Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

A Comparative Study of Machine Learning Classifiers with SMOTE for Predicting Purchase Intention

Khairunnisa Khairunnisa*, Sopian Soim, Lindawati Lindawati

Teknik Elektro, Teknik Telekomunikasi, Politeknik Negeri Sriwijaya, Palembang, Indonesia
Email:[1,*]wahyudikhairunnisa@gmail.com, [2]sopiansoim@gmail.com, [3]lindawati@polsri.ac.id
Correspondence Author Email: wahyudikhairunnisa@gmail.com
Submitted: **17/06/2025**; Accepted: **01/09/2025**; Published**: 02/09/2025**

Abstract−The rapid growth of e-commerce has made it increasingly important for online platforms to understand user behavior, particularly in predicting purchasing intention. This study examines the implementation of three machine learning models: Logistic Regression, Random Forest, and Gradient Boosting, to classify purchase intention using real transaction session data. One of the primary obstacles confronted in this investigation is the matter of class imbalance found in the dataset, where 10422 records indicate no purchase while only 1908 indicate a completed purchase. This disparity may result in a biased model performance that prioritizes the dominant class and limits the ability to accurately detect minority class behavior, which in this case is the actual purchase. To resolve this matter, During the data preprocessing phase, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented. Accuracy, precision, recall, and F1-score metrics were implemented to assess each model's functionality. The results indicate that following the implementation of SMOTE, the Random Forest model attained the best accuracy of 93%, succeeded by Gradient Boosting at 90% and Logistic Regression with 84%. These findings demonstrate that the use of SMOTE significantly improves model sensitivity and balance. This study provides useful insights into designing fairer and more effective predictive systems in the field of e-commerce.

Keywords: Class Imbalance; E-Commerce; Machine Learning; Predictive Modeling; SMOTE

1. INTRODUCTION

The exponential expansion of e-commerce in recent years has dramatically changed how consumers interact with online platforms [1], [2]. Understanding consumer behavior, particularly predicting whether a user intends to make a purchase, has become a crucial aspect of data-driven decision-making for online businesses[3], [4]. The ability to accurately predict purchase intention not only enables better targeting of marketing strategies but also supports optimization of user experiences and maximizes conversion rates [5], [6]. The application of machine learning techniques to develop prediction models has garnered considerable interest owing to their capacity to discern intricate patterns from extensive behavioral data [7], [8].

This study is focused on building a predictive model to identify whether a user intends to complete a purchase based on their session data. The target outcome is a binary label indicating purchase or no purchase. One of the primary obstacles in this task persists despite the intrinsic disparity inside real-world datasets, where the quantity of users who complete a purchase is significantly lower than those who do not. This imbalance often results in biased models that perform well on the majority class but fail to identify meaningful instances of the minority class, which in this case are the actual purchasers. Therefore, the study also incorporates a data preprocessing technique known as Synthetic Minority Over-sampling Technique, commonly referred to as SMOTE, which is designed to improve classifier performance by balancing the class distribution [9], [10], [11]. Each classifier was evaluated under two scenarios, before and after SMOTE implementation to thoroughly observe the impact of class balancing on model performance. Accordingly, this study aims to evaluate and compare the performance of logistic regression, random forest, and gradient boosting models in predicting purchase intention, and to investigate the effectiveness of SMOTE in addressing class imbalance.

The research compares three machine learning classifiers, namely logistic regression, random forest, and gradient boosting. These models were selected based on their popularity, interpretability, and diverse learning strategies. Logistic regression is a recognized statistical model that calculates the likelihood of a binary result through a linear amalgamation of input variables. It is frequently employed as a benchmark because to its simplicity and clarity of interpretation [12], [13], [14], [15]. Random forest, conversely, is an ensemble learning technique that generates numerous decision trees and consolidates their predictions. It is recognized for its resilience to overfitting and its capacity to elucidate intricate linkages within the data [16], [17]. Gradient boosting constructs an ensemble of trees sequentially, with each subsequent tree endeavoring to rectify the flaws of its predecessors. This approach generally attains superior predicting accuracy and can capture nuanced patterns within the data [18], [19].

This study utilizes the Online Shoppers Purchasing Intention Dataset, that constitutes a real-world dataset available on Kaggle [20]. It contains various features that represent user activity, such as the number of administrative and informational pages visited, session duration, bounce rates, exit rates, and categorical variables including month, visitor type, and weekend status. The target variable, labeled as Purchase Intention, indicates whether a user made a purchase. Categorical features were encoded prior to modeling to ensure compatibility with the machine learning algorithms. This dataset is particularly appropriate for the study because it presents a clear case of class imbalance and includes both numerical and categorical features, making it suitable for evaluating the robustness of different classifiers and the impact of oversampling methods.

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

SMOTE is applied in the preprocessing pipeline to address the imbalance in the dataset through the creation of synthetic samples for the minority class, utilizing the feature space similarities observed among existing minority instances [21], [22], [23]. This method ensures that classifiers encounter a more equitable distribution of classes throughout the training process, enhancing their capability to identify genuine purchasers rather than merely non-buyers. The application of SMOTE in this study seeks to investigate the impact on model performance through various classification metrics, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve.

The inclusion of both pre-SMOTE and post-SMOTE evaluations is a critical aspect of this research. It enables a direct comparison of model performance under imbalanced and balanced conditions, offering insights into how each algorithm adapts to class distribution shifts. It also helps reveal the trade-offs involved when employing synthetic sampling strategies. All classifiers in this study were trained using their default hyperparameters, allowing for a fair baseline comparison without additional tuning.

The choice of the three machine learning models is based on their widespread use in both academic research and industry applications. Logistic regression provides a strong foundation due to its transparency and theoretical grounding [15]. The random forest algorithm is favored due to being able to manage highly dimensional data sets and its inherent robustness against noise and overfitting [17], [24]. Gradient boosting is often considered a high-performance model suitable for structured datasets like the one used in this study [25]. By selecting these models, the study provides a fair and comprehensive comparison across a range of classification approaches, from simple linear to complex non-linear learners.

In summary, this study seeks to assess and contrast the efficacy of logistic regression, random forest, and gradient boosting classifiers in predicting purchase intention using an imbalanced e-commerce dataset. It investigates how SMOTE can enhance model performance in identifying true purchasers, which are typically underrepresented in such datasets. The study provides practical implications for data scientists and business analysts working in digital commerce environments who seek to improve the predictive accuracy of their user behavior models. The findings are expected to support the development of more balanced and effective recommendation and marketing systems that ultimately contribute to better business outcomes.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study adopts the SEMMA methodology to guide the stages of data mining and predictive modeling. SEMMA, which stands for Sample, Explore, Modify, Model, and Assess, is a structured framework developed by the SAS Institute that is widely used in data science and machine learning research [26]. Each stage in the SEMMA methodology corresponds to a specific step in the knowledge discovery process, allowing for systematic data preparation, modeling, and evaluation. The process in Figure 1, begins with the Sample phase, involving data collection and preparation, followed by Explore to analyze feature distributions and detect anomalies. The Modify phase includes data preprocessing, feature engineering, and handling class imbalance using SMOTE. The Model phase involves training three machine learning classifiers, such as Logistic Regression, Random Forest, and Gradient Boosting. Finally, the Assess phase evaluates the performance of each model using appropriate classification metrics.

