Comparative Analysis of Classification Models for Sales Prediction in E-commerce: Decision Tree, Random Forest, SVM, Naive Bayes, and KNN

Eko Purwanto*[1], Bangun Prajadi Cipto Utomo[2], Hanifah Permatasasi[3], Farahwahida Mohd[4]

[1]Information System, Universitas Duta Bangsa Surakarta, Indonesia
[2]Management, Universitas Duta Bangsa Surakarta, Indonesia
[3]Information System, Universitas Duta Bangsa Surakarta, Indonesia
[4]Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Malaysia

Email: [1]eko_purwanto@udb.ac.id

Received : Aug 3, 2025; Revised : Sep 9, 2025; Accepted : Dec 31, 2025; Published : Dec 31, 2025

Abstract

The swift expansion of e-commerce has markedly heightened the necessity for precise sales forecasting, essential for efficient marketing tactics and inventory control. This research evaluates five classification models—Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors (KNN)—to predict sales outcomes using e-commerce transaction data. The models were assessed utilizing criteria including accuracy, precision, recall, F1-score, AUC, and Log Loss. The findings indicate that Random Forest exceeds the performance of the other models, with an accuracy of 97.5% and an AUC of 0.991, markedly outperforming the alternatives. This study presents a unique contribution by contrasting these classification models in the realm of e-commerce in Indonesia, yielding significant insights for the advancement of more effective predictive algorithms in informatics. The results not only enhance the optimization of marketing strategies but also enrich the comprehension of machine learning applications in sales forecasting. This study underscores the necessity of choosing the appropriate model for enhanced sales forecasting, with considerable ramifications for data-driven decision-making in the e-commerce sector.

Keywords: Classification models, E-commerce, Machine learning, Predictive analytics, Random Forest, Sales prediction

1. INTRODUCTION

The increasing reliance on big data has fundamentally reshaped decision-making processes in various industries, particularly in e-commerce and retail. Businesses can now gather, process, and analyze extensive volumes of sales transaction data, offering significant opportunity for improving marketing tactics, enhancing customer experiences, and refining inventory management. In such a competitive business landscape, accurately predicting sales outcomes has become crucial for developing more targeted and effective strategies [1], [2]. Predictive analytics, driven by advanced machine learning models, is essential for delivering important insights that inform key business choices. [3]–[5]

Sales transactions are affected by various factors, including product features, pricing methods, and consumer characteristics. Data collected from each transaction, including transaction dates, product names, customer information, pricing, and quantity sold, contains significant potential for identifying patterns that can influence future purchasing behavior [6]. For instance, analyzing transaction dates can reveal seasonal purchasing trends, while attributes such as product names, prices, and categories provide insights into customer preferences and demand for specific products [7]. Additionally, information regarding the quantity sold and total sales amounts can help assess the profitability of products and guide inventory decisions [8].

Nonetheless, prior research on sales forecasting has predominantly concentrated on conventional models, such as Decision Trees and Naive Bayes, which frequently fail to encapsulate the intricate, non-linear correlations characteristic of e-commerce data [9]–[13]. Moreover, numerous current methodologies inadequately tackle the challenges presented by high-dimensional datasets, resulting in problems such as overfitting or suboptimal performance. This research offers a systematic comparative assessment of five machine learning models—Decision Tree, Random Forest, SVM, Naive Bayes, and K-Nearest Neighbors (KNN)—utilizing various evaluation criteria on e-commerce transaction data, in contrast to prior studies that concentrated on individual models [14]–[16]. This study will elucidate the most effective method for forecasting sales outcomes within the specific context of Indonesian e-commerce through a comparative analysis of different models.

The objective of predictive analytics in sales is to comprehend the interplay of many aspects and their influence on one another to accurately forecast future sales outcomes [1]. Classification models, which are widely used in predictive analytics, offer a powerful tool for grouping transactions based on relevant features and predicting the likelihood of future sales outcomes. These models can assist businesses in identifying patterns in past transactions, enabling them to forecast potential sales trends with higher accuracy. The ability to accurately predict sales allows companies to optimize their marketing campaigns and adjust their inventory management strategies to meet future demand [2], [17].

This study aims to evaluate and compare several classification models, including Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbors (KNN). These models were chosen for their capacity to manage intricate and high-dimensional datasets, exemplified by sales transactions. Random Forests and Decision Trees are proficient in managing extensive datasets and discerning non-linear patterns, but Support Vector Machines (SVM) are superior in scenarios with distinct margins between categories. Naive Bayes, despite being relatively simple, performs well in cases where the data follows a probabilistic distribution, and KNN can be effective when the dataset is relatively homogeneous [18]–[20].

The categorization models used in this study differ in their analytical methodologies. Decision Trees are intuitive and provide a transparent decision-making framework, Random Forests consolidate numerous decision trees to improve accuracy and mitigate overfitting [21], [22]. Support Vector Machines (SVM) are highly effective in high-dimensional spaces and complex datasets, making them ideal for intricate classification tasks [23]. Naive Bayes, grounded in probability theory, is particularly useful for predicting outcomes when features are independent, whereas KNN classifies data based on proximity within the feature space, making it a non-parametric model adept at capturing complex relationships[24].

This research seeks to create and evaluate models for predicting sales outcomes, concentrating on sales transactions that encompass critical characteristics such as product name, unit price, amount sold, transaction date, customer profile, and sales representative. By evaluating the performance of these models, the study seeks to provide insights into which methods offer the most reliable predictions for sales outcomes in an e-commerce setting. Moreover, the research will explore the practical implications of applying these models for optimizing marketing and inventory management strategies [5], [25].

This study aims to achieve two primary outcomes: first, to determine the most precise categorization model for forecasting sales results, and second, to provide actionable recommendations for firms seeking to integrate predictive analytics into their operations [2], [26]. Accurate sales predictions are essential for minimizing stockouts and overstock situations, enhancing customer satisfaction, and improving overall business efficiency [1]. By using historical transaction data, companies can forecast demand more precisely, guaranteeing the availability of appropriate products at the optimal time and in the correct amounts.

This study aims to advance predictive analytics in sales and marketing by comparing classification models. This research will elucidate the benefits and constraints of various machine learning techniques, enabling organizations to make more informed decisions regarding the selection of predictive models. The findings from this study will guide future research on the utilization of machine learning in sales forecasting and data-driven decision-making in the retail and e-commerce industries.

2. METHOD

This study seeks to create and evaluate various categorization models for forecasting sales outcomes derived from sales transaction data. The methodology includes multiple stages: data gathering, preprocessing, model implementation, and evaluation[27], [28]. Each phase is essential for ensuring the dependability and validity of the model's predictions. The next sections outline the detailed methods utilized in this research, as illustrated in Figure 1.

