PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

# CLASSIFICATION OF CUSTOMERS' REPEAT ORDER PROBABILITY USING DECISION TREE, NAÏVE BAYES AND RANDOM FOREST

**Amelia Citra Dewi[1*]; Arief Hermawan; Donny Avianto[3]**

Master of Information Technology[1,2,3]
Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia[1,2,3]
https://uty.ac.id/[1,2,3]
ameliacitradewi@gmail.com[1*]; ariefdb@uty.ac.id[2]; donny@uty.ac.id[3]
(*) Corresponding Author



***Abstract***—*Limited customer information in sales data on e-commerce in Indonesia hinders companies in determining targeted marketing strategies, especially in targeting groups of potential customers to make repeat purchases. Sales data in the form of customers' names and cellphone numbers has been hidden by e-commerce, and only data is available in the form of products purchased, number of purchases, and customer addresses. So far, the methods used to determine potential customers mostly use more complete data features. Research that uses limited e-commerce data to determine potential customers is scarce. Several algorithms for predicting repeat purchases in e-commerce also have been widely used. However, the comparison of the performance of these methods in the context of e-commerce in Indonesia with limited data has yet to be discovered. In this research, the Decision Tree, Naive Bayes, and Random Forest methods were compared to classify potential customers using Maschere brand sales data from two e-commerce sites, namely Tokopedia and Shopee. The research results show that the Decision Tree algorithm achieved an accuracy of 90.91%, Naive Bayes achieved an accuracy of 37.50%, and Random Forest achieved the best level of accuracy, namely 93.94%. These results show that the Random Forest method is the best method for classifying customers' probability of repeat purchases. In the future, the results of this research can be developed again as a decision-making system to determine potential customers.*

***Keywords:*** *customer classification, decision tree, e-commerce, naive bayes, random forest.*

**Abstrak**—Keterbatasan informasi pelanggan dalam data penjualan pada e-commerce di Indonesia menghambat perusahaan dalam menentukan strategi marketing yang terarah, terutama dalam menargetkan kelompok calon pelanggan potensial untuk melakukan pembelian berulang. Data penjualan berupa nama dan nomor handphone pelanggan sudah disembunyikan pihak e-commerce, dan hanya tersedia data berupa produk yang dibeli, jumlah pembelian, beserta alamat pelanggan. Selama ini metode yang digunakan untuk menentukan pelanggan potensial sebagian besar menggunakan fitur data yang lebih lengkap. Adapun penelitian yang menggunakan keterbatasan data e-commerce untuk menentukan pelanggan potensial sangat jarang dijumpai. Beberapa algoritma untuk memprediksi pembelian berulang di e-commerce sudah banyak digunakan, namun perbandingan performa metode tersebut dalam konteks e-commerce di Indonesia dengan keterbatasan data belum diketahui. Pada penelitian ini, dibandingkan metode Decision Tree, Naive Bayes, dan Random Forest untuk mengklasifikasikan calon pelanggan potensial dengan menggunakan data penjualan merk Maschere dari dua e-commerce, yaitu Tokopedia dan Shopee. Hasil penelitian menunjukkan algoritma Decision Tree mencapai akurasi 90.91%, Naive Bayes memiliki mencapai akurasi 37.50%, dan Random Forest mencapai tingkat akurasi yang terbaik yaitu sebesar 93.94%. Dari hasil tersebut diketahui bahwa metode Random Forest menjadi metode terbaik dalam mengklasifikasikan probabilitas pelanggan untuk pembelian berulang. Di masa mendatang, hasil penelitian ini dapat dikembangkan kembali sebagai sistem penentu keputusan untuk menentukan calon pelanggan potensial.

***Kata Kunci****: klasifikasi pelanggan, decision tree, e-commerce, naive bayes, random forest.*

## INTRODUCTION

E-commerce helps a company increase the reach and number of potential customers it can target for product sales. This is because the transaction process of exchanging goods in e-commerce involves internet and computer networks (Farras et al., 2022). E-commerce groupings are divided based on the parties involved in the transaction. Business to consumer (B2C) e-commerce occurs when transactions occur in retail, where the business consumers are individuals. Business to Business (B2B) e-commerce occurs when both parties in transactions are organizations, and Consumer to Consumer (C2C) occurs when both parties are individuals (Man, 2020).

