Volume 8 Number 4 November 2023
E-ISSN 2622 - 3627
P-ISSN 2502 - 5872

MATHLINE

Jurnal Matematika dan Pendidikan Matematika

Universitas Wiraloda

ISSN 2502-5872 (Print)
ISSN 2622-3627 (Elektronik)

EDITORIAL TEAM

Harina Fitriyani (Scopus ID: 57200642252) Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Eka Fitria Ningsih (Scopus ID: 57209460133) Institut Agama Islam Ma'arif NU (IAIMNU) Metro Lampung, Indonesia
Hasan Hamid (Scopus ID: 57197729527) Universitas Khairun, Ternate, Indonesia
Nofriyandi (Scopus ID: 57211266096) Universitas Islam Riau, Riau, Indonesia
Tommy Tanu Wijaya (Scopus ID: 57218281226) Beijing Normal University, China
Anim (Scopus ID: 57205055902) Universitas Asahan, Kisaran, Indonesia
Salwah (Scopus ID: 57193709774) Universitas Cokrominoto Palopo, Palopo, Indonesia
Eline Yanty Putri Nasution (Scopus ID: 57222535960) Institut Agama Islam Negeri Kerinci, Kerinci, Indonesia
Evan Farhan Wahyu Puadi (Scopus ID: 57216952539) STKIP Muhammadiyah Kuningan, Kuningan, Indonesia
Nurina Hidayah (Scopus ID: 57217607293) Universitas Pekalongan, Pekalongan, Indonesia
Herri Sulaiman (Scopus ID: 57207961751) Universitas Swadaya Gunung Djati, Cirebon, Indonesia
Nurul Mukhlisah Abdal (Scopus ID: 57197827676) Universitas Negeri Makassar, Makassar, Indonesia
Fadilah Ilahi (Scopus ID: 56646356300) UIN Sunan Gunung Djati, Bandung, Indonesia
Lilik Muzdalifah (Scopus ID: 57218170246) Universitas PGRI Ronggolawe Tuban, Indonesia
Ali Shodikin (Scopus ID: 57207958861) Universitas Negeri Surabaya, Surabaya, Indonesia
Denni Ismunandar (Scopus ID: 57221586999) Universitas Wiralodra, Indramayu, Indonesia

Daftar Isi Mathline Volume 8 Number 4, November 2023

ISSN 2502-5872 (Print)
ISSN 2622-3627 (Elektronik)

Volume 8 Nomor 4, November 2023, 1359-1372

CUSTOMER SEGMENTATION ANALYSIS USING RANDOM FOREST & NAÏVE BAYES METHOD IN THE CASE OF MULTI-CLASS CLASSIFICATION AT PT. XYZ

Sofia Debi Puspa[1*], Fani Puspitasari[2], Joko Riyono[1], Christina Eni Pujiastuti[1], David Leon Bijlsma[1], Joseph Andrew Leo[3]

[1]Departement of Mechanical Engineering, Universitas Trisakti, Jakarta Province, Indonesia
[2]Departement of Industrial Engineering, Universitas Trisakti, Jakarta Province, Indonesia
[3]Departement of Computer Science & Business Administration, University of Southern California, United States

*Correspondence: sofia.debi.puspa@trisakti.ac.id

ABSTRACT

Cases of the COVID-19 pandemic are gradually decreasing every day in Indonesia, but the impact of the COVID-19 pandemic has greatly affected various sectors, especially the economy and business. Sales transactions have not yet reached the company's target due to weak public purchasing power. The accuracy of customer segmentation analysis and attractive promo voucher offers are needed to increase the opportunity for people's purchasing power for a product. This study aimed to predict the level of customer purchasing power using the random forest and naïve Bayes methods in the case of multi-class data classification at PT. XYZ. The classification is carried out to determine the type of promo voucher suitable to be offered to customers according to the level of customer purchasing power. The data used is historical daily transaction data from January 1, 2022, to December 31, 2022, which is the transition period for the COVID-19 pandemic. Evaluation using the random forest method produces an accuracy of 99.99%, while the naïve Bayes method produces an accuracy of 92.99%. The random forest and naïve Bayes methods can work very well on large data volumes. However, from the comparison results, it can be concluded that the performance of the random forest method is better than the naïve Bayes method in the multi-class classification case in predicting the level of customer purchasing power at PT. XYZ.
Keywords: Classification, Random Forest, Naïve Bayes, Multi-Class, Customer Segmentation

How to Cite: Puspa, S. D., Puspitasari, F., Riyono, J., Pujiastuti., Bijlsma, D. L., & Leo, J. A. (2023). Customer Segmentation Analysis Using Random Forest & Naïve Bayes Method In The Case of Multi-Class Classification at PT. XYZ. *Mathline: Jurnal Matematika dan Pendidikan Matematika*, *8*(