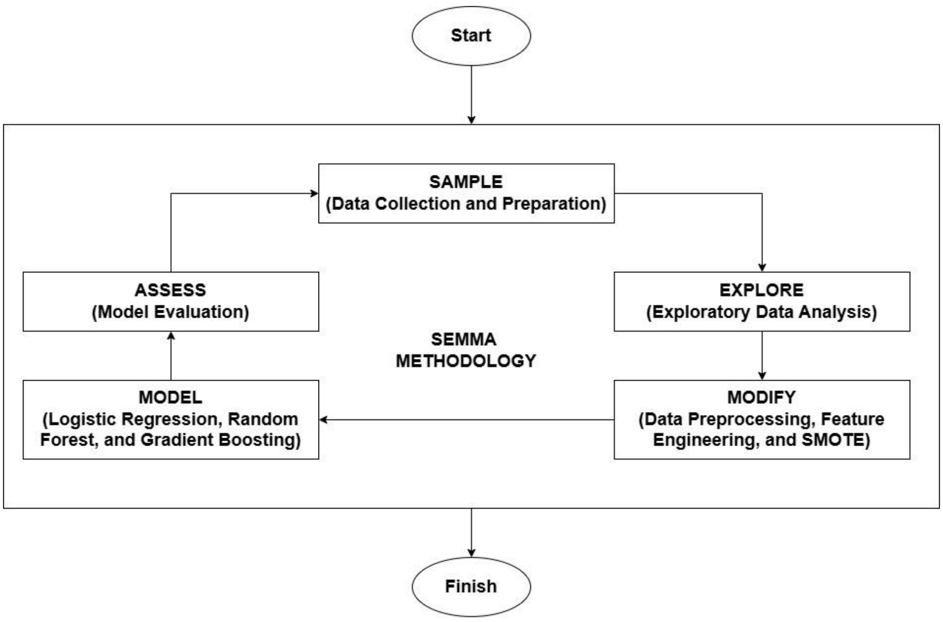


Figure 1. SEMMA Methodology

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

2.2 Sample

This research utilizes the publicly accessible Online Shoppers Purchasing Intention Dataset from Kaggle [20]. It contains session-level behavioral data of users interacting with an e-commerce website. This dataset consists of 12,330 rows and 18 attributes, including both numerical and categorical variables. Each row in the dataset represents a single user session and includes variables related to the number and duration of page visits, user engagement metrics, and technical information such as browser, operating system, and traffic source. The dataset includes a variety of features that capture multiple dimensions of user interaction. These features are categorized into three main types: integer, categorical, and the target variable. Features labeled as integer represent numerical data, including counts and durations of page visits as well as engagement metrics. These consist of variables such as Administrative, Informational, and ProductRelated, which reflect the number of pages viewed in each respective category.

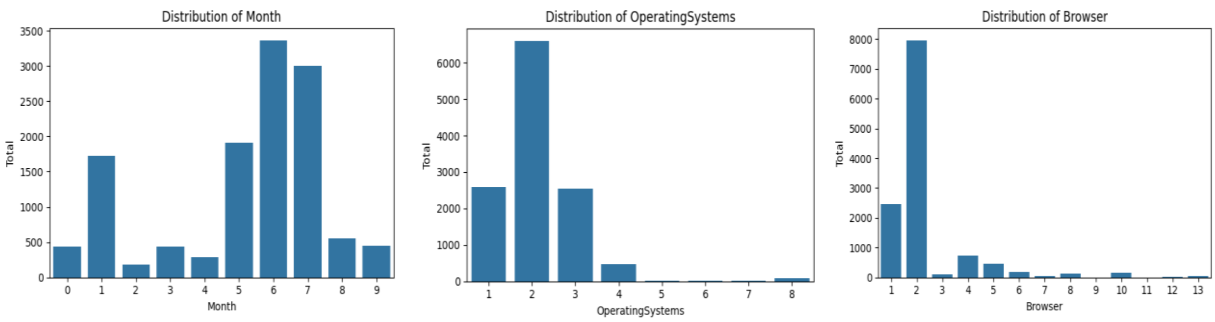
Table 1. Data Type

Feature	Data Type
Administrative	Numeral
Adiminstrative_Durations	Numeral
Informational_	Numeral
Informational_Durations	Numeral
ProductsRelated_	Numeral
ProductRelated_Durations	Numeral
Bounces_Rates	Numeral
Exit_Rates	Numeral
Pages_Values	Numeral
Specials_day	Numeral
Months_	Categorical
Operating_Systems	Numeral
Browsers_	Numeral
Region_	Numeral
Traffic_Types	Numeral
Visitor_Types	Categorical
Weekends	Categorical
Revenues	Class

2.3 Explore

The Explore phase in the SEMMA methodology aims to understand the dataset's structure, feature types, distributions, and potential issues such as missing values or anomalies. The dataset contains 12,330 rows and 18 columns, with no missing values. It includes 10 numerical features (e.g., Administrative, PageValues, BounceRates) and 4 categorical features (Month, VisitorType, Weekend, and Revenue). The target variable Revenue indicates purchase intention (True or False) and was label-encoded for modeling.

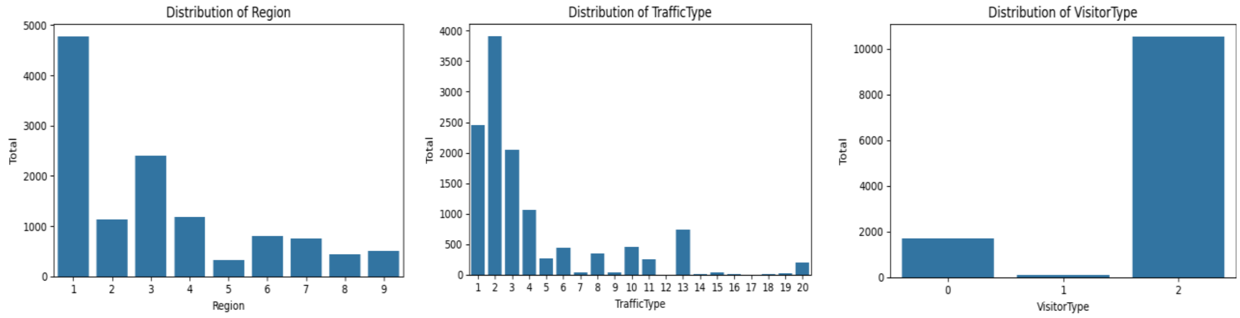
To analyze feature distributions, bar plots were grouped into Figure 2, Figure 3, and Figure 4, each showing six features. In Figure 2, categorical attributes Month (a), OperatingSystems (b), and Browser (c) are shown to be highly imbalanced, as are Region (d), TrafficType (e), and VisitorType (f). Figure 3 shows Weekend (a), Revenue (b), and numerical features such as Administrative (c), Administrative_Duration (d), Informational (e), and Informational_Duration (f), which exhibit right-skewed distributions. Figure 4 presents additional numerical features: ProductRelated (a), ProductRelated_Duration (b), BounceRates (c), ExitRates (d), PageValues (e), and SpecialDay (f), many of which show long-tail or sparse distributions. The class distribution of the target variable Revenue is also analyzed. As shown in Figure 3(b), the dataset is highly imbalanced, Approximately 84.5 percent of sessions do not lead to a purchase, while only 15.5 percent do result in a purchase. This disparity will be addressed in the Modify phase through class rebalancing techniques.