Maschere is one of the brands that carries out B2C sales through e-commerce in Indonesia, such as Tokopedia and Shopee. In an effort to increase sales, Maschere carries out marketing strategies in the form of discounts on specific dates or promotional events run by e-commerce parties. Apart from discount prices, Maschere is also active in e-commerce marketing by implementing a cost-per-click advertising system. However, this method is considered less efficient. It does not have a long-term impact on sales because many buyers only hunt for discounts or do window shopping: browsing online stores and looking at the products on offer without making a purchase (Ma et al., 2020).

In business practice, companies use sales data to monitor sales and forecast trends. Historical sales data supports analysis to optimize prices, predict demand, manage stock, and plan marketing strategies (Chee et al., 2022; Farras et al., 2022). However, sales data from the Tokopedia e-commerce platform is currently limited. With the implementation of the buyer's data protection policy on Tokopedia and Shopee, some information on sales data is disguised (Pusat Edukasi Penjual Tokopedia, 2023; Pusat Edukasi Penjual Shopee Indonesia, 2023). This can influence the company in determining business strategy, considering that marketing strategy is critical for increasing sales (Djami et al., 2023).

Finding patterns, trends, and useful information from databases, in this case sales data, can be done using the Knowledge Discovery in Database (KDD) process. In KDD, data mining methods are used to identify patterns in data (Kotawadekar, 2022).

Classification is a technique in data mining that is used to divide data into categories, classes, or groups based on specific characteristics (Kotawadekar, 2022). In the process, classification techniques analyze data sets that already have labels to determine rules for classifying new data into existing class labels (Djami et al., 2023).

Decision Tree, Random Forest, and Naive Bayes are three algorithms that can be used in classification techniques. Decision Tree is an algorithm that uses a decision tree structure to classify objects based on a series of questions asked of the object, where each node represents a question and each branch represents the answer to that question (Charbuty & Abdulazeez, 2021; Talekar, 2020). Meanwhile, Random Forest is an algorithm that uses several decision trees to make predictions. Each decision tree is built separately from other decision trees, and the final prediction in this algorithm is made based on combining the predictions from all decision trees (Papakyriakou & Barbounakis, 2022). Naive Bayes stands out as a straightforward yet effective classification algorithm. Naive Bayes classifier assumes that each feature in the data is independent. Naive Bayes can estimate the probability that an object belongs to a particular class by multiplying the probabilities of each feature of the object (Wickramasinghe & Kalutarage, 2021).

These three algorithms were selected to explore their comparative effectiveness in accurately classifying customers' repeat order probabilities, a crucial aspect for targeted marketing strategies in e-commerce, especially in the context of limited customer information availability from e-commerce in Indonesia.

Previous research has shown how machine learning algorithms are used in predicting sales or classifying customer profiles. Naive Bayes has been used to predict the number of sales based on the best-selling to least-selling product categories with an accuracy rate of 54% (Djami et al., 2023). This algorithm is also used to classify the distribution of customer locations with an accuracy level of 92% (Putro et al., 2020). Decision Trees have been used to predict customer purchasing patterns to maintain product stock availability with an accuracy level of 98.86% (Budilaksono et al., 2021). Random Forest has also been used to predict when a customer will make their next purchase with an accuracy rate of up to 89% (M K et al., 2021). However, all of these studies use more complete datasets sourced from sample datasets from Kaggle repository (M K et al., 2021), or internal company records which do not originate from third-party data like e-commerce, and are not used to analyze the possibility of repeat orders by customers (Budilaksono et al., 2021; Djami et al., 2023; Putro et al., 2020).