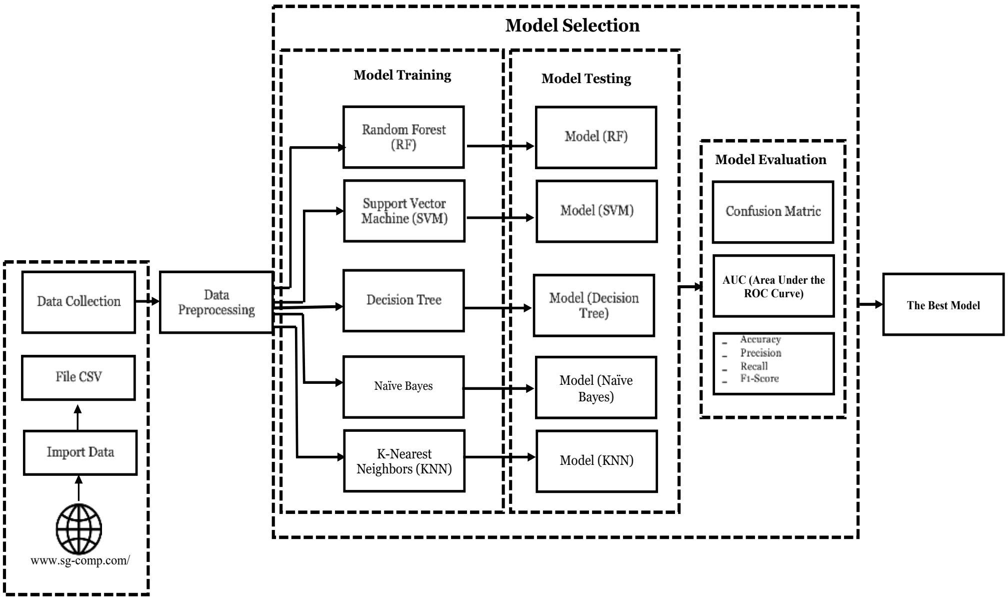


Figure 1. Comparison Model Methodology

2.1. Data Collection

This research employs sales transaction data sourced from the e-commerce platform www.sg-comp.com. Every transaction is characterized by a collection of essential features, as seen in Table 1.

These transaction data offer extensive insights into customer behavior, sales patterns, and product performance, all of which are crucial for building effective predictive models. The dataset spans several months and consists of 1,000 records, ensuring a varied range of transaction data for model training.

Table 1. Key Attributes Dataset

No	Attribute	Information
1	Transaction Date	Indicates the date of the sale
2	Product Name	Identifies the product sold
3	Customer Name	Specifies the customer making the purchase
4	Selling Price	The price at which the product is sold
5	Quantity Sold	The number of units of the product purchased
6	Total Transaction Amount	The total value of the transaction
7	Sales Representative Name	The name of the sales representative

2.2. Data Preprocessing

Before the implementation of machine learning models, several preparation procedures were conducted to ensure the data's quality and suitability for analysis, as illustrated in Figure 2.

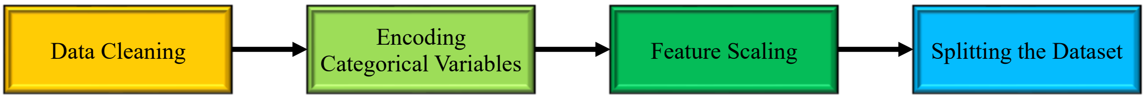


Figure 2. Data Preprocessing Steps

a. Data Cleaning
Identified incomplete or duplicate records were eliminated to ensure analytical correctness. Missing values in critical variables were imputed using appropriate methods, such as mean or median imputation for numerical data and mode imputation for categorical variables.

b. Encoding Categorical Variables
Categorical variables, including Product Name, Customer Name, and Sales Representative Name, were converted into numerical values by one-hot encoding or label encoding. This modification guarantees the machine learning algorithms can process the data effectively.

c. Feature Scaling
Continuous variables such as Selling Price, Quantity Sold, and Total Transaction Amount were standardized by Min-Max scaling or Robust Scaling to provide uniform contribution of all features to the model and mitigate bias arising from varying units or scales.

d. Splitting the Dataset
The dataset was partitioned into training and testing subsets utilizing an 80/20 split ratio. The training set was employed to develop the classification models, whereas the testing set was utilized for model assessment.

2.3. Model Selection

This study selected five prevalent classification models for comparison, each possessing distinct strengths and limits in managing complicated, high-dimensional sales transaction data:

a. Decision Tree
A computational framework that produces a hierarchical tree structure for decision-making predicated on attribute values. Decision Trees are readily interpretable and proficiently manage both numerical and categorical data. In a Decision Tree model, the algorithm segments the data based on the feature that provides the most information gain, utilizing metrics such as Gini impurity or Entropy.
Entropy Formula:

$$\boldsymbol{Entropy(S) = -\sum_{i=1}^{k} p_i \log_2 p_i} \tag{1}$$

Where S is the data set to be split, p_i is the probability of class i in dataset S, and k is the number of classes in the dataset.

Gini Index Formula:

$$\boldsymbol{Gini(S) = 1 - \sum_{i=1}^{k} {p_i}^2} \tag{2}$$

Where p_i is the probability of class i in dataset S.

b. Random Forest

An ensemble learning method that combines many Decision Trees to improve forecast accuracy and reduce overfitting. It is particularly beneficial for handling vast amounts of information and discerning non-linear relationships. Random Forest comprises an ensemble of several decision trees. Each tree utilizes a stochastic selection of attributes and information to alleviate overfitting. Random Forest produces predictions by consolidating the predominant vote from all trees in the ensemble.
Random Forest Prediction Formula:

$$\hat{y}_{RF} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \tag{3}$$

Where $\hat{y}_i$ is the prediction from the i -th tree, N is the number of trees in the ensemble, and $\hat{y}_{RF}$ is the average prediction of the ensemble.

c. Support Vector Machine (SVM)

A supervised learning model that determines the optimal hyperplane for class separation inside the dataset. Support Vector Machines (SVM) excel in high-dimensional spaces and are noted for their ability to handle complex data. Support Vector Machine identifies a hyperplane that separates the classes with the biggest margin. The decision function for the Support Vector Machine (SVM) is articulated as:
SVM Decision Function;

$$\boldsymbol{f(x) = \ w^T x + b} \tag{4}$$

Where w is the weight vector, x is the input feature vector, and b is the bias term. The goal is to maximize the margin by solving the following optimization problem:
Maximizing Margin:

$$Maximize \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{5}$$

Subject to:

$$\boldsymbol{y_i\left(w^T x_i + b\right) \geq 1 \ for\ all\ i} \tag{6}$$

Where $\boldsymbol{y_i}$ is the class label for data point $\boldsymbol{x_i}$.