(a) (b) (c)

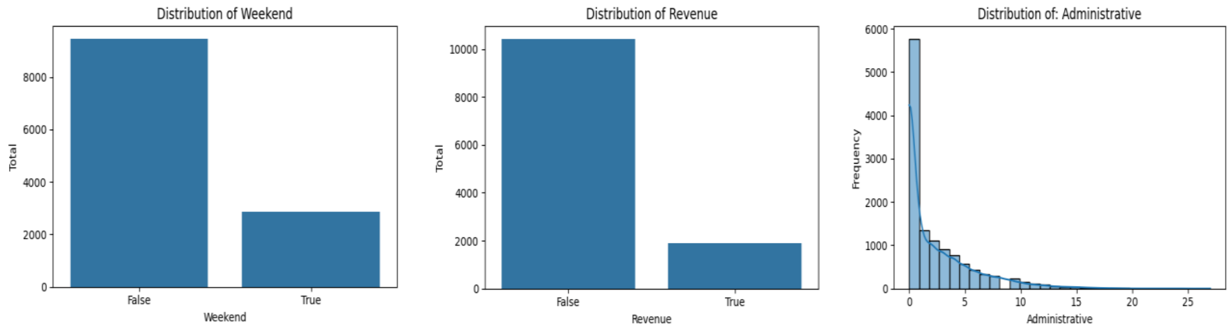
Copyright © 2025 **Author**, Page 995
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

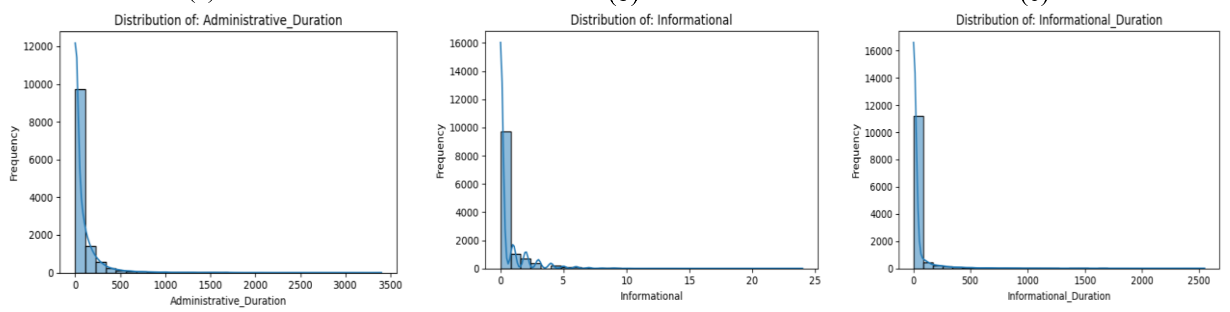


(d) (e) (f)

Figure 2. Distribution of Month, OperatingSystems, Browser, Region, TrafficType, and VisitorType

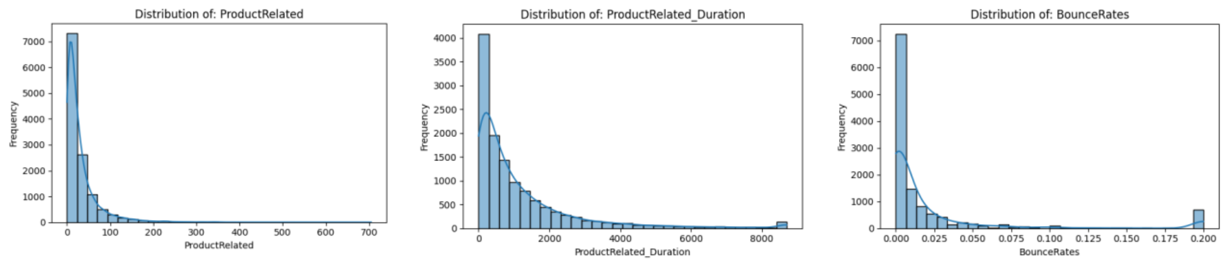


(a) (b) (c)

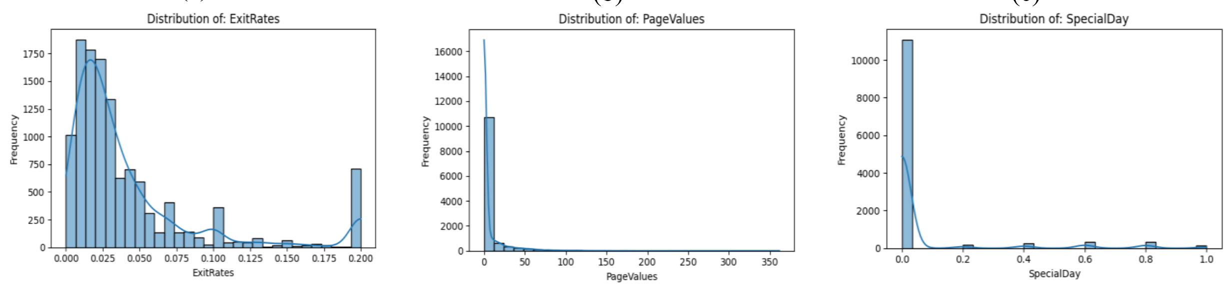


(d) (e) (f)

Figure 3. Distribution of Weekend, Revenue, Administrative, Administrative_Duration, Informational, Informational_Duration



(a) (b) (c)



(d) (e) (f)

Figure 4. Distribution of ProductRelated, ProductRelated_Duration, BounceRates, ExitRates, PageValues , and SpecialDay

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

The relationships between features were further examined using Pearson correlation analysis. The results of this analysis are visualized using a correlation heatmap, as shown in Figure 5. From the correlation heatmap, it is observed that ProductRelated and ProductRelated_Duration show a high correlation coefficient of 0.85, while BounceRates and ExitRates have the highest correlation at 0.91. These strong correlations are consistent with feature definitions, as longer interaction with product pages often increases product duration, and bounce activity typically aligns with session exits. Most other feature pairs show weak correlations, which suggests low multicollinearity across the dataset.

