JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR COSMETIC SALES PREDICTION ON TOKOPEDIA

Mutia Sahira[1], Ken Ditha Tania[1*], Mira Afrina[1]
[1]Faculty of Computer Science, Universitas Sriwijaya
email: *kenya.tania@gmail.com

Abstract: The rapid growth of the cosmetics industry on e-commerce platforms has intensified competition, creating a critical need for effective, data-driven marketing strategies. This study aims to conduct a comparative analysis of machine learning algorithms to predict the sales categories (High, Medium, Low) of cosmetic products on the Tokopedia marketplace. Four classification models; Random Forest, XGBoost, Logistic Regression, and Naive Bayes were trained and evaluated on data collected via web scraping. The methodology incorporates the Synthetic Minority Over-sampling Technique (SMOTE) to address significant class imbalance and GridSearchCV for hyperparameter optimization to ensure a fair and robust comparison. The experimental results conclusively show that the Random Forest model achieved the best performance, yielding the highest F1-Score Macro Average of 0.75 and an accuracy of 85.3%. The superior model was subsequently implemented in a simple recommendation system to simulate optimal discount strategies, demonstrating its practical utility in providing actionable insights for business decisions.

Keywords: classification; comparative analysis; machine learning; sales prediction; SMOTE

Abstrak: Pertumbuhan pesat industri kosmetik pada platform e-commerce telah membuat persaingan ketat, sehingga menciptakan kebutuhan krusial akan strategi pemasaran yang efektif dan berbasis data. Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap algoritma machine learning untuk memprediksi kategori penjualan (Tinggi, Sedang, Rendah) produk kosmetik di marketplace Tokopedia. Empat model klasifikasi, yaitu Random Forest, XGBoost, Regresi Logistik, dan Naive Bayes, dilatih dan dievaluasi menggunakan data yang dikumpulkan melalui web scraping. Metodologi penelitian ini menerapkan Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas yang signifikan dan GridSearchCV untuk optimisasi hyperparameter guna memastikan perbandingan yang adil. Hasil eksperimen menunjukkan bahwa model Random Forest mencapai performa terbaik, dengan menghasilkan F1-Score Macro Average tertinggi sebesar 0,75 dan akurasi 85,3%. Model unggul ini kemudian diimplementasikan dalam sebuah sistem rekomendasi sederhana untuk menyimulasikan strategi diskon yang optimal, yang menunjukkan kegunaan praktisnya dalam memberikan wawasan yang dapat ditindaklanjuti untuk pengambilan keputusan bisnis.

Kata kunci: analisis komparatif; klasifikasi; machine learning; prediksi penjualan; SMOTE

INTRODUCTION

The rapid development of the beauty industry, especially on e-commerce platforms, has created a highly competitive business environment. Various brands, including MakeOver, compete fiercely to attract consumer attention and increase sales volume. In this competition, effective marketing

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

strategies are the key to competitive advantage. Previous research has consistently shown that factors such as price discounts and promotions have a positive and significant influence on purchase decisions and sales volume of beauty products [1], [2].

Along with the shift in consumer behavior to the digital realm, platforms like Tokopedia serve not only as sales platforms but also as an abundant and crucial source of data for businesses, including Micro, Small, and Medium Enterprises (MSMEs) [3]. Historical sales data, product attributes, prices, and digital promotion labels contain hidden patterns regarding consumer purchasing behavior [4]. However, conventional analysis is often insufficient to uncover these complex dynamics. As stated by [5], models that rely solely on historical sales data are often inaccurate because they fail to incorporate contextual information such as marketing campaigns.

To address this challenge, the application of machine learning offers a solution proven to be superior to traditional statistical methods for sales classification in e-commerce [6]. By using machine learning, companies can predict a product's sales category, a task known as multiclass classification, and identify its main driving factors [7]. Algorithms such as Random Forest and XGBoost have proven effective in various cases, demonstrating their capability in handling complex data [8], [9].

Nevertheless, the implementation of machine learning on real-world sales data faces two main technical challenges. One of the primary issues is imbalanced data, where the number of products in the high sales category (minority class) is far less than in other categories (majority class). This condition can cause the model to become biased and perform poorly in predicting the minority class, which is often the most important [10]. This phenomenon is similar to cases in other domains, such as coffee quality classification, where one quality category dominates the data, making predictions on rare categories unreliable [11]. The Synthetic Minority Over-sampling Technique (SMOTE) has been proven effective in addressing this issue by balancing the class distribution [12], thereby improving accuracy, recall, and F1-score for the minority class [11], [13], [14].