d. Naïve Bayes

A probabilistic model based on Bayes' theorem, which asserts that features are independent. Despite its simplicity, Naive Bayes demonstrates robust performance on datasets with clear probabilistic distributions. Naïve Bayes is a probabilistic model based on Bayes' Theorem, which assumes that features are conditionally independent. The class probability C_k is defined by:
Bayes' Theorem:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{7}$$

Where $P(C_k|X)$ is the posterior probability of class C_k given the features X, $(X|C_k)$ is the likelihood, the probability of observing X given class C_k, $P(C_k)$ is the prior probability of class C_k, and $P(X)$ is the evidence, the total probability of the features X (normalization). The likelihood $P(C_k|X)$ is computed under the independence assumption:

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k) \qquad (8)$$

Where $x_i$is the value of the i -th feature.

e. K-Nearest Neighbors (KNN)

A non-parametric method that classifies a sample based on the dominating class of its nearest neighbors inside the feature space. The performance of KNN is affected by the choice of distance metric and may result in considerable computing expenses. KNN is a non-parametric method that classifies a data point based on the majority class of its k nearest neighbors.
KNN Prediction Formula:

$$\boldsymbol{y}^{\wedge}(\boldsymbol{x}) = \boldsymbol{majority}(\{\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_k}\}) \qquad (9)$$

Where $y_1, y_2, \dots, y_k$are the class labels of the k nearest neighbors, and $y^{\wedge}(x)$ is the predicted class for data point x. The distance between data points is often computed using the Euclidean distance:
Euclidean Distance:

$$\boldsymbol{d(x, x')} = \sqrt{\boldsymbol{\sum_{i=1}^{n}(x_i - x_i')^2}} \qquad (10)$$

Where x_i and x_i' are the feature values of two data points.

2.4. Hyperparameter Tuning

Default hyperparameters were initially employed for the training of each classification model. To enhance model performance, Grid Search was utilized for hyperparameter optimization. Grid Search entails a comprehensive examination of a manually defined hyperparameter grid to assess the performance of each combination through cross-validation. This procedure was utilized to ascertain the ideal configurations for hyperparameters such as max_depth, n_estimators, C (for SVM), and k (for KNN), hence enhancing the models' prediction performance and ensuring superior generalization to novel data.

2.5. Model Evaluation

The efficacy of each categorization model was assessed using various evaluation measures, facilitating a thorough appraisal of model performance:

a. **Accuracy**

The proportion of correctly predicted instances out of the total instances in the test set.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (11)$$

Where TP is True Positive (TP): Accurately predicted positive instances; True Negative (TN): Accurately predicted negative cases; False Positive (FP): Incorrectly predicted positive cases; False Negative (FN): Incorrectly predicted negative cases.

b. **Precision**

The ratio of accurate positive predictions to the total positive predictions generated by the model.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

c. **Recall**

The ratio of genuine positive predictions to the total number of real positive cases in the dataset.

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

d. **F1-Score**

The harmonic mean of precision and recall offers a compromise between these two criteria.

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \tag{14}$$

e. **AUC (Area Under the ROC Curve)**

A statistic that assesses the model's capacity to differentiate between positive and negative classes, with elevated values signifying superior performance.

$$AUC = \int_0^1 TPR(FPR)\, dFPR \tag{15}$$

Where TPR is True Positive Rate (also Recall), and $FPR\ is$ False Positive Rate.

f. **Log Loss**

A metric that evaluates the precision of probability forecasts, imposing more penalties for erroneous classifications associated with higher confidence levels.

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1 - y_i)log(1 - p_i)] \tag{16}$$

Where $N\ is$ Number of samples, y_i is True label for instance i, and p_i is Predicted probability for class 1 for instance i.

g. **Confusion Matrix**

A tabular depiction of true positives, true negatives, false positives, and false negatives, facilitating a comprehensive analysis of categorization efficacy.

	Predicted Positive	Predicted Negative
Actual Positive	*TP*	*FN*
Actual Negative	*FP*	*TN*

Where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative.

These measures were selected to evaluate the overall predictive efficacy of the models and their capacity to accurately classify both positive and negative outcomes.

3. RESULT

This section outlines the evaluation results for the classification models, including Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN), employing multiple performance metrics: Accuracy, Precision, Recall, F1 Score, AUC, and Log Loss.

Additionally, we assess the results using Confusion Matrices and ROC Curves to provide further insights into model effectiveness.

3.1. Model Evaluation

Table 2 below presents the evaluation results for each model, including Accuracy, Precision, Recall, F1 Score, AUC, and Log Loss.

Table 2. Comparison Model Evaluation

Model	Accuracy	Precesion	Recall	F1 Score	AUC	Log Loss
Decision Tree	0.98	1.00	0.90	0.95	0.95	0.72
Random Forest	0.97	1.0	0.88	0.94	0.99	0.09
SVM	0.92	1.00	0.64	0.78	0.92	0.22
Naïve Bayes	0.94	0.97	0.74	0.84	0.92	0.24
KNN	0.93	1.00	0.69	0.82	0.88	1.52

Table 2 illustrates that the Decision Tree model achieved an accuracy of 0.98, with precision and recall metrics indicating strong effectiveness in identifying positive classifications. An AUC value of 0.95 signifies the model's effectiveness in distinguishing between positive and negative classes; however, improvements in recall could enhance its ability to identify more genuine positives. The Random Forest model demonstrated an accuracy of 0.97 and a precision of 1.0, closely following the Decision Tree. It demonstrated a balanced performance, attaining a recall of 0.88 and an F1 score of 0.94. The AUC of 0.99 and Log Loss of 0.09 indicate that the Random Forest model is highly reliable in distinguishing between classes and producing precise predictions.

The SVM model achieved an accuracy of 0.92 and a precision of 1.0. Nonetheless, its recall was significantly lower (0.64), indicating that the model neglected a considerable number of actual positive occurrences. An F1 score of 0.78 signifies that the model struggled to reliably identify positive cases. The AUC of 0.92 and Log Loss of 0.22 indicate that the model demonstrates commendable performance; nonetheless, there is room for improvement, especially for recall.