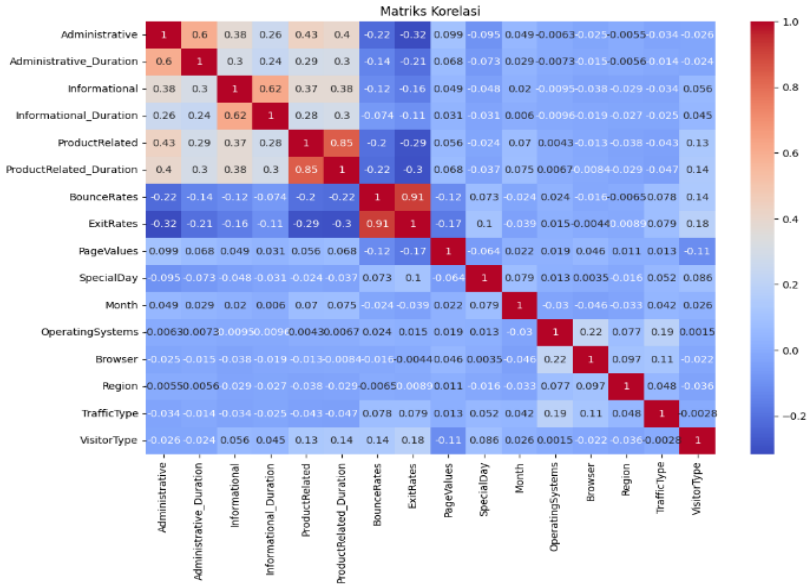


Figure 5. Correlation Heatmap

Outliers in the collected data were detected using the Interquartile Range (IQR) approach. A value is deemed an outlier if it deviates outside the lower or upper bounds [27], which are calculated using the following formulas:

$$UpperBound = Q3 + (1{,}5 \times IQR) \quad (1)$$

$$LowerBound = Q1 - (1{,}5 \times IQR) \quad (2)$$

The number and percentage of outliers detected per feature are summarized in Table 2. shows the number and percentage of data points identified as outliers for each feature in the dataset. Outliers are values that are significantly different from most of the other data and can negatively affect the performance of machine learning models. As shown in the table, the feature Browser has the highest number of outliers, with 4,369 entries, accounting for 35.43% of the data in that feature. Other features with a high number of outliers include PageValues (22.14%) and Informational (21.33%). This analysis is important to understand which features contain extreme values. The identified outliers were identified during this stage and later removed in the Modify phase using the Interquartile Range (IQR) method.

Table 2. Data Outlier

Feature	Outlier	Percentase (%)
Administratives	404	3.27%
Adiminstrative_Durations	1172	9.50%
Informational_	2631	21.33%
Informational_Durations	2405	19.50%
ProductsRelated_	987	8.01%
ProductRelated_Durations	961	7.79%
Bounces_Rates	1551	12.57%
Exit_Rates	1099	8.91%
Pages_Values	2730	22.14%
Specials_day	1251	10.14%
Months_	2160	17.51%
Operating_Systems	111	0.90%
Browsers_	4369	35.43%
Region_	511	4.14%
Traffic_Types	2101	17.03%
Visitor_Types	1779	14.42%

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

2.4 Modify

Modify is the third stage in the SEMMA methodology. It involves diverse data preparation methods, including the management of absent values, value substitution, categorical variable encoding, and executing feature engineering procedures such as oversampling and normalization. The following steps were carried out in this phase:

a. Outlier Removal
   Outliers identified during the Explore phase were removed using the IQR method. Upper and lower thresholds were calculated using:

$$UpperBound = Q3 + (1{,}5 \times IQR) \tag{3}$$

$$LowerBound = Q1 - (1{,}5 \times IQR) \tag{4}$$

b. Data Splitting
   After outlier removal, the dataset was partitioned into training and testing subgroups in an 80:20 ratio. The train_test_split function from scikit-learn was used with stratification to preserve the distribution of class labels in both subgroups. This ensures that the training and testing sets accurately reflect the original distribution of purchase and non-purchase sessions.
c. Feature Scaling
   Numerical characteristics were normalized with the StandardScaler technique. This transformation standardizes the data, ensuring each feature has a median value of zero and an average variability of one. The scaling algorithm was exclusively constructed on the training data and thereafter applied to both the training and testing datasets to avert any information leakage from the test set into the model during training.
d. Feature Selection
   Dimensionality was reduced and essential information was emphasized by feature selection with the SelectKBest technique, employing the ANOVA F-test as the scoring function. The first 20 characteristics were chosen according to their statistical significance to the target variable. This procedure enhances computing efficiency and mitigates the danger of overfitting.
e. Class Balancing with SMOTE
   To rectify the class imbalance identified during the experimental phase, the Synthetic Minority Over-sampling Technique (SMOTE) was used. SMOTE creates fresh synthetic samples of the minority class by interpolating between analogous existing cases in the feature space. This approach yields a more equitable dataset, enabling classification algorithms to learn proficiently from both classes and enhancing the model's capacity to identify minority class instances.

2.4 Model

In the Model phase of the SEMMA methodology, predictive algorithms are trained using the prepared dataset from the Modify phase. This stage is critical for identifying the model that best captures the underlying patterns of user behavior and effectively predicts purchase intention. To explore a variety of learning strategies, this study implements three different classification algorithms such as Logistic Regression, Random Forest, and Gradient Boosting. Each approach has a unique methodology for complexity, comprehension, and learning capability.

2.4.1 Logistic Regression

Logistic Regression is a classification method used to assess the likelihood of a binary result. It converts a linear assortment of input variables onto an estimated value ranging from zero to one using the logistic function. The log-odds of the target variable being equal to one is expressed as follows [28].

$$In\left(\frac{p}{1-p}\right) = B_0 + B_1X \tag{3}$$

Where p represents the probability of a positive outcome, $B0$ is the intercept, and $B1$ is the coefficient associated with the feature X. The probability itself is calculated by applying the sigmoid function to the linear equation

$$p = \frac{1}{1+e^{-(B_0+B_1X)}} \tag{4}$$

This model is widely used due to its simplicity and ease of interpretation. It is particularly useful when a straightforward relationship is assumed between independent variables and the target. Logistic Regression is often used as a baseline model in binary classification tasks.