This research aims to fill in possible differences in results due to differences between the data used in previous research and the current research: internal data that can be adjusted and

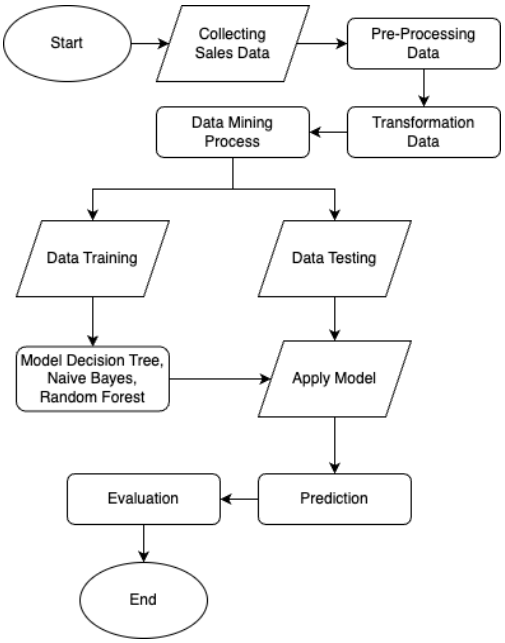
PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

curated by the company and e-commerce sales data that the company cannot change. In this research, a classification of potential customers in e-commerce is proposed based on the probability of repeat customer purchases. The dataset that will be used is Maschere Indonesia sales data on Tokopedia and Shopee as the third-party data providers. The three classification methods that will be used are Naive Bayes, Random Forest, and Decision Tree.

By identifying which buyers have the potential to become loyal customers and make repeat purchases, companies can target specific promotions only to potential loyal customers so that promotional costs can be more efficient in boosting sales. This research also aims to contribute to the broader field by serving as a valuable reference for similar companies that conduct sales through e-commerce platforms like Tokopedia and Shopee. This research provides insights and methodologies that can be adapted to improve sales strategies and promotional efficiency for other companies operating in the digital marketplace. This approach offers a strategic advantage by enabling more targeted marketing efforts and optimizing promotional expenditures, thereby potentially enhancing customer loyalty and sales outcomes in the competitive e-commerce landscape.

## MATERIALS AND METHODS

In this paper, Knowledge Discovery in Databases (KDD) is used in conjunction with Data Mining to discover specific patterns from data sets (Kotawadekar, 2022). The flow of the research methods used can be seen in Figure 1.




Source : (Research Results, 2024)

Figure 1. Research Method Flow Chart

In Figure 1, Sales data collection is the first step in this research, followed by data pre-processing, which involves cleaning and transforming data, data training, and data testing with data mining models such as: Decision Tree, Naive Bayes, and Random Forest, data testing, then finally evaluating classification results (Maryoosh & Hussein, 2022). The Maschere brand sales data used in this research is sales data for the period August to October 2023.

### A. Collecting Sales Data

Sales data for the Maschere brand was obtained by providing limited access to the Tokopedia and Shopee e-commerce seller dashboards. On the Tokopedia and Shopee seller dashboard pages, sales data for the period August to October 2023 can be downloaded in .csv file format.

In that period, a total of 496 records were found, of which 396 records came from sales via Tokopedia and 100 records came from Shopee sales. The downloaded data contains information in the form of Customer Name, Cell Phone Number, Email, Street Name, Province, City, Country, Date Registered as a customer based on date of first purchase, Total Orders, and Total Nominal Spending.

### B. Pre-processing Sales Data

Before it can be used for this research, sales data is then entered into the data pre-processing stage. Sales data from the two e-commerce sites is first combined into one file. Next, sales data needs to go through a data cleansing process first (Fan et al., 2021).

At this stage, unnecessary data, such as empty, incomplete, incorrect, or irrelevant data for analysis purposes, is cleaned (Fan et al., 2021). This is done to reduce noise or interference with research results due to inaccurate data (Arhami & Nasir, 2020). Empty column data without any information such as: Email and Street Name are removed from the dataset.