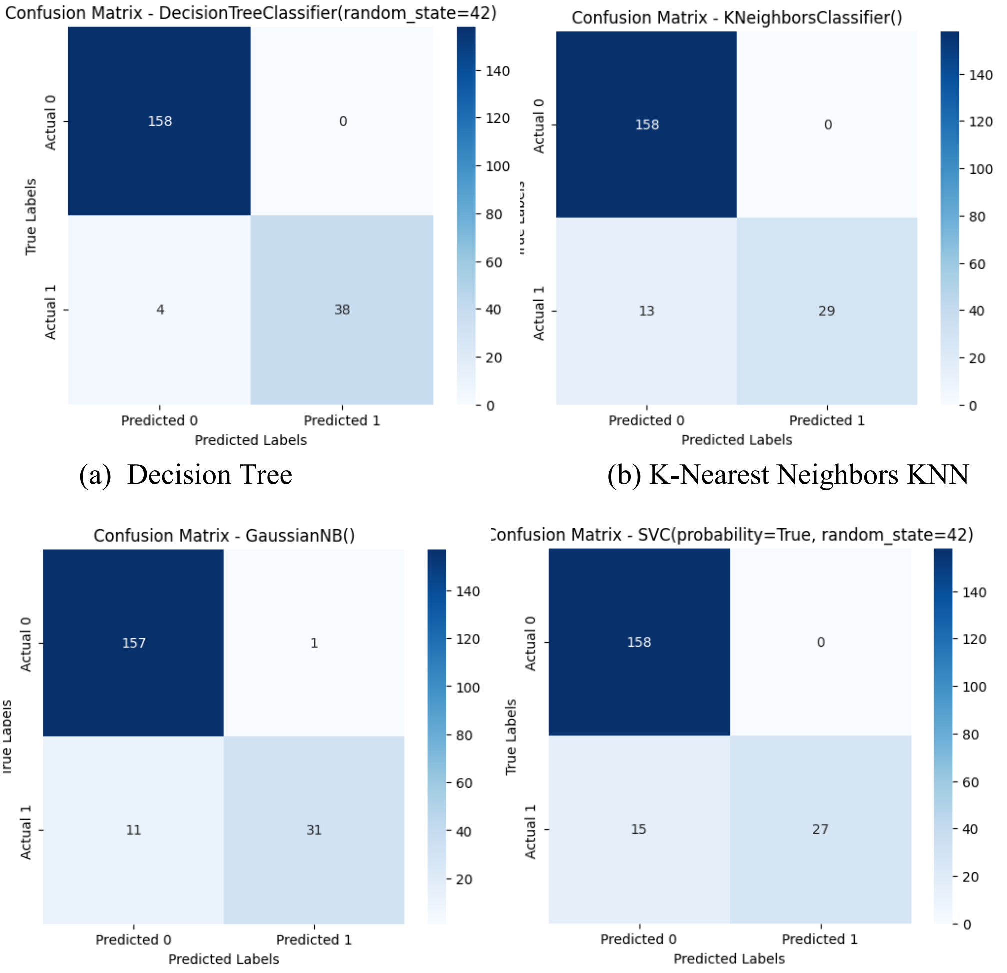
The Naïve Bayes model achieved an accuracy of 0.94, a precision of 0.97, and a recall of 0.74. The F1 score of 0.84 indicates a balanced performance, however it is inferior to that of the Decision Tree or Random Forest models. The AUC of 0.92 is strong, however the Log Loss of 0.24 suggests a reliable confidence in its predictions. The KNN model achieved an accuracy of 0.93 and a precision of 1.0. However, its recall of 0.69 and F1 score of 0.82 were the lowest among the models. An AUC of 0.88 signifies a reasonable ability of the model to differentiate between positive and negative classifications. The Log Loss of 1.52 was the highest across all models, indicating that KNN generated the least assured predictions.

3.2. Confusion Matrix

Figure 3 presents a comparative analysis of the Confusion Matrices for each model evaluated in this study.

The Confusion Matrix for the Decision Tree model (Figure 3a) indicates that, although it accurately identifies the majority of occurrences, it exhibits a number of false negatives. The Random Forest model (Figure 3e) demonstrates a reduced rate of misclassifications and an increased count of true positives, resulting in superior recall. The SVM (Figure 3d) has difficulty in accurately recognizing true positives, resulting in diminished recall. The KNN model (Figure 3b) exhibits a significant number of false negatives, further demonstrating its difficulty in identifying positive cases. The confusion matrix for the Decision Tree model demonstrates that the model correctly identified most positive and negative instances. Nonetheless, there were multiple false negatives, indicating that certain genuine positive occurrences were disregarded by the model. This is evident in the somewhat diminished recall relative to precision. A deeper investigation into the misclassified cases could provide insights into improving recall without sacrificing precision. The confusion matrix for Random Forest shows excellent performance with very few misclassified cases. It performed better in identifying true positives compared to Decision Tree, as seen in its higher recall value. Most of the misclassifications were false positives, suggesting that the model could benefit from further fine-tuning in terms of threshold selection to reduce unnecessary classifications of negative cases as positive.

The confusion matrix for the SVM indicates that the model accurately identified negative examples; nevertheless, it failed to recognize numerous real positives, as evidenced by the low recall. This suggests that SVM may require further parameter tuning, such as kernel optimization or adjusting the margin of error, to improve its sensitivity to positive cases. The confusion matrix for Naïve Bayes indicates that the model effectively differentiates positive examples, although it frequently misclassifies certain negative occurrences as positive. This is evident from its relatively low recall**,** which may be improved by exploring additional features or adjusting the decision threshold. The confusion matrix for KNN shows a high number of false negatives, contributing to its low recall. This suggests that KNN struggles to correctly identify positive instances, particularly when the data is noisy or the class distribution is skewed. The high Log Loss also indicates that KNN's predictions were less confident compared to other models.

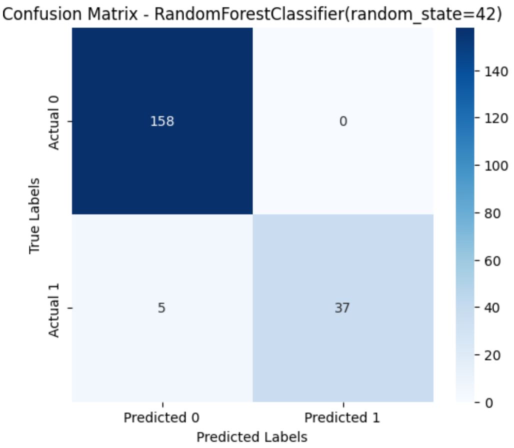


(a) Decision Tree

(b) K-Nearest Neighbors KNN

(c)Naïve Bayes

(d) Support Vector Machine (SVM)