2.4.2 Random Forest

Random Forest is a method that utilizes ensemble learning to build numerous decision trees, each based on various subsets of the training data. The method utilizes the bootstrap framework sampling and random feature selection to promote diversity among the trees. Every tree within the forest generates a prediction, and the ultimate result is established through majority voting among all the trees. This method minimizes variance and enhances model stability through the integration of several weak learners. Random Forest demonstrates strong capabilities in managing high-

Copyright © 2025 **Author**, Page 998
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

dimensional data and shows resilience against overfitting. This approach offers an assessment of feature significance by analyzing the extent to which each feature aids in diminishing impurity throughout the process of tree construction [29].

2.4.3 Gradient Boosting

Gradient Boosting is an iterative ensemble method that constructs models in a sequential manner. Every new tree addresses the prediction errors of the prior model by aligning with the not positive gradient of the decrease function. This leads to a steady enhancement of the overall predictive performance. The logistic loss is a widely utilized loss function for binary classification, and it can be articulated as follows[15].

$$-log\ L1 = -\sum_{i=1}^{N} y_i log(odds) + log(1 + e^{odds}) \quad (5)$$

Gradient Boosting is known for its ability to capture complex patterns in structured data and is often used in data mining and competition settings due to its high accuracy. Regularization techniques such as learning rate and tree depth control are typically applied to prevent overfitting.

2.4 Assess

The assess stage is an essential step in the SEMMA methodology, aimed at measuring the performance of classification models. Accuracy evaluation determines how closely the predicted labels match the actual values in the dataset. Among various evaluation methods, one commonly used indicator is Overall Accuracy (OA), which reflects the probability that a randomly selected instance is correctly classified. To obtain a more complete evaluation, this study also considers other performance metrics, namely Precision, Recall, and F1-score. These metrics are especially important when dealing with imbalanced data, where relying on accuracy alone may be misleading.

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (6)$$

$$F1 - Score = \frac{2 \times TP}{2 \times TP+FP+FN} \quad (7)$$

$$Precision = \frac{TP}{TP+FN} \quad (8)$$

$$Recall = \frac{TN}{TN+FP} \quad (9)$$

3. RESULT AND DISCUSSION

3.1 Modify

Modify represents the third phase of the SEMMA methodology, which involves a series of preprocessing techniques aimed at refining the dataset before modeling. At this stage, several essential processes are performed to enhance data quality and representativeness. These include outlier detection and removal, feature scaling, feature selection, and class balancing through the SMOTE technique. The goal is to ensure that the model built on this data can learn optimally and does not exhibit bias toward the majority class. Outlier detection and elimination are conducted using the Interquartile Range (IQR) method. Initially, the dataset was found to contain a total of 26,826 outliers across various features, as shown in Table 2. The highest proportions of outliers were identified in the features Browser (35.43 percent), PageValues (22.14 percent), and Informational (21.33 percent). These outliers have the potential to distort learning patterns, especially in models that are sensitive to the distribution of feature values. Consequently, the dataset was cleansed by removing records that fell beyond the upper and lower IQR bounds. Following the removal process, the total number of outliers was significantly reduced to 819 entries, as displayed in Table 3. This demonstrates a substantial improvement in data quality, dropping from the initial 26,826 outliers. Features such as Informational, ExitRates, PageValues, and SpecialDay no longer contained any outliers, while features like Administrative_Duration and BounceRates exhibited substantial reductions. This step ensures that the dataset used for training is free from extreme values that may otherwise bias the learning algorithms.

Table 3. Outlier Remover

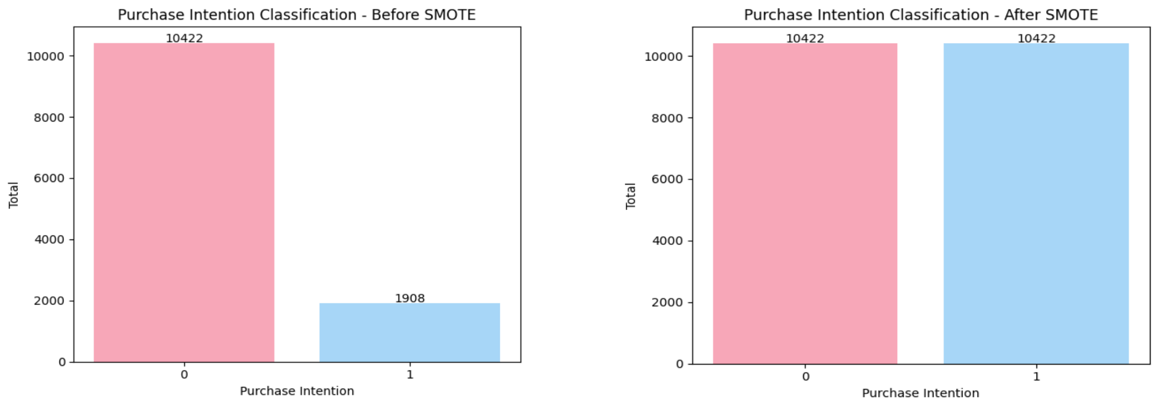
Feature	Outlier	Percentase (%)
Administratives	127	4.88%
Adiminstrative_Durations	196	7.54%
Informational_	0	0%
Informational_Durations	0	0%
ProductsRelated_	84	3.23%
ProductRelated_Durations	74	2.84%
Bounces_Rates	228	8.72%
Exit_Rates	0	0%

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

Feature	Outlier	Percentase (%)
Pages_Values	0	0%
Specials_day	0	0%
Months_	0	0%
Operating_Systems	0	0%
Browsers_	0	0%
Region_	0	0%
Traffic_Types	0	0%
Visitor_Types	110	4.23%

After addressing outliers, the dataset proceeds to the next step of class balancing using the Synthetic Minority Over-sampling Technique (SMOTE). This technique plays a critical role in tackling the class imbalance issue found in the original dataset. Prior to oversampling, the class distribution was highly skewed, with 10,422 samples in the non-purchase class (label 0) and only 1,908 samples in the purchase class (label 1), as shown in Figure 6(a). Such a significant imbalance caused classification models to be biased toward the majority class, resulting in high accuracy for predicting non-purchase sessions but poor performance in detecting actual purchase behavior.



(a) (b)

Figure 6. Class Distribution of Purchase Intention Before and After SMOTE Application

To address this, SMOTE was applied to generate synthetic samples for the minority class. The result of this balancing process is depicted in Figure 6(b), where both classes now contain an equal number of samples 10,422 for each class. This balanced distribution enables the learning algorithm to better generalize and improves the model's ability to detect purchasing intentions more effectively during the modeling phase. This adjustment is expected to enhance the model's capacity to correctly classify purchase behavior, especially in terms of recall and precision metrics that were previously low pertaining to the underrepresented category. Furthermore, the use of SMOTE was shown to enhance overall accuracy and F1-score performance.