After the pre-processing stage, the sales data is selected into 10 features. These features are determined by their contribution to the research objective. It was critical to perform machine learning with these attributes to capture customer purchasing behavior dynamics, which may influence the repeat order probability. As displayed in Table 1, “Name” and “Phone Number” represent the customer’s name and phone number, “Province” is the customer’s originating province, “City” is the customer’s originating city, “Country” is the customer’s originating country, “Date Regist.” is the date registration when the customer made first purchased at Maschere through e-commerce and made another purchased between August to

October 2023, "Total Orders Shopee" and "Total Orders Tokopedia" represent the number of times a customer has placed orders on each e-commerce platform, "Total Amount" is the total amount of the customer's purchases across all e-commerce platforms, and "Cust. Category" is the label for customer categorization.

Table 1. Maschere's Sales Data After Pre-Processing

| Name | Phone Number | Province | City | Country | Date Regist. | Total Orders Shopee | Total Orders Tokopedia | Total Amount | Cust. Category |
|---|---|---|---|---|---|---|---|---|---|
| Digda | *********0451 | Jawa Barat | Kota Bekasi | Indonesia | 2023-05-31 16:32:37 | 0 | 2 | 291300 | Regular |
| M******i | ******47 | Banten | Kab. Pandeglang | ID | 2023-06-01 01:39:39 | 9 | 0 | 1919093 | Loyal |
| L******i | *******46 | Jawa Barat | Kota Bekasi | ID | 2023-06-22 19:56:31 | 2 | 0 | 64382 | Non-commited |
| Albertus | *********4987 | D.I. Yogyakarta | Kab. Sleman | Indonesia | 2023-05-31 16:32:26 | 0 | 1 | 139000 | Regular |
| D**n | ******00 | Jambi | Kota Jambi | ID | 2023-06-01 01:40:01 | 20 | 0 | 1191691 | Occasionals |

Source : (Research Results, 2024)

Based on Maschere sales data, Table 2 shows that 8.47% are loyal customers, 14.93% are regular customers, 55.85% are occasional customers, and 20.77% are non-attached customers.

Table 2. Customer Ratio Data by Categories

| Cust. Category | Data Ratio | Number of Customer |
|---|---|---|
| Loyal | 8.47 % | 42 |
| Regular | 14.92 % | 74 |
| Occasional | 55.85 % | 277 |
| Non-Commited | 20.77 % | 103 |

Source : (Research Results, 2024)

### C. Transformation Sales Data

Classification of potential customers will be carried out using RapidMiner Studio Version 10.3 software running on the Linux Debian 12 operating system. The distribution of sales data is 80% for training data and 20% for test data. Customer status in sales data will be converted into 4 class labels: Level 1 for loyal customers, Level 2 for Regular customers, Level 3 for occasional customers, and Level 4 for unattached customers.

### D. Data Mining Methods

In this research, the data mining classification algorithms that will be used are Decision Tree, Naive Bayes, and Random Forest. Although there are many other algorithms in machine learning, limiting the comparison to these three algorithms is also influenced by time and resource constraints, enabling a comprehensive and focused comparison.

Decision Tree is an algorithm that uses a decision tree structure to classify objects based on a series of questions asked of the object, where each node represents a question and each branch represents the answer to that question (Charbuty & Abdulazeez, 2021).

Naive Bayes is a classification algorithm based on the Bayes Theorem, which uses the assumption of independence between features in expressing posterior probability relationships (Papakyriakou & Barbounakis, 2022).

$$P(c|B) = \frac{P(c \cap B)}{p(B)} = \frac{P(c).P(B|c)}{P(B)} \quad \text{..... (1)}$$

where $P(c|B)$ is the probability that data $B$ is included in class $c$, $P(c)$ is the probability that data is included in class, and $P(B|c)$ is the probability that data $B$ has features that correspond to class $c$.

Meanwhile, in the Random Forest algorithm, object classification is carried out using several decision trees, which are built separately from other decision trees to avoid overfitting. The final classification results in the Random Forest model are made based on combining all decision trees and using the Gini Index indicator (Papakyriakou & Barbounakis, 2022).