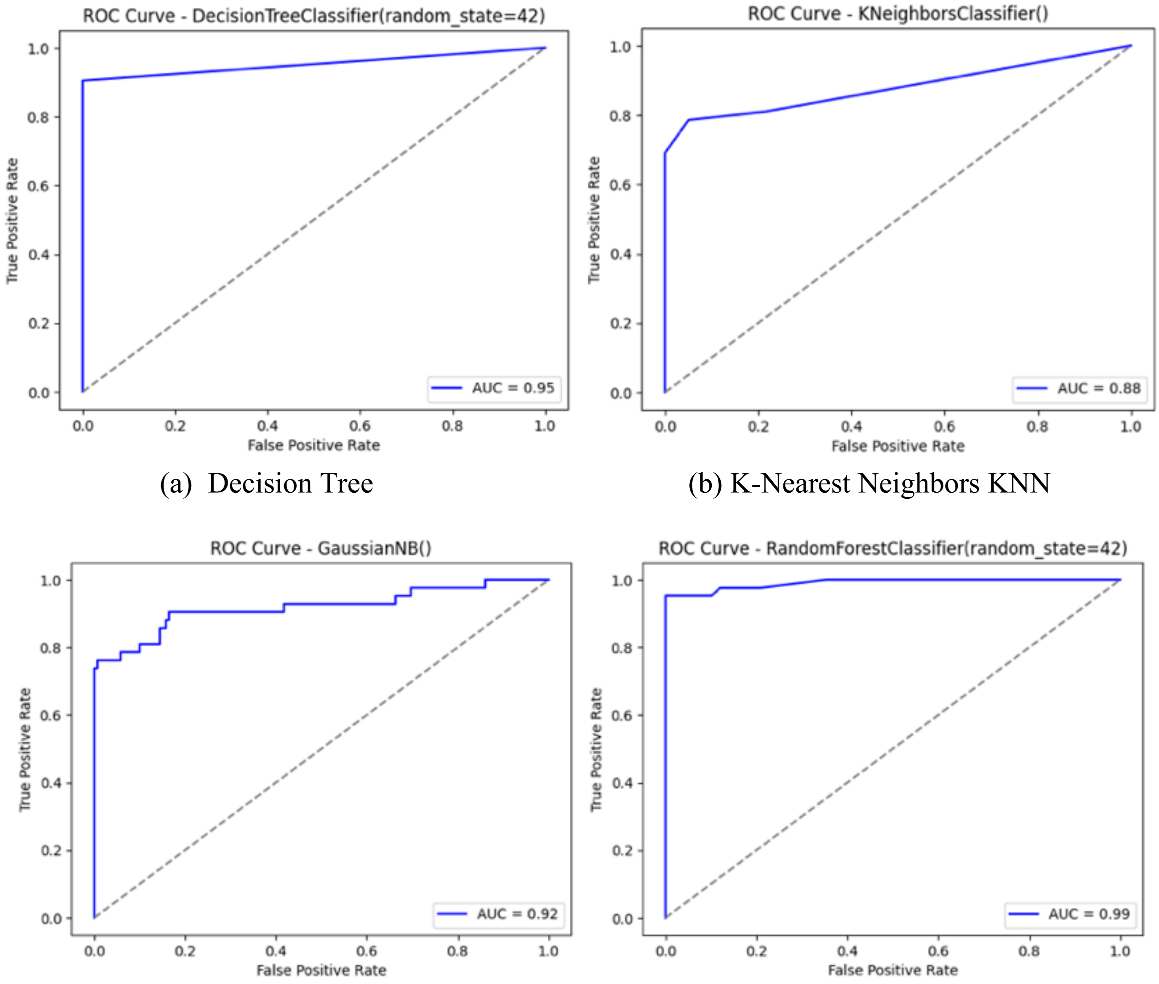


(e) Random Forest

Figure 3. Comparison Confusion Matric

3.3. AUC (Area Under the ROC Curve)

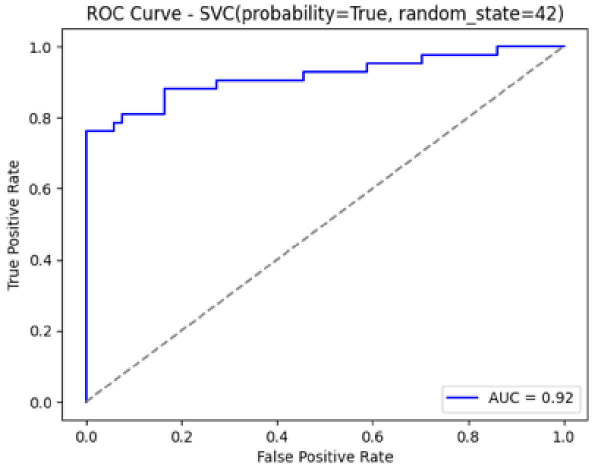
The AUC (Area Under the ROC Curve) assesses the model's capacity to differentiate between positive and negative classifications, as illustrated in Figure 4.



(a) Decision Tree

(b) K-Nearest Neighbors KNN

(c)Naïve Bayes (d) Random Forest



(e) Support Vector Machine (SVM)

Figure 4. Comparison AUC (Area Under the ROC Curve)

The ROC curve for the Decision Tree (4a) has a strong ability to distinguish between positive and negative classes, with an AUC of 0.952, signifying the model's considerable discriminatory power. The model has strong performance across several thresholds, while minor enhancements in recall may be attained by threshold optimization. The ROC curve for Random Forest (4d) distinctly illustrates its elevated AUC of 0.991, signifying outstanding classification proficiency. The curve remains around the top-left corner, demonstrating the model's strong discriminatory ability in predicting positive and negative outcomes. The ROC curve for SVM (4e) demonstrates moderate efficacy, with an AUC of 0.917. The curve demonstrates a degree of class separation, although implies potential for enhancement. The comparatively flat slope indicates that SVM may be less effective in differentiating positive and negative situations than alternative models such as Random Forest. The ROC curve for Naïve Bayes (4c) demonstrates strong performance with an AUC of 0.919, signifying the model's efficacy in class differentiation. Nonetheless, it does not exhibit the same efficacy as Random Forest or Decision Tree, indicating that more advanced models may produce superior outcomes. The ROC curve for KNN (4b) demonstrates the model's reasonable proficiency in differentiating between positive and negative cases, yielding an AUC of 0.879. The curve's position suggests that KNN possesses restricted discriminatory capability in comparison to models such as Random Forest and Decision Tree.

4. DISCUSSIONS

This study evaluates models—Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN)—that exhibit varying degrees of effectiveness in predicting sales outcomes from transaction data. This section evaluates the findings from the Results and examines the effectiveness of each model using confusion matrices, ROC curves, and various performance metrics.

The Random Forest model exhibited superior performance, achieving an accuracy of 0.975, a precision of 1.0, and an outstanding AUC of 0.991, indicating its remarkable ability to distinguish between positive and negative events. The confusion matrix for Random Forest indicated a minimal number of misclassifications, especially in false positives, implying that the model is proficient in accurately classifying both positive and negative transactions [29], [30]. This corresponds with prior research [31]–[34] indicating that Random Forest outperformed alternative models in e-commerce predictions.