While machine learning shows promise, its application to brand-specific e-commerce data such as MakeOver remains underexplored. In the case of MakeOver's official store on Tokopedia, sales distribution is highly uneven, with only a few products achieving consistently high sales while most fall into medium and low categories. This imbalance complicates the identification of effective product attributes and promotional strategies, risking inefficient decisions. Prior studies highlight the effectiveness of machine learning in e-commerce sales prediction, the role of SMOTE in addressing class imbalance, and the strong performance of ensemble models such as Random Forest and XGBoost. However, these studies often use aggregated or non-brand-specific datasets and lack a fair comparative framework, leaving a gap in understanding how different algorithms perform when applied systematically to brand-specific data.

Since algorithms have distinct strengths; Random Forest offering interpretability and XGBoost excelling in non-linear feature interactions, relying on a single model may not be optimal. This

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

study addresses the gap by implementing a structured pipeline with SMOTE, GridSearchCV, and a comparative evaluation of four machine learning algorithms, ensuring fairness in model selection and practical applicability.

Therefore, this study aims to build a comparative machine learning framework using SMOTE and GridSearchCV to identify the best predictive model for MakeOver's Tokopedia sales categories, providing reliable insights for data-driven pricing and promotion strategies.

METHOD

This study employs a quantitative data mining approach using classification algorithms to compare the performance of several machine learning models in classifying MakeOver cosmetic product sales on Tokopedia into three categories: High, Medium, and Low. The objective is to determine the most accurate model that can serve as the foundation for a strategic recommendation system.

Research Framework

The research process follows a structured framework that guides method design, data collection, analysis, and presentation [15].

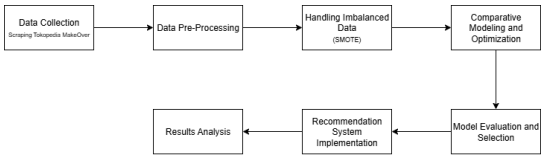

Image 1. Research Framework

Data Collection

The dataset consists of all products listed in MakeOver's official Tokopedia store, collected in June 2025 to capture a market snapshot. Purposive sampling was applied to include all available products during the collection period. Data were obtained using a custom Python web scraping script, extracting attributes relevant to sales and promotion.

Research Variables

The variables used in this research are divided into two categories: the dependent variable (target) and the independent variables (predictors).

The dependent variable in this study is the sales category (sales_category), which is a categorical variable with three levels. The High category refers to products sold in quantities exceeding 100,000 units, the Medium category represents products sold between 1,001 and 100,000 units, while the Low category includes products sold in quantities of 1,000 units or less.

The independent variables are the features used to predict the dependent variable. These variables are presented in the following table/section.

Table 1. Research Independent Variables

Variable Name	Data Type
discount_percent	Numeric
current_price	Numeric
original_price	Numeric
price_diff	Numeric
category	Categorical
is_flash_sale	Binary
is_bundle	Binary
is_best_seller	Binary

Research Methods

This study follows a structured data mining workflow consisting of four stages: data preparation, handling imbalance, comparative modeling, and evaluation. The raw data obtained via scraping was cleaned, converted, and transformed into independent features (Table 1), while the dependent variable

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

(sales_category) was derived from sales quantity.

To address class imbalance, SMOTE oversampling was applied only to the training set within the pipeline to avoid data leakage. The dataset was then split into training (80%) and test (20%) sets using stratified sampling. A pipeline integrating StandardScaler (for numerical features) and OneHotEncoder (for categorical features) was built to ensure consistent preprocessing. The StandardScaler standardizes features by removing the mean and scaling to unit variance using the formula (1) where z is the standardized value, x is the original value, μ is the mean, and σ is the standard deviation.

$$z=\frac{x-\mu}{\sigma} \quad (1)$$

Four algorithms; Random Forest, XGBoost, Logistic Regression, and Naive Bayes were optimized with GridSearchCV. The mathematical foundations of the primary models are as follows:

Logistic Regression models the probability of a categorical outcome using the sigmoid function (2)

$$\sigma(z)=\frac{1}{1+e^{-z}} \quad (2)$$

Naive Bayes classifier is based on Bayes' theorem with an assumption of feature independence (3)

$$P(y \mid x_1,...,xn)=\frac{P(y)\Pi_{i=1}^{n}P(xi|y)}{P(x_1,\ldots,xn)} (3)$$

Random Forest and XGBoost are ensemble methods that operate by combining multiple decision trees to improve predictive accuracy and do not have a single defining formula.