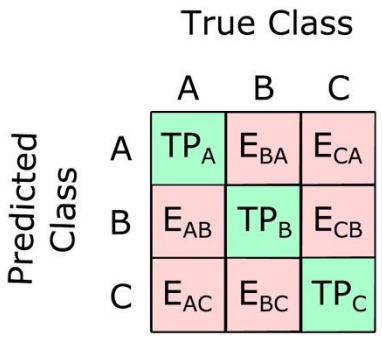
$$G = 1 - \sum_{i=1}^{c} (p(i))^2 \quad \text{..... (2)}$$

where C represent the total number of classes, and $p(i)$ denote the probability of selecting a data point belonging to class $i$.

The classification results obtained through the Decision Tree, Naive Bayes, and Random Forest algorithms will then be evaluated using the confusion matrix table (Tharwat, 2021). In Figure 2,

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 20, No. 1 March 2024 | DOI: 10.33480/pilar.v20i1.5243

$TP_A$ is the amount of class A data that is correctly predicted as class A. $E_{AB}$ is the amount of class A data that is incorrectly classified as class B, and so on (Tharwat, 2021).




Source : (Tharwat, 2021)
Figure 2. Confusion Matrix for a Multi-Class Classification

The model test results can be derived from the information in the confusion matrix table, including accuracy, precision, recall, and F1-Score values. Accuracy is a performance measure that reflects the percentage of data that the model predicts correctly. Precision assesses the ratio of accurate positive predictions relative to all positive predictions, while recall evaluates the proportion of correctly predicted positive data among all actual positive data. Meanwhile, F1-Score is used to measure classification performance by combining precision and recall values (Varoquaux & Colliot, 2023).

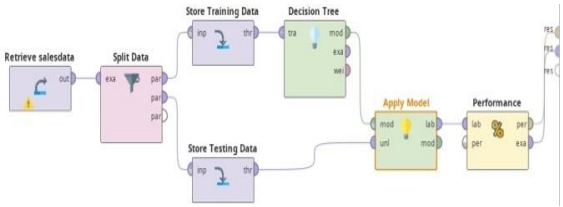
The selection of Accuracy, Precision, Recall, and F1 Score as evaluation indicators in this research is based on their relevance and effectiveness in measuring the performance of classification models. The chosen four indicators represent critical aspects of model performance: overall prediction accuracy (Accuracy), the model's ability to accurately identify positive instances (Precision), the model's ability to find all actual positive cases (Recall), and the harmony between Precision and Recall (F1 Score). These four indicators provide a comprehensive view of the model's quality in the context of this research, allowing researchers and practitioners to understand the model's capability in accurately predicting customer repeat purchases. Although other indicators could add to the understanding of model performance, combining these four indicators is sufficient to provide a robust and representative evaluation of the research objectives.

By carrying out an evaluation, you can find out which algorithm produces the best output so that it can be used to determine potential customers for the Maschere brand.

## RESULTS AND DISCUSSION

In this research, three main experiments will be carried out to determine the classification of potential customers based on repeat order probability using Decision Tree, Naive Bayes, and Random Forest algorithms. Figure 3 displays the operator architecture and functions used in this research. There are several operators used in training and testing data, such as (RapidMiner, 2023):

a. The Retrieve operator is used to import sales data.
b. The Split Data operator splits imported sales data into two data subsets. In this research, the data split is divided into 80:20 for training and test data.
c. Store Data Operators are used to store sales data that has been split into a database.
d. Algorithm models are used to determine algorithms in learning training data for classification and prediction.
e. The Apply Model operator is used to apply the training model to test data to produce classifications or predictions.
f. The Performance Classification operator is used to evaluate model performance results on test data.



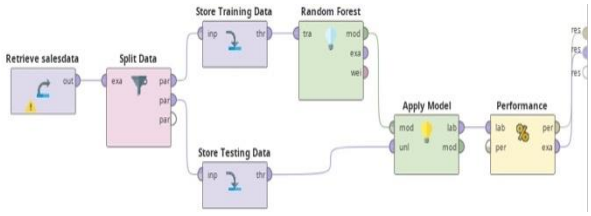

Source : (Research Results, 2024)
Figure 3. Training and Testing Data Process with Decision Tree Algorithm

In Figure 3, the training data is trained using the Decision Tree algorithm. The training is output as a model, which test data can then use. After the training model is applied to the test data, performance evaluation can then be carried out to check the accuracy of the test data's results.