Comparing Random Forest to Decision Tree, the latter showed a very similar performance with an accuracy of 0.98, but its slightly lower recall value indicates it missed a few true positive instances. This difference highlights the inherent advantage of Random Forest in handling data with more complexity, thanks to its ensemble approach, which reduces the likelihood of overfitting and improves generalization. Despite this, Decision Tree offers greater interpretability, which may be beneficial in environments where model transparency is important [35]–[37]. Nevertheless, Decision Tree offers greater interpretability, which may be beneficial in decision-making environments that require transparency, as seen in the study by [38], [39]. This makes Decision Tree particularly useful in contexts where model transparency is crucial, such as regulatory environments or decision-making processes that require clear and understandable rules.

The SVM attained an accuracy of 0.925; however, its recall of 0.643 was significantly inferior to that of the other models. The confusion matrix indicated that the SVM model has difficulty accurately classifying true positive cases, leading to false negatives. This constraint arises from the model's decision boundary, which may inadequately represent intricate relationships within the data. Future endeavors may involve enhancing the SVM model via kernel selection or hyperparameter optimization to improve its accuracy in identifying good outcomes [40]. Similar findings were noted by [41], where SVM struggled with data that did not have clear margin separation. Future work may include hyperparameter optimization and kernel selection to improve SVM’s performance on more complex datasets.

Naïve Bayes, while simpler and faster, showed reasonable performance with an accuracy of 0.94, but it exhibited a similar recall issue, with a value of 0.738. Its confusion matrix also indicated misclassifications, particularly among negative cases. The model's assumptions about feature independence may not be ideal in the presence of correlated features, as seen in this dataset, which limits its ability to capture more complex relationships [42]–[44]. This highlights a fundamental limitation of Naïve Bayes, especially in cases where features are correlated and cannot be treated as independent.

KNN showed the weakest performance among the models, with an accuracy of 0.935 and the highest Log Loss of 1.519, indicating its relatively low confidence in its predictions. The confusion matrix for KNN revealed a high number of false negatives, and the model had difficulty in detecting true positives. This is a common limitation of KNN, especially in high-dimensional data where the choice of distance metric significantly impacts performance. The ROC curve for KNN demonstrated modest discriminatory capability, although its AUC of 0.879 signifies worse effectiveness in differentiating between positive and negative classes relative to Random Forest and Decision Tree [45], [46]. This limitation is especially pronounced when working with high-dimensional data, where distance metrics become less reliable in accurately separating data points.

The ROC curves for each model visually depicted their capacity to differentiate between positive and negative cases. As expected, Random Forest demonstrated the greatest AUC value (0.991), indicating its exceptional capacity to distinguish between the classes accurately. The Decision Tree exhibited an AUC of 0.952, demonstrating robust performance, albeit with marginally inferior discrimination capability compared to the Random Forest [47], [48].

On the other hand, SVM and Naïve Bayes had moderate AUC values (0.917 and 0.919, respectively), demonstrating that while these models could distinguish between classes, they were not as robust as the ensemble-based methods. KNN showed the lowest AUC of 0.879, confirming its weaker ability to separate positive and negative instances in this specific dataset [49].

The findings indicate that Random Forest is the most appropriate model for forecasting sales outcomes, particularly in e-commerce contexts where precision and the capacity to manage intricate, non-linear interactions are crucial. The Decision Tree, while slightly less accurate, offers a high level of interpretability, making it useful for decision-making processes that require clear rules or explanations [34], [50].

For businesses that prioritize speed and computational efficiency, Naïve Bayes might be an option, but its performance may be suboptimal when feature dependencies exist. Similarly, SVM can be considered for datasets where margin separation is clear, but its ability to handle imbalanced datasets could be improved [51]. KNN's limitations in handling high-dimensional or noisy data make it less suited for applications where prediction accuracy is critical.

Notwithstanding its advantages, Random Forest is computationally demanding, particularly with extensive datasets. Future research may investigate improving the model for enhanced scalability or experimenting with hybrid models that include the advantages of various classifiers.. Additionally, feature engineering could be explored to extract more informative features that could further enhance model performance, particularly for SVM and KNN.

Additionally, subsequent research should investigate the effects of hyperparameter adjustment and cross-validation to enhance model performance, especially for algorithms such as SVM and Naïve Bayes, which may gain from these optimizations.

5. CONCLUSION

This study determined that Random Forest is the most effective model for predicting sales results, demonstrating superior accuracy, AUC, and overall performance. The Decision Tree demonstrated commendable performance, providing a high degree of interpretability, albeit with marginally reduced memory efficacy. Conversely, SVM and Naïve Bayes exhibited intermediate performance, with SVM displaying reduced recall and Naïve Bayes demonstrating efficiency although limited efficacy for intricate data patterns. The KNN model demonstrated the poorest performance, encountering misclassifications and elevated Log Loss.

The findings underscore Random Forest as the optimal choice for predictive tasks in e-commerce. Additionally, improvements in SVM and Naïve Bayes could be achieved through hyperparameter tuning or feature engineering. Future research should explore hybrid models and optimization strategies to enhance predictive accuracy in dynamic e-commerce settings.

This research significantly contributes to the field of Informatics by advancing the application of machine learning techniques in predictive analytics. It enhances our understanding of how data-driven models can optimize decision-making processes in various industries, especially e-commerce. The study provides a foundation for further research into machine learning optimization, hybrid models, and their integration into real-world systems, marking a meaningful contribution to the scientific community's understanding of applied machine learning in business decision-making.

CONFLICT OF INTEREST

The authors assert that there are no conflicts of interest pertaining to the publishing of this paper. All authors have made substantial contributions to the described work, and there are no financial or personal relationships that could improperly affect or distort the content of this publication.

ACKNOWLEDGEMENT

The authors express their profound gratitude to Universitas Duta Bangsa Surakarta for the support and facilities offered throughout the completion of this research. Our heartfelt appreciation also goes to the community partners and respondents for generously sharing their time and valuable insights, which significantly contributed to the success of this study.

We extend our heartfelt gratitude to the Ministry of Higher Education, Science, and Technology (Kemendikitsaintek) for their essential financial support, which was pivotal to the successful completion of this research. Without their funding, the completion of this study would not have been feasible.

REFERENCES

[1] M. Gafarov, "Predictive Analytics for Sales Forecasting and Inventory Management," *Next Gener. J. Young Res.*, vol. 8, no. 1, p. 109, 2024, doi: 10.62802/7t6wq430.

[2] A. R. Jakkula, "Predictive Analytics in E-Commerce: Maximizing Business Outcomes," *J. Mark. Supply Chain Manag.*, vol. 2, no. 2, pp. 1–3, 2023, doi: 10.47363/jmscm/2023(2)158.