3.2 Model

In this phase, three machine learning models were implemented and evaluated to predict purchase intention, namely Logistic Regression, Random Forest, and Gradient Boosting. The selection of each of them was based on their popularity, interpretability, and proven performance in classification tasks. Each model was trained and tested using the same dataset that had previously undergone a series of preprocessing steps as described in the Modify phase. To examine the effect of class imbalance on classification performance, each model was tested in two conditions: using the original imbalanced data and using the balanced data generated by the SMOTE technique. The evaluation metrics used in this study include accuracy, precision, recall, and F1-score. Additionally, confusion matrices were analyzed to provide a more detailed view of classification outcomes and model sensitivity across both classes. A comparative analysis of the results for each model under both conditions is described in the following subsections.

3.2.1 Logistic Regression

The performance results for the Logistic Regression model, both with and without the application of SMOTE, are presented in Table 4. Without SMOTE, the model achieved an overall accuracy of 88 percent. It demonstrated a strong recall for the negative class, reaching 98 percent. However, it performed poorly in detecting the positive class, with a recall of only 31 percent, indicating that the model struggled to identify instances of purchase behavior and was biased toward the majority class. Upon applying SMOTE, the recall for the positive class significantly improved to 78 percent. This improvement was accompanied by an increase in precision and a substantial rise in the F1-score. Although the overall accuracy slightly decreased to 84 percent, the model's ability to classify the minority class improved significantly, resulting in a more balanced and fair model. Table 4 illustrates the effectiveness of SMOTE

This Journal is licensed under a Creative Commons Attribution 4.0 International License

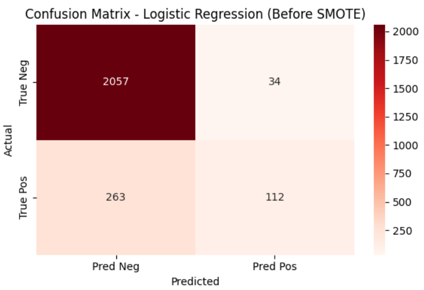
Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

in creating a more equitable learning environment for the Logistic Regression model, which allowed the classifier to better detect positive instances.

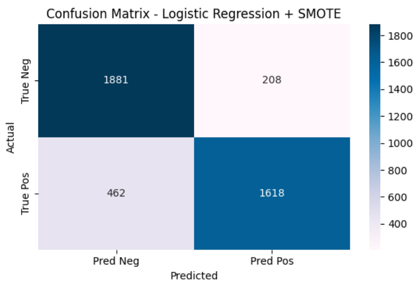
Table 4. Comparison of Logistic Regression Before and After SMOTE

Model	Sentiment	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	Negative	88%	89%	98%	93%
	Positive		79%	31%	45%
Logistic Regression + SMOTE	Negative	84%	80%	90%	85%
	Positive		89%	78%	83%

The impact of SMOTE on model performance is shown in the confusion matrices in Figure 7. Before applying SMOTE, in Figure 7(a) the model produced 2057 true negatives and only 112 true positives, with 263 false negatives. After SMOTE was applied, in Figure 7(b) the number of true positives increased significantly to 1618, although false positives also rose to 208. This trade-off led to improved model sensitivity, making the classifier more effective in detecting purchase intentions.



(a)



(b)

Figure 7. Comparison of Logistic Regression Confusion Matrix

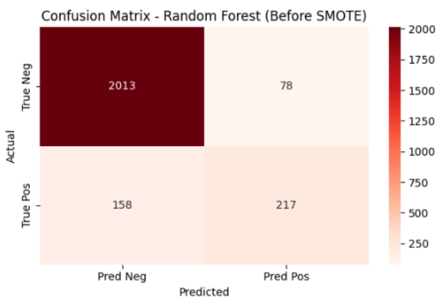
3.2.2 Random Forest

Table 5 summarizes the performance of the Random Forest model under both conditions. Without SMOTE, the model showed high accuracy at 90 percent, and demonstrated excellent classification performance for the negative class, with a recall of 96 percent and an F1-score of 94 percent. However, the recall for the positive class was limited to 57 percent. After implementing SMOTE, the model achieved substantial gains in performance for the positive class. The recall improved to 95 percent, and precision rose to 91 percent, resulting in an overall accuracy of 93 percent and balanced F1-scores for both classes above 90 percent. These results highlight Random Forest's strong capability in leveraging balanced data to enhance classification robustness.

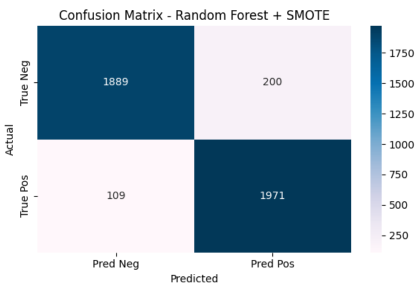
Table 5. Comparison of Random Forest Before and After SMOTE

Model	Sentiment	Accuracy	Precision	Recall	F-1 Score
Random Forest	Negative	90%	93%	96%	94%
	Positive		73%	57%	64%
Random Forest + SMOTE	Negative	93%	95%	90%	92%
	Positive		91%	95%	93%

A similar improvement is observed in the Random Forest model, as shown in Figure 8. Before applying SMOTE in Figure 8(a), the model correctly predicted 217 true positives and 2013 true negatives. After SMOTE was applied in Figure 8(b), the number of true positives increased significantly to 1971, with a slight drop in true negatives to 1889. This indicates that SMOTE effectively enhanced the model's ability to detect the minority class without sacrificing much precision.



(a)



(b)

Figure 8. Comparison of Random Forest Confusion Matrix

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

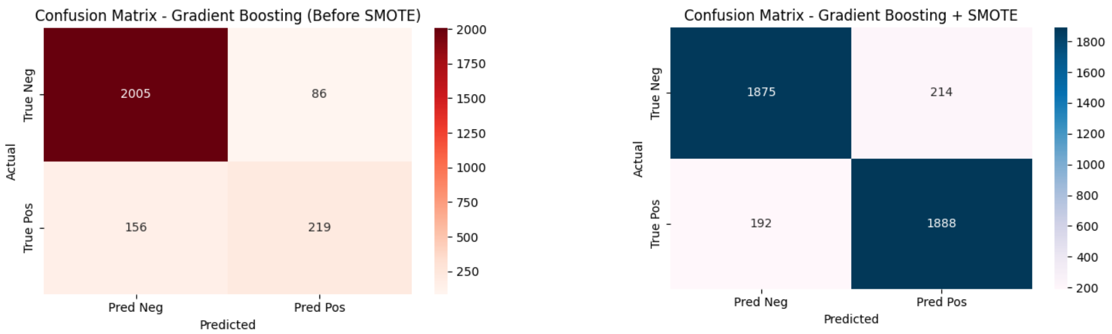
3.2.3 Gradient Boosting

Table 6 reports the classification results for the Gradient Boosting model. Initially, without SMOTE, the model obtained an accuracy of 90 percent. It performed very well on the negative class with a recall of 96 percent, but the positive class suffered from a recall of only 59 percent. This indicates a strong bias toward the majority class. After applying SMOTE, the model achieved balanced performance across both classes. The recall for the positive class increased to 91 percent and the F1-score reached 90 percent for both classes, indicating that the model could generalize well after balancing.