Table 2. Hyperparameter Search Space in GridSearchCV

Algorithm	Hyperparameters (GridSearchCV Ranges)
Random Forest	n_estimators: [100, 200, 300]; max_depth: [10, 20, None]; min_samples_split: [2, 5, 10]; criterion: [“gini”, “entropy”]
XGBoost	n_estimators: [100, 200, 300]; learning_rate: [0.01, 0.1, 0.3]; max_depth: [3, 6, 10]; subsample: [0.8, 1.0]; colsample_bytree: [0.8, 1.0]
Logistic Regression	penalty: [“l1”, “l2”]; C: [0.01, 0.1, 1, 10]; solver: [“liblinear”, “saga”]
Naive Bayes	var_smoothing: [1e-09, 1e-08, 1e-07, 1e-06]

Finally, model performance was evaluated on the test data using Accuracy, Precision, Recall, and Macro F1-Score. Given the imbalanced nature of the dataset, Macro F1-Score was emphasized as the primary selection criterion, calculated as (4) where N denotes the number of classes, and $Precision_i$ and $Recall_i$ represent the precision and recall for each class i. The model with the highest Macro F1-Score was selected as the best predictive model for sales classification.

$$F1macro=\frac{1}{N}\sum_{i=1}^{N}\frac{2\times Precision_i\times Recall_i}{Precision_i+Recall_i} \quad (4)$$

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

RESULT AND DISCUSSION

The experimental process began with hyperparameter optimization followed by a comparative performance evaluation to determine the most superior predictive model. The optimization process, using GridSearchCV with 3-fold cross-validation and the f1_macro metric, yielded the optimal hyperparameter configurations for each algorithm. Determining these hyperparameters is a crucial step to ensure that the performance comparison conducted is fair and that each model has reached its maximum potential.

The performance of the four optimized models was then evaluated on the test data. The evaluation results are presented in Table 3. The primary metric of focus is the F1-Score, particularly the Macro Average, which provides a balanced overview of performance across all classes, a vital aspect given the class imbalance in the original dataset.

Table 3. Comparison of Classification Model Performance on Test Data

Model	Model Performance Evaluation					
	Metric	High	Low	Medium	Accuracy	Macro AVG (F1)
Random Forest	Precision	0.33	0.88	0.93	85.3%	0.75
	Recall	1.00	0.94	0.76		
	F1-Score	0.50	0.91	0.84		
XGBoost	Precision	0.25	0.88	0.92	82.4%	0.70
	Recall	1.00	0.94	0.71		
	F1-Score	0.40	0.91	0.80		
Logistic Regression	Precision	0.25	0.82	0.85	76.5%	0.66
	Recall	1.00	0.88	0.65		
	F1-Score	0.40	0.85	0.73		
Naive Baiyes	Precision	0.09	0.90	0.62	52.9%	0.46
	Recall	1.00	0.56	0.47		
	F1-Score	0.17	0.69	0.53		

From the comparison results, Random Forest conclusively demonstrates the best performance among the four models. With an F1-Score Macro Average of 0.75 and an accuracy reaching 85.3%, this model proves to be the most reliable. The superiority of Random Forest lies in its ability to effectively balance precision and recall across all three classes, especially for the minority 'High' class (F1-Score 0.50) and the majority classes 'Low' (F1-Score 0.91) and 'Medium' (F1-Score 0.84).