[3] E. K. Akinyemi, A. I. Audu, O. A. O. A. P., and D. O. Ighawho, "Machine Learning Techniques in Predicting Sales a Case Study of Jumia," *Int. J. Res. Innov. Appl. Sci.*, vol. IX, no. XII, pp. 623–628, 2025, doi: 10.51584/ijrias.2024.912053.

[4] S. Bhujbal, V. Rajure, S. Shinde, D. Singh, Y. Bagul, and P. Agarkar, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," *Int. J. Sci. Technol. Eng.*, vol. 12, no. 5, pp. 4313–4316, 2024, doi: 10.22214/ijraset.2024.62603.

[5] H. Upadhyay, S. Shekhar, A. Vidyarthi, R. Prakash, and R. Gowri, "Sales Prediction in the Retail Industry Using Machine Learning: A Case Study of BigMart," in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 2023, pp. 1–6. doi: 10.1109/elexcom58812.2023.10370313.

[6] H. Chen, S. Yu, F. Huang, B. Zhu, L. Gao, and C. Qian, "Spatio-temporal Analysis of Retail Customer Behavior based on Clustering and Sequential Pattern Mining," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2020, pp. 284–288. doi: 10.1109/ICAIBD49809.2020.9137465.

[7] C. Xia and Y. Ma, "Sales data analysis and product layout analysis model based on association rule mining algorithm," in *Proc. SPIE 13447, International Conference on Mechatronics and Intelligent Control (ICMIC 2024)*, 2025, p. 133. doi: 10.1117/12.3045746.

[8] M. H. Rifqo, G. Gunawan, D. Sunardi, and I. Nurazizah, "Implementation Of Data Mining To Find Product Sales Patterns Using The Apriori Algorithm (Case Study: Warung Dini)," *J. Komputer, Inf. dan Teknol.*, vol. 4, no. 2, p. 11, 2024, doi: 10.53697/jkomitek.v4i2.2147.

[9] Q. Li and M. Yu, "Achieving Sales Forecasting with Higher Accuracy and Efficiency: A New Model Based on Modified Transformer," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 4, pp. 1990–2006, 2023, doi: 10.3390/jtaer18040100.

[10] Y. Rajalakshmi, T. Ammannamma, and D. Gudibandla, "Improving sales projections: a neural prophet-based approach for weekly forecasting," in *Futuristic Trends in Computing Technologies and Data Sciences Volume 3 Book 7*, IIP Series, 2024, pp. 115–120. doi: 10.58532/v3bkct7p1ch10.

[11] S. Jothiraj, S. I. Chellam, V. Rajeshwari, and C. K. Sri, "A Comprehensive Analysis of Predicting Future Sale and Forecasting Using Random Forest Regression," in *Industry Applications of Thrust Manufacturing: Convergence with Real-Time Data and AI*, IGI Global Scientific Publishing, 2024, pp. 177–196. doi: 10.4018/979-8-3693-4276-3.ch007.

[12] P. Ganguly and I. Mukherjee, "Enhancing Retail Sales Forecasting with Optimized Machine Learning Models," in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 2024, pp. 884–889. doi: 10.48550/arxiv.2410.13773.

[13] X. X. Liang *et al.*, "Product Sales Forecasting Model Driven by Multi-Source Data Integration Based on XGBoost," in *Frontiers in artificial intelligence and applications*, A. J. Tallón-Ballesteros, Ed. 2024, pp. 166–179. doi: 10.3233/faia241416.

[14] B. R. D. E. Oliveira, D. D. E. Simas, E. M. Frazzon, and M. Kück, "E-commerce sales forecasting: a product-based approach using consumer browsing data," in *ENEGEP 2024*, 2024, pp. 1–13. doi: 10.14488/enegep2024_tn_wpg_413_2033_47846.

[15] M. S. Chowdhury *et al.*, "Optimizing E-Commerce Pricing Strategies: A Comparative Analysis of Machine Learning Models for Predicting Customer Satisfaction," *Am. J. Eng. Technol.*, vol. 06, no. 09, pp. 6–17, 2024, doi: 10.37547/tajet/volume06issue09-02.

[16] L. Kumari, K. Bhattacharjee, N. Sharma, S. Kumar, and A. Kumari, "Machine Learning Models

in Customer Behaviour Prediction: A Comparative Analysis," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024, pp. 957–959. doi: 10.1109/ic3i61595.2024.10828637.

[17] G. R. Shrivas, "A Study of Impact and Applications of Predictive Analytics in Sales Forecasting," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 2, no. 3, pp. 12–20, 2023, doi: 10.22214/ijraset.2023.57535.

[18] J. E. Simarmata, G.-W. Weber, and D. Chrisinta, "Performance Evaluation of Classification Methods on Big Data: Decision Trees, Naive Bayes, K-Nearest Neighbors, and Support Vector Machines," *J. Mat. Stat. dan Komputasi*, vol. 20, no. 3, pp. 623–638, 2024, doi: 10.20956/j.v20i3.32970.

[19] Y. Oktafriani, G. Firmansyah, B. Tjahjono, and A. M. Widodo, "Analysis of Data Mining Applications for Determining Credit Eligibility Using Classification Algorithms C4.5, Naïve Bayes, K-NN, and Random Forest," *Asian J. Soc. Humanit.*, vol. 1, no. 12, pp. 1139–1158, 2023, doi: 10.59888/ajosh.v1i12.119.

[20] R. Suryawanshi, S. Musale, and S. Bhosale, "Comparative Analysis of use of Machine Learning Algorithm for Prediction of Sales," *J. Electr. Syst.*, vol. 20, no. 3, pp. 851–863, 2024, doi: 10.52783/jes.1383.

[21] X. Chang, "Comparative Analysis of Machine Learning, Decision Trees, and K-Nearest Neighbors for Heart Disease Prediction," *Appl. Comput. Eng.*, vol. 82, no. 1, pp. 188–192, 2024, doi: 10.54254/2755-2721/82/20241186.

[22] E. Halabaku and E. Bytyçi, "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests," *Intell. Autom. Soft Comput.*, vol. 39, no. 6, pp. 987–1006, 2024, doi: 10.32604/iasc.2024.059429.

[23] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, no. 1, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.

[24] G. Airlangga, "Analysis of Machine Learning Classifiers for Speaker Identification: A Study on SVM, Random Forest, KNN, and Decision Tree," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 1, pp. 430–438, 2024, doi: 10.47709/cnahpc.v6i1.3487.

[25] D. S. AbdElminaam, M. A. Mohamed, S. Khaled, F. Hany, M. Magdy, and Y. Sherif, "Leveraging Machine Learning for Accurate Store Sales Prediction: A Comparative Study," in *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2024, pp. 355–362. doi: 10.1109/miucc62295.2024.10783509.

