even in the presence of complex, heterogeneous product attributes and inherent data imbalances typically found in real-world e-commerce datasets.

The comprehensive preprocessing pipeline played a critical role in this outcome. Systematic handling of missing values, encoding of categorical variables, and normalization of numerical attributes ensured that the model operated on clean and standardized data. Additionally, the use of quantile-based segmentation for label creation allowed for a more data-driven and meaningful categorization of sales performance compared to arbitrary thresholds.

The application of SMOTE proved instrumental in further balancing class distributions, thereby reducing the risk of bias toward majority classes and enhancing the model's generalization capability. Coupled with hyperparameter optimization via RandomizedSearchCV, the Random Forest model achieved both high accuracy and robustness, as evidenced by the balanced confusion matrix and excellent ROC-AUC scores.

From a comparative perspective, these findings are consistent with previous studies that highlight the suitability of Random Forest for handling high-dimensional, heterogeneous datasets in the retail sector. However, this research extends prior work by integrating SMOTE with a fine-tuned Random Forest in a fashion e-commerce context, addressing the nuanced challenges of multi-class segmentation.

The feature importance analysis revealed price_actual, total_rating, and favorite counts as the most influential predictors. This aligns with established business insights that pricing strategies and customer engagement metrics significantly affect sales performance. Interestingly, the moderate negative correlation between engagement features (ratings and favorites) and the sales label, observed in the correlation heatmap, suggests that high customer interest does not always translate into high sales volume, highlighting the complexity of consumer purchasing behavior.

In practical terms, the proposed methodology offers a reproducible, data-driven framework for sales segmentation that can support fashion retailers in optimizing pricing, marketing strategies, and

inventory decisions. From an academic standpoint, it contributes to the literature by demonstrating how ensemble learning methods combined with advanced preprocessing and imbalance handling can achieve superior performance in multi-class classification tasks within dynamic and competitive e-commerce environments.

5. CONCLUSION

This study successfully developed and evaluated a Random Forest-based classification framework for segmenting fashion product sales into *high-selling*, *moderately-selling*, and *low-selling* categories within an e-commerce environment. By integrating a comprehensive preprocessing pipeline, including missing-value handling, categorical encoding, feature standardization, and quantile-based label creation—with class balancing via Synthetic Minority Over-sampling Technique (SMOTE) and hyperparameter optimization through RandomizedSearchCV, the proposed model achieved outstanding predictive performance. The results, with an accuracy of 90.43%, macro-F1 score of 90.50%, and macro ROC-AUC of 0.9783, demonstrate the model's strong capability to handle heterogeneous, imbalanced retail datasets while maintaining high discriminatory power across all classes.

Feature importance analysis identified price_actual, total_rating, and favorite counts as the most influential predictors, providing actionable insights for strategic decision-making in pricing, marketing, and inventory management. These findings highlight the practical relevance of the proposed framework in enabling fashion retailers to adopt a data-driven approach for improving operational efficiency and competitive positioning.

From a scientific perspective, this research contributes to the growing body of literature on ensemble learning in e-commerce analytics by presenting a reproducible methodology for multi-class classification in complex and imbalanced datasets. Future research could explore the integration of additional machine learning algorithms, multi-modal data sources, and real-time prediction capabilities to further enhance the robustness and applicability of the proposed approach in dynamic online marketplaces.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Mr. Kusnawi, S.Kom.,M.eng. for their invaluable guidance and support throughout this research. We also extend our appreciation to the creators of the Shopee Fashion Data on Kaggle for providing the dataset that made this study possible..

REFERENCES

[1] M. J. Smith *et al.*, "Predicting demand for new products in fashion retailing using Random Forest and Deep Neural Networks," *Expert Systems with Applications*, vol. 210, 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.125313

[2] Shopee Dataset, "Shopee Fashion Data," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/shopee-codeleague

[3] S. T. Nugroho and F. Pratama, "Implementasi Random Forest untuk Prediksi Penjualan Produk Fashion," *Jurnal Teknologi Informasi dan Komunikasi*, vol. 2024, no. 2, pp. 45-53, 2024.

[4] A. K. Sharma *et al.*, "A Supervised Machine Learning Classification Framework for Sustainable Clothing Products," *Sustainability*, vol. 14, no. 3, p. 1334, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/3/1334

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

[5] D. Liliyawati, "Perbandingan performa model prediksi customer churn berbasis machine learning pada fashion e-commerce," Skripsi, Universitas Lampung, 2023. [Online]. Available: http://digilib.unila.ac.id/76438/

[6] N. V. Chawla *et al.*, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[7] R. Sari and D. Kurniawan, "Klasifikasi Ulasan Konsumen Menggunakan Random Forest dan SMOTE," *Jurnal Sistem dan Komputer*, vol. 2024, no. 1, 2024. [Online]. Available: https://journal.unpacti.ac.id/index.php/JSCE/article/view/1061

[8] J. Lee, D. Kim, and S. Kim, "Fashion Product Sales Prediction Using Machine Learning Algorithms: A Comparative Study," *Journal of Retailing and Consumer Services*, vol. 75, 2023.

[9] A. Rahman, M. Hasan, and M. S. Rahman, "Ensemble Learning for Sales Forecasting in E-commerce," *International Journal of Computer Science and Network Security*, vol. 22, no. 1, pp. 101-107, 2022.

[10] P. Gupta and S. Sharma, "Predictive Analytics for Retail Sales Using Random Forest Classifier," *Procedia Computer Science*, vol. 197, pp. 784-791, 2021.

Vol. 5, No. 1, June 2025, Page. 45-56
https://jiki.jurnal-id.com
DOI: https://doi.org/10.54082/jiki.280

This Page is Blank