In the first experiment, the Decision Tree model set the depth value to a maximum of 10, split the data 80:20, and did not use pruning or pre-pruning. In terms of performance results, the Decision Tree algorithm shows an accuracy level of 90.91%, a precision of 94.05%, a recall of 88.33%, and an F1 score of 0.91.

The second experiment, seen in Figure 4, uses a Random Forest model with the depth value set to a maximum of 10, split data 80:20, and does not use pruning. The performance results of the Random Forest algorithm show an accuracy level of

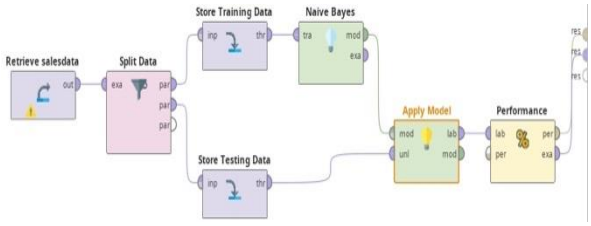
93.94%, precision of 95.14%, recall of 91.90%, and F1 Score of 0.93.




Source : (Research Results, 2024)

Figure 4. Training and Testing Data Process with Random Forest Algorithm

In the third experiment depicted in Figure 5, the Naive Bayes algorithm was applied with an 80:20 data split. The outcomes indicate an accuracy rate of 36.36%, precision reaching 51.70%, recall at 47%, and an F1 Score of 0.49.




Source : (Research Results, 2024)

Figure 5. Training and Testing Data Process with Random Forest Algorithm

Based on these three experiments, classification using Maschere sales data shows that the Random Forest algorithm has the highest level of accuracy compared to other algorithms, namely 93.94%. Then followed by the Decision Tree algorithm, which achieved an accuracy level of 87.88%, and Naive Bayes with the lowest accuracy level of 36.36%. The higher the accuracy level, the higher the correct model prediction value.

Table 3. Data Test Results with Naive Bayes, Decision Tree and Random Forest Algorithms

| Algorithms | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Naive Bayes | 36.36 % | 51.70 % | 47.00 % | 0.49 |
| Decision Tree | 90.91 % | 94.05 % | 88.33 % | 0.91 |
| **Random Forest** | **93.94 %** | **95.14 %** | **91.90 %** | **0.93** |

Source : (Research Results, 2024)

The Random Forest algorithm yielded the highest precision, recall, and F1 scores among the tested models. This shows that the Random Forest algorithm has good performance in predicting classes.

There are many reasons why the Random Forest algorithm can have a better level of accuracy than Decision Tree. Random Forest can produce more accurate models than Decision Trees for various types of data, including complex data. By building several decision trees randomly, Random Forest can also provide more accurate prediction results by combining the prediction results from each decision tree.

Naive Bayes provides results that are much lower in accuracy compared to other algorithms. In the fourth experiment, training and retesting were carried out on the Naive Bayes algorithm with several types of composition of training data and test data. Split data is carried out with a ratio of 70:30, 60:40, 50:50, 40:60, and 30:70 for training data and test data respectively. Table 4 displays the results of the data testing.