This finding is in line with previous work [16] who reported Random Forest achieving the highest predictive accuracy ($R^2 \approx 0.94$) in retail sales forecasting with imbalanced and seasonal demand patterns, outperforming simpler models such as Linear Regression and ARIMA. Similarly [17], further optimized Random Forest for complex retail data, achieving an R^2 of 0.945 and RMSLE of 1.172, surpassing Gradient Boosting, SVR, and XGBoost while excelling in interpretability and non-linear discount impact handling. Together, these studies reinforce the reliability of Random Forest for

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XI No 4, September 2025, hlm. 701 – 708 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/10.33330/jurteksi.v11i4.4187
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

capturing cosmetic sales complexities where traditional or parametric models like Logistic Regression often fail. These results collectively support the robustness of Random Forest in retail and e-commerce contexts, including cosmetic sales data where imbalanced categories are prevalent.

The practical implication of selecting this best model is realized through the implementation of a discount strategy recommendation system. Utilizing the superior Random Forest model, a simulation function was developed to predict the probability of a product achieving the 'High' sales category under various discount scenarios. For example, for a product with an original price of Rp 150,000, the model predicts that applying a 50% discount will increase the probability of achieving 'High' sales to 20.43%, the highest value among the tested scenarios (Image 2).

For example, for a product with an original price of Rp 150,000, the model predicts that applying a 50% discount will increase the probability of achieving 'High' sales to 20.43%, which is the highest value among the tested scenarios.

```
--- Rekomendasi Strategi Diskon untuk Produk ---
   Diskon (%)  Harga Baru (Rp) Probabilitas Penjualan Tinggi
0          10          135000                         6.08%
1          15          127500                         6.15%
2          20          120000                         5.85%
3          25          112500                         4.73%
4          30          105000                        11.73%
5          35           97500                        14.73%
6          40           90000                        18.67%
7          50           75000                        20.43%

>> Rekomendasi Terbaik <<
Diskon (%)                          50
Harga Baru (Rp)                  75000
Probabilitas Penjualan Tinggi    20.43%
Name: 7, dtype: object
```

Image 2. Visualization of 'High' Sales Probability Based on Discount Scenarios

This finding aligns with previous research [18], which used Random Forest to predict women's clothing retail sales during crises, achieving over 85% accuracy and simulating discount impacts that boosted minority class F1-Scores from 0.55 to 0.90. Such evidence validates the strategic utility of Random Forest in recommendation systems for volatile retail environments, strengthening the argument that cosmetic brands can rely on this model for more data-driven promotional planning.

CONCLUSION

This research concludes that the Random Forest model is the most superior method for predicting cosmetic product sales categories on e-commerce, following a systematic comparison with other algorithms and the application of SMOTE and GridSearchCV techniques. This model not only demonstrates the highest predictive performance but also reveals that variables related to pricing strategy, such as price difference and discounted price, are the most significant factors in determining the sales category. Nevertheless, this research has limitations related to the use of a static dataset, and future research can be developed by using longitudinal data to capture sales trends over time, enriching the model with external variables such as advertising campaign data, and exploring deep learning architectures for deeper analysis. For business practice, the findings highlight that companies can leverage Random Forest-based predictions to design more effective discount strategies and promotional planning, enabling data-driven decisions to boost competitiveness in highly dynamic e-commerce markets.

BIBLIOGRAPHY

[1] E. N. Lathifah, "Pengaruh Potongan Harga Terhadap Volume Penjualan Produk Skincare Pada Klinik Kecantikan Anye Medical Estetik Di Kelurahan Peranap, Kecamatan Peranap, Kabupaten Indragiri Hulu," vol. 10, no. 1, 2024.

[2] I. Rahmania dan A. Waris, "Pengaruh Price Discount dan Konsep Diri terhadap Keputusan Pembelian Produk Skintific pada Tiktok Shop," *CEMERLANG J. Manaj. Dan Ekon. Bisnis*, vol. 4, no. 4, hlm. 143–155, Okt 2024, doi: 10.55606/cemerlang.v4i4.3247.

[3] R. Johannes dan A. Alamsyah, "Sales Prediction Model Using Classification Decision Tree Approach For Small Medium Enterprise Based on Indonesian E-Commerce Data," *Eprint ArXiv210303117*, 2021, doi: https://doi.org/10.48550/arXiv.2103.03117.

[4] M. L. Prayugo, D. A. Wibowo, M. S. Hidajat, E. Mintorini, dan R. R. Ali, "Data Mining Application for Analyzing Pattern of Customer Purchase Using Apriori Algorithm," 2024.