Table 6. Comparison of Gradient Boosting Before and After SMOTE

Model	Sentiment	Accuracy	Precision	Recall	F-1 Score
Gradient Boosting	Negative	90%	93%	96%	94%
	Positive		72%	59%	65%
Gradient Boosting + SMOTE	Negative	90%	91%	90%	90%
	Positive		90%	91%	90%

As shown in Figure 9(a), the Gradient Boosting model initially produced only 219 true positives and missed 156 purchase sessions. After applying SMOTE, the true positives increased significantly to 1888, as seen in Figure 9(b). Although false positives also rose slightly to 214, this improvement led to a more balanced and fair classification outcome. Although there was a slight increase in false positives, the improvement in overall balance and classification fairness outweighed the minor trade-offs.



(a) (b)

Figure 9. Comparison of Gradient Boosting Confusion Matrix

3.3 Discussion

The comparative analysis across the three machine learning models, namely Logistic Regression, Random Forest, and Gradient Boosting, reveals significant insights regarding model behavior, particularly in the presence of class imbalance and the effectiveness of the SMOTE technique. Table 7 presents a consolidated summary of model performance metrics including accuracy, precision, recall, and F1 score.

From Table 7, it is evident that the application of SMOTE led to considerable improvements in recall and F1 score for Logistic Regression. Initially, the Logistic Regression model showed an accuracy of 88 percent and a precision of 84 percent, but its recall was limited to only 65 percent. This suggests that while the model was relatively accurate in classifying non purchasing users, it underperformed in identifying purchase intentions. The use of SMOTE helped the model achieve better balance by increasing recall to 84 percent and F1 score to 84 percent as well. The overall accuracy slightly dropped to 84 percent, but this is acceptable given the significant gains in recall, indicating a more inclusive learning of the minority class.

Table 7. Model Accuracy

Model	Accuracy	Precision	Recall	F-1 Score
LR	88%	84%	65%	69%
LR + SMOTE	84%	84%	84%	84%
RF	90%	83%	77%	79%
RF + SMOTE	93%	93%	93%	93%
GB	90%	82%	78%	80%
GB + SMOTE	90%	90%	90%	90%

The Random Forest model showed strong baseline performance without SMOTE, with 90 percent accuracy, 83 percent precision, and 77 percent recall. However, applying SMOTE enhanced its effectiveness further. The Random Forest with SMOTE version achieved 93 percent across all metrics, including accuracy, precision, recall, and F1 score, demonstrating excellent generalization. The consistent increase in every metric confirms that Random Forest not only adapted well to balanced data but also benefited from it in terms of model fairness and robustness. Gradient

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Boosting also delivered competitive results in both conditions. Without SMOTE, the Gradient Boosting model achieved 90 percent accuracy, 82 percent precision, and 78 percent recall, with an F1 score of 80 percent. After SMOTE was applied, all evaluation metrics aligned at 90 percent, indicating that the model became more balanced and reliable. Unlike Logistic Regression, the Gradient Boosting model maintained its overall accuracy even after oversampling, showcasing its capacity to accommodate synthetic minority data without overfitting.

Across all models, the impact of SMOTE is clearly observable in recall and F1 score metrics, reinforcing the notion that while raw accuracy remains a useful metric, it is insufficient for evaluating classification performance in imbalanced datasets. Recall and F1 score offer a more nuanced view of a model's ability to capture the minority class, which in this context represents users who intend to make purchases. The confusion matrices for each model further support these findings, with Logistic Regression showing a drastic increase in true positives after SMOTE, highlighting improved sensitivity. Similarly, Random Forest and Gradient Boosting demonstrated a notable reduction in false negatives and an increase in correctly classified positive samples, which reinforces the effectiveness of class balancing. Additionally, it is worth noting that both ensemble methods, Random Forest and Gradient Boosting, consistently outperformed the simpler Logistic Regression across most scenarios, likely due to their capacity to model complex feature interactions and nonlinear relationships, particularly when the class distribution is balanced.

Despite the promising results, this study has several limitations. First, the experiment was conducted on a single dataset specific to e-commerce user sessions, which may limit the generalizability of the findings across different industries or user behaviors. Second, only three classification models were tested without hyperparameter tuning, and the class balancing relied solely on SMOTE without comparison to other resampling techniques such as ADASYN or undersampling. For future research, additional experiments using diverse datasets from other domains such as finance, healthcare, or clickstream data are recommended. Moreover, the inclusion of more advanced models such as XGBoost or deep learning classifiers, along with hyperparameter optimization and comparisons of multiple balancing techniques, may yield further insights and improve predictive performance.

4. CONCLUSION

This study aimed to evaluate the performance of various machine learning models in predicting online shopper purchasing intentions using the SEMMA methodology, which involves sequential stages of sampling, exploring, modifying, modeling, and assessing data. In the modify stage, outlier removal and SMOTE-based oversampling were applied to enhance data quality and address class imbalance. The experimental results demonstrated that data preprocessing significantly influenced model performance, especially in improving classification accuracy and sensitivity toward the minority class. Before balancing, the models tended to favor the dominant class of non-purchase sessions, resulting in high accuracy but low recall and F1 scores for the purchase class. After SMOTE was applied, the class distribution became balanced, leading to a substantial improvement in the ability of all models to detect purchasing intentions. Among the three models tested, Random Forest and Gradient Boosting consistently outperformed Logistic Regression in nearly all performance metrics, both before and after the application of SMOTE. These ensemble-based models are more effective in capturing complex patterns and handling feature interactions. The effectiveness of SMOTE in enhancing the model's ability to generalize and detect minority class instances is evident from the improved recall and F1-score values. However, limitations remain, including the static nature of the dataset and the lack of temporal features. Future research could address these limitations by incorporating real-time data, exploring advanced oversampling techniques like ADASYN, or leveraging deep learning approaches. Researchers are also encouraged to ensure ethical use of behavioral data and to focus on model interpretability to promote transparency in machine learning applications.

REFERENCES

[1] M. B. Gulfraz, M. Sufyan, M. Mustak, J. Salminen, and D. K. Srivastava, "Understanding the impact of online customers' shopping experience on online impulsive buying: A study on two leading E-commerce platforms," *Journal of Retailing and Consumer Services*, vol. 68, Sep. 2022, doi: 10.1016/j.jretconser.2022.103000.