Table 4. Naive Bayes Data Test Results with Different Split Data Composition

| Split Data | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 90:10 | 30.61 % | 59.03 % | 43.93 % | 0.50 |
| 70:30 | 26.17 % | 38.95 % | 34.34 % | 0.37 |
| 60:40 | 25.13 % | 40.59 % | 34.21 % | 0.37 |
| **50:50** | **37.50 %** | **47.96 %** | **44.92 %** | **0.46** |
| 40:60 | 27.27 % | 38.45 % | 37.71 % | 0.48 |
| 30:70 | 21.61 % | 39.83 % | 29.23 % | 0.34 |

Source : (Research Results, 2024)

Even though several different data split compositions have been carried out, the Naive Bayes algorithm was only able to achieve the highest accuracy of 37.50% at a 50:50 data split composition. Even when using a 90:10 data split, the accuracy only reached 30.61%. This can be caused by imbalanced data, where the number of comparisons between classes in Maschere sales data is not evenly distributed. Higher precision in the 90:10 split data does not correspond to higher overall accuracy, indicating that Naive Bayes has a known weakness when predicting the majority class in imbalanced datasets. Because the Naive Bayes algorithm assumes independence for each attribute, it tends to predict the majority class and is not suitable for use for all types of data.

In the next stage, testing the Decision Tree and Random Forest training models was carried out by adding pruning settings. In the fifth experiment, Table 5 and Table 6 show the results of testing the Decision Tree and Random Forest algorithms, which are adjusted with additional pruning at Confidence levels ranging from 0.1 to 0.5.

Table 5. Decision Tree Algorithm Data Test Results with Pruning

| Pruning | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.1 | 88.89 % | 90.61 % | 87.42 % | 0.89 |
| **0.2** | **90.91 %** | **94.05 %** | **88.33 %** | **0.91** |
| **0.3** | **90.91 %** | **94.05 %** | **88.33 %** | **0.91** |
| 0.4 | 88.89 % | 90.61 % | 87.42 % | 0.89 |
| 0.5 | 88.89 % | 90.61 % | 87.42 % | 0.89 |

Source : (Research Results, 2024)

Table 6. Random Forest Algorithm Data Test Results with Pruning

| Pruning | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 0.1 | 91.92 | 91.73 | 91.00 | 0.91 |
| **0.2** | **93.94** | **95.14** | **91.90** | **0.93** |
| **0.3** | **93.94** | **95.14** | **91.90** | **0.93** |
| 0.4 | 91.92 | 91.73 | 91.00 | 0.91 |
| 0.5 | 91.92 | 91.73 | 91.00 | 0.91 |

Source : (Research Results, 2024)

Based on the results of the fifth experiment, data testing using the Decision Tree and Random Forest algorithms with pruning did not significantly impact increasing accuracy. The accuracy, recall, precision, and F1-Score values of both Decision Tree and Random Forest algorithms with pruning settings remain the same as for tests without pruning. This condition is caused by Maschere's sales data needing to be more balanced. The model accuracy level also decreased in the Decision Tree algorithm to 88.89% and in the Random Forest to 91.92%, which could be caused by pruning, which removed too many decision tree branches. This causes the model to be too simple and unable to capture data patterns well. In this case, the use of pruning can have a significant impact if used on data with overfitting conditions or more complex data.

## CONCLUSION

In this research, the Random Forest method shows better performance compared to the Decision Tree and Naive Bayes methods in classifying customers based on the probability of repeat orders. The high level of accuracy, namely 93.94%, indicates that the Random Forest method can be used as an effective tool in identifying potential customers. This is particularly important as limited customer information in e-commerce sales data poses an obstacle to determining effective marketing strategies. Although this research was conducted using sales data from the Maschere brand on two e-commerce sites in Indonesia, namely Tokopedia and Shopee, its results can provide general insights useful for other companies selling via e-commerce to optimize their marketing strategies. Sales data stored in a database can be a valuable source of information for predicting sales trends, optimizing prices, managing stock, and planning marketing strategies. This research lays the groundwork for further development in the form of a decision-making system to support companies in better identifying and targeting potential customers in the future. However, the results of this research are contextual and can be influenced by factors that are not included in the analysis model, such as economic factors and changes in market trends. Further research and regular data updates are recommended to ensure the continued success of potential customer classification.

## REFERENCE

Arhami, M., & Nasir, M. (2020). *Data Mining—Algoritma dan Implementasi*. Penerbit Andi.