[26] W. Ma, "Advanced Analytics for Retail Inventory and Demand Forecasting," *Trans. Econ. Bus. Manag. Res.*, vol. 10, no. 1, pp. 113–119, 2024, doi: 10.62051/jme9b319.

[27] U. Kulkarni *et al.*, "Future Sales Prediction Using Regression and Deep Learning Techniques," in *Lecture Notes in Electrical Engineering*, Singapore: Springer Science+Business Media, 2024, pp. 435–451. doi: 10.1007/978-981-99-7633-1_33.

[28] M. U. Ashraf, "A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques," *Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 04, pp. 23–33, 2022, doi: 10.54692/lgurjcsit.2022.0604399.

[29] B. Zhang, X. Qiao, H. Yang, and Z. Zhou, "A Random Forest Classification Model for Transmission Line Image Processing," in *International Conference on Computer Science and Education*, 2020, pp. 613–617. doi: 10.1109/ICCSE49874.2020.9201900.

[30] M. Maindola *et al.*, "Utilizing Random Forests for High-Accuracy Classification in Medical Diagnostics," in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024, pp. 1679–1685. doi: 10.1109/ic3i61595.2024.10828609.

[31] X. Shi, "The Application of Machine Learning in Online Purchasing Intention Prediction," in *International Conference on Big Data*, 2021, pp. 21–29. doi: 10.1145/3469968.3469972.

[32] R. Kasemrat and T. Kraiwanit, "Benchmarking Machine Learning Models for Predictive Analytics in E-Commerce," *Artif. Intell. eJournal*, vol. 12, no. 10, p. 429, 2024, doi: 10.2139/ssrn.4832967.

[33] S. M and D. N, "Regression Analysis-Based Predictive Model for E-Commerce Application," in

2023 International Conference on Networking and Communications (ICNWC), 2023, pp. 1–7. doi: 10.1109/ICNWC57852.2023.10127390.

[34] W. Zheng, “Customer Online Purchase Behavior Prediction and Performance Analysis Using Decision Tree and Random Forest,” *Sci. Technol. Eng. Chem. Environ. Prot.*, vol. 1, no. 6, pp. 1–8, 2024, doi: 10.61173/pncab928.

[35] A. Dorador, “Improving the Accuracy and Interpretability of Random Forests via Forest Pruning,” in *Proceedings of Machine Learning Research (PMLR)*, 2024, vol. 1, p. 240. doi: 10.48550/arxiv.2401.05535.

[36] A. Testas, “Random Forest Classification with Scikit-Learn and PySpark,” in *Distributed Machine Learning with PySpark*, Apress, 2023, pp. 243–258. doi: 10.1007/978-1-4842-9751-3_9.

[37] B. Gulowaty and M. Wozniak, “Extracting Interpretable Decision Tree Ensemble from Random Forest,” in *International Joint Conference on Neural Network*, 2021, pp. 1–8. doi: 10.1109/IJCNN52387.2021.9533601.

[38] B. Žlahtič, J. Završnik, H. B. Vošner, and P. Kokol, “Transferring Black-Box Decision Making to a White-Box Model,” *Electronics*, vol. 13, no. 10, pp. 1–16, 2024, doi: 10.3390/electronics13101895.

[39] J. Jagannathan, A. K., N. Labhade-Kumar, R. Rastogi, M. V. Unni, and K. K. Baseer, “Developing interpretable models and techniques for explainable AI in decision-making,” *Sci. Temper*, vol. 14, no. 04, pp. 1324–1331, 2023, doi: 10.58414/scientifictemper.2023.14.4.39.

[40] Jumanto *et al.*, “Optimizing Support Vector Machine Performance for Parkinson’s Disease Diagnosis Using GridSearchCV and PCA-Based Feature Extraction,” *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 1, pp. 38–50, 2024, doi: 10.20473/jisebi.10.1.38-50.

[41] M. Mittal, H. M. Al–Jawahry, N. Varshney, S. P. Kumar, J. J. Michaelson, and R. Reddy, “Improving Support Vector Machine Performance with Advanced Kernel Methods,” in *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)*, 2024, pp. 1749–1754. doi: 10.1109/ic3i61595.2024.10828664.

[42] P. J. B. Pajila, B. G. Sheena, A. Gayathri, J. Aswini, M. Nalini, and S. R, “A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications,” in *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, 2023, pp. 1228–1234. doi: 10.1109/icosec58147.2023.10276274.

[43] M. Schonlau, “The Naive Bayes Classifier,” in *Applied Statistical Learning*, Springer Nature, 2023, pp. 143–160. doi: 10.1007/978-3-031-33390-3_8.

[44] D. Prabha, J. Aswini, B. Maheswari, R. Subramanian, R. Nithyanandhan, and P. N. Girija, “A Survey on Alleviating the Naive Bayes Conditional Independence Assumption,” in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 2022, pp. 654–657. doi: 10.1109/ICAISS55157.2022.10011103.

[45] F. Acito, “k Nearest Neighbors,” in *Predictive Analytics with KNIME*, Springer Nature, 2023, pp. 209–227. doi: 10.1007/978-3-031-45630-5_10.

[46] A. M.D and P. K.K, “Addressing K-Nn Limitations Through Boosted Multi-Algorithm Nearest Neighbour Ensembles,” in *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, 2024, pp. 1804–1809. doi: 10.1109/iccpct61902.2024.10673192.

[47] Y. Gu, “A Comparative Analysis Study of Stock Prediction Based on Random Forest and Decision Tree,” in *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, 2024, pp. 96–100. doi: 10.1109/icedcs64328.2024.00022.

[48] E. Helmud, F. Fitriyani, and P. Romadiana, “Classification Comparison Performance of Supervised Machine Learning Random Forest and Decision Tree Algorithms Using Confusion Matrix,” *J. Sist. Inf. dan Komput.*, vol. 13, no. 1, pp. 92–97, 2024, doi: 10.32736/sisfokom.v13i1.1985.

[49] H.-H. Nguyen, “An Efficient Ensemble Algorithm for Boosting k-Nearest Neighbors Classification Performance via Feature Bagging,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 6, pp. 767–776, 2024, doi: 10.14569/ijacsa.2024.0150677.

[50] H. K. Alghamdi, S. M. Omar, and H. Namankani, “Predicting the Customer Behaviour Utilizing Tree Based Machine Learning Algorithms,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 12, no.

11, pp. 125–130, 2023, doi: 10.17148/ijarcce.2023.121118.

[51] M. R. Pahlawan, A. Setyanto, and M. Arief, “A Comprehensive Review of Clasifier used with Imbalanced Data in Machine Learning,” *J. Electr. Eng. Comput.*, vol. 6, no. 1, pp. 177–185, 2024, doi: 10.33650/jeecom.v6i1.8510.