[5] G. G. Pessanha dan E. A. Soares, "Apenas uma postagem? previsões de vendas diárias de empresas varejistas de beleza e cosmético a partir da influência de mídias sociais," *ReMark - Rev. Bras. Mark.*, vol. 20, no. 4, hlm. 241–266, Nov 2021, doi: 10.5585/remark.v20i4.17914.

[6] H. Jain, V. Dattpalsinh, S. K. Ray, dan Dr. Vishal, "Sales Prediction using Machine Learning," dalam Proceedings of the KILBY 100 7th International Conference on Computing Sciences 2023 (ICCS 2023), India: SSRN / KILBY 100 Committee, Apr 2024, hlm. 1–5. doi: http://dx.doi.org/10.2139/ssrn.4495850.

[7] F. S. Aditama, D. Krismawati, dan S. Pramana, "Multiclass Classification Of Marketplace Products With Machine Learning," MEDIA Stat., vol. 17, no. 1, hlm. 25–35, Okt 2024, doi: 10.14710/medstat.17.1.25-35.

[8] F. Fiddin, M. Y. Syahbarna, dan M. Ridwan, "Penggunaan Supervised Learning untuk Prediksi Validitas Ulasan Negatif Aplikasi Tokopedia Berdasarkan Pengalaman Pengguna Ahli," J. SAINTIKOM J. Sains Manaj. Inform. Dan Komput., vol. 23, no. 2, hlm. 409–417, Agu 2024, doi: 10.53513/jis.v23i2.10030.

[9] M. Syukron, R. Santoso, dan T. Widiharih, "Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class DatA," J. Gaussian, vol. 9, no. 3, hlm. 227–236, Agu 2020, doi: 10.14710/j.gauss.v9i3.28915.

[10] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, dan F. Zoromi, "Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang," MATRIK J. Manaj. Tek. Inform. Dan Rekayasa Komput., vol. 21, no. 3, hlm. 677–690, Jul 2022, doi: 10.30812/matrik.v21i3.1726.

[11] H. A. Fatan, T. Widiharih, Dan Sudarno, "Klasifikasi Kualitas Kopi Arabika Dengan Metode Random Forest Dan K-Nearest Neighbor Pada Imbalanced Dataset," vol. 14, no. 1, 2025.

[12] E. Constancio dan K. D. Tania, "Penerapan Metode Supervised Learning dan Teknik Resampling untuk Prediksi Penipuan Transaksi Keuangan," Build. Inform. Technol. Sci. BITS, vol. 6, no. 3, hlm. 1427–1439, Des 2024, doi: 10.47065/bits.v6i3.6110.

[13] D. S. Jayanthi, D. T. S. Kumari, S. Inturi, D. B. Nathan, J. Sathya, dan D. K. Karmakonda, "Predicting E-Commerce Revenue with SHAP Insights: A Comparative Study of SMOTE-Enhanced Machine Learning Models," Panam. Math. J., vol. 35, no. 4, 2025.

[14] M. M. R. Mubarak, Y. H. Chrisnanto, dan P. N. Sabrina, "Enrichment: Journal of Multidisciplinary Research and Development," 2023.

[15] E. Yolanda, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Data Pasien Rehabilitasi Narkoba," vol. 4, no. 1, hlm. 183, 2023, doi: 10.30865/klik.v4i1.1107.

[16] O. O. Mustapha dan Dr. T. Sithole, "Forecasting Retail Sales using Machine Learning Models," Am. J. Stat. Actuar. Sci., vol. 6, no. 1, hlm. 35–67, Apr 2025, doi: 10.47672/ajsas.2679.

[17] P. Ganguly dan I. Mukherjee, "Enhancing Retail Sales Forecasting with Optimized Machine Learning Models," dalam 2024 4th International Conference on Sustainable Expert Systems (ICSES), Okt 2024, hlm. 884–889. doi:10.1109/ICSES63445.2024.10762950.

[18] K. T. Kizgin, S. Alp, N. Aydin, dan H. Yu, "Machine learning-based sales forecasting during crises: Evidence from a Turkish women's clothing retailer," *Sci. Prog.*, vol. 108, no. 1, hlm. 00368504241307719, Jan 2025, doi: 10.1177/00368504241307719.