Budilaksono, S., Jupriyanto, J., Suwarno, M. A. S., Suwartane, I. G. A., Azhari, L., Fauzi, A., … Effendi, M. S. (2021). Customer Profilling for Precision Marketing using RFM Method, K-MEANS algorithm and Decision Tree. *SinkrOn*, *6*(1), 191–200. https://doi.org/10.33395/sinkron.v6i1.11225

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chee, C. C. F., Leng Chiew, K., Sarbini, I. N. B., & Jing, E. K. H. (2022). Data Analytics Approach for Short-term Sales Forecasts Using Limited Information in E-commerce Marketplace. *Acta Informatica Pragensia*, *11*(3), 309–323. https://doi.org/10.18267/j.aip.196

Djami, A. S. M., Utami, N. W., & Paramitha, A. I. I. (2023). The Prediction Of Product Sales Level Using K-Nearest Neighbor and Naive Bayes Algorithms (Case Study: PT Kotamas Bali). Jurnal Pilar Nusa Mandiri, 19(2), 77-84. https://doi.org/10.33480/pilar.v19i2.4420

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*, 652801. https://doi.org/10.3389/fenrg.2021.652801

Farras, M., Friscily, F., Gabriela, G., Sera, S., Sherinne, S., & Mulyawan, B. (2022). *Implementation of Big Data in E-Commerce to Improve User Experience:* Presented at the 3rd Tarumanagara International Conference on the Applications of Social Sciences and Humanities (TICASH 2021), Jakarta, Indonesia. Jakarta, Indonesia. https://doi.org/10.2991/assehr.k.220404.326

Kotawadekar, M. S. V. (2022). Data Mining and Knowledge Discovery Process. *International Journal of Creative Research Thoughts*, *10*(11).

M K, S., Gupta, R., & Gupta, K. (2021). Predicting Customers' Next Order. *International*

*Journal of Engineering Research and Technology*, *10*(6), 418–421. https://doi.org/10.17577/IJERTV10IS060166

Ma, L., Zhang, X., Ding, X., & Wang, G. (2020). How Social Ties Influence Customers' Involvement and Online Purchase Intentions. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(3), 395–408. https://doi.org/10.3390/jtaer16030025

Man, L. (2020). Comparison between Community Shopping Mode and Traditional Cross-border E-commerce Modes. *International Journal of Scientific Engineering and Science*, *4*(3), 8–10. https://doi.org/10.5281/zenodo.3749669

Maryoosh, A. A., & Hussein, E. M. (2022). A Review: Data Mining Techniques and Its Applications. *International Journal of Computer Science and Mobile Applications*, *10*(3), 1–14. https://doi.org/10.47760/ijcsma.2022.v10i03.001

RapidMiner. (2023, October). Operator Manual—RapidMiner Documentation. Retrieved January 16, 2024, from https://docs.rapidminer.com/latest/studio/operators/

Papakyriakou, D., & Barbounakis, I. S. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications*, *183*(48), 5–19. https://doi.org/10.5120/ijca2022921884

Pusat Edukasi Penjual Tokopedia. (2023, 14 Desember). Penerapan Kebijakan Pelindungan Data Pribadi Pembeli. Retrieved January 16, 2024, from Pusat Seller website: https://seller.tokopedia.com/edu/kebijakan-data-pembeli/kebijakan-data-pembeli

Pusat Edukasi Penjual Shopee Indonesia. (2023, March 20). Perubahan Tampilan Informasi Pembeli. Retrieved January 16, 2024, from https://seller.shopee.co.id/edu/article/14910

Putro, H. F., Vulandari, R. T., & Saptomo, W. L. Y. (2020). Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan. *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, *8*(2). https://doi.org/10.30646/tikomsin.v8i2.500

Talekar, B. (2020). A Detailed Review on Decision Tree and Random Forest. *Bioscience Biotechnology Research Communications*, *13*(14), 245–248. https://doi.org/10.21786/bbrc/13.14/57

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Varoquaux, G., & Colliot, O. (2023). Evaluating Machine Learning Models and Their Diagnostic Value. In O. Colliot (Ed.), *Machine Learning for Brain Disorders* (pp. 601–630). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-3195-9_20

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6