[2] A. Rahaman, P. Hulgutte, S. Shaligram, and S. P. Pawar, "Advancements in Diagnostic Strategy of Neurological and Neuropsychiatric Disorders: From Conventional Methods to Point-of-Care Approaches," *British Journal of Multidisciplinary and Advanced Studies*, vol. 5, pp. 1–14, Sep. 2024, doi: 10.37745/bjmas.2022.04184.

[3] M. S. Azad, S. S. Khan, R. Hossain, R. Rahman, and S. Momen, "Predictive modeling of consumer purchase behavior on social media: Integrating theory of planned behavior and machine learning for actionable insights," *PLoS One*, vol. 18, Dec. 2023, doi: 10.1371/journal.pone.0296336.

[4] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decis Support Syst*, vol. 149, Oct. 2021, doi: 10.1016/j.dss.2021.113622.

[5] D. C. Gkikas and P. K. Theodoridis, "Predicting Online Shopping Behavior: Using Machine Learning and Google Analytics to Classify User Engagement," *Applied Sciences (Switzerland)*, vol. 14, Dec. 2024, doi: 10.3390/app142311403.

[6] S. Jayanthi, D. Rajeshwari, N. M. Goud, R. Geetha, S. B. Franklin, and P. Rajyalakshmi, "Optimizing Purchase Intention Prediction in E-Commerce," in *2024 1st International Conference for Women in Computing, InCoWoCo 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/InCoWoCo64194.2024.10863606.

[7] R. Gupta, A. Sharma, and T. Alam, "Building Predictive Models with Machine Learning," in *Studies in Big Data*, vol. 145, Springer Science and Business Media Deutschland GmbH, 2024, pp. 39–59. doi: 10.1007/978-981-97-0448-4_3.

Copyright © 2025 **Author**, Page 1003
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 993−1004
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7615

[8] M. Arunkumar, K. Rajkumar, W. R. Salem Jeyaseelan, and N. A. Natraj, "Data Mining, Machine Learning, and Statistical Modeling for Predictive Analytics with Behavioral Big Data," *Tehnicki Vjesnik*, vol. 32, pp. 72–77, 2025, doi: 10.17559/TV-20231102001073.
[9] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowl Based Syst*, vol. 248, Jul. 2022, doi: 10.1016/j.knosys.2022.108839.
[10] S. A. Alex, J. Jesu Vedha Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl Soft Comput*, vol. 156, May 2024, doi: 10.1016/j.asoc.2024.111491.
[11] F. Kamalov, A. F. Atiya, and D. Elreedy, "Partial Resampling of Imbalanced Data," Jul. 2022.
[12] F. E. Harrell, *Regression Modeling Strategies*. in Springer Series in Statistics. New York, NY: Springer New York, 2001. doi: 10.1007/978-1-4757-3462-1.
[13] E. Tahirovic and S. Krivic, "Interpretability and Explain ability of Logistic Regression Model for Breast Cancer Detection," in *International Conference on Agents and Artificial Intelligence*, Science and Technology Publications, Lda, 2023, pp. 161–168. doi: 10.5220/0011627600003393.
[14] A. Cemiloglu, L. Zhu, A. B. Mohammednour, M. Azarafza, and Y. A. Nanehkaran, "Landslide Susceptibility Assessment for Maragheh County, Iran, Using the Logistic Regression Algorithm," *Land (Basel)*, vol. 12, Jul. 2023, doi: 10.3390/land12071397.
[15] N. A. Saran and F. Nar, "Fast binary logistic regression," *PeerJ Comput Sci*, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2579.
[16] M. Mohammadagha, "Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction: A Comparative Study of AdaBoost, Decision Trees, and Random Forest," *SSRN Electronic Journal*, 2025, doi: 10.2139/ssrn.5226457.
[17] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.
[18] A. Thakur *et al.*, "Product Length Predictions with Machine Learning: An Integrated Approach Using Extreme Gradient Boosting," *SN Comput Sci*, vol. 5, Aug. 2024, doi: 10.1007/s42979-024-02999-8.
[19] J. Li, P. Liu, L. Chen, W. Pedrycz, and W. Ding, "An Integrated Fusion Framework for Ensemble Learning Leveraging Gradient Boosting and Fuzzy Rule-Based Models," *IEEE Transactions on Artificial Intelligence*, 2024, doi: 10.1109/TAI.2024.3424427.
[20] A. Shamim, "Predictive Modeling of E-Commerce Purchase Intent," https://www.kaggle.com/datasets/adilshamim8/online.
[21] B. Ghojogh, M. Crowley, F. Karray, and A. Ghodsi, "Adversarial Autoencoders," in *Elements of Dimensionality Reduction and Manifold Learning*, Springer International Publishing, 2023, pp. 577–596. doi: 10.1007/978-3-031-10602-6_21.
[22] A. Bernardo and E. Della Valle, "An extensive study of C-SMOTE, a Continuous Synthetic Minority Oversampling Technique for Evolving Data Streams," *Expert Syst Appl*, vol. 196, Jun. 2022, doi: 10.1016/j.eswa.2022.116630.
[23] S. A. Alex, J. Jesu Vedha Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl Soft Comput*, vol. 156, May 2024, doi: 10.1016/j.asoc.2024.111491.
[24] G. Kunapuli, *Ensemble Methods for Machine Learning*. Simon and Schusters, Manning, 2023. Accessed: Jun. 08, 2025. [Online]. Available: https://search.worldcat.org/title/1266357525
[25] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, Apr. 2023, doi: 10.1002/eng2.12599.
[26] F. Sulianta, *Basic Data Mining from A to Z - Feri Sulianta - Google Books*. 2023. Accessed: Jun. 23, 2025. [Online]. Available: https://books.google.co.id/books?hl=en&lr=lang_en&id=JcLhEAAAQBAJ&oi=fnd&pg=PA1&dq=metodologi+semma&ots=VnDoPkWIrp&sig=BzMu92d48476WZ6-oo7fQfLEUYw&redir_esc=y#v=onepage&q=metodologi%20semma&f=false
[27] M. E. Lestari, I. Asror, and I. L. Sardi, "Penerapan PCA (Principal Component Analysis) pada Deteksi Outlier untuk Data Text," *eProceedings of Engineering*, vol. 10, no. 3, Jun. 2023, doi: 10.1016/j.jsb.2012.10.010.
[28] F. D. Pramakrisna, "Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic Regression." TEKNIKA, vol. 11, no. 2, 2025. Available: https://ejournal.ikado.ac.id/index.php/teknika/article/view/466/206
[29] K. A. Khalim, U. Hayati, and A. Bahtiar, "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest Dan Naïve Bayes," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 498–504, Mar. 2023, doi: 10.36040/JATI.V7I1.6376.

Copyright © 2025 **Author**, Page 1004
This Journal is licensed under a Creative Commons Attribution 4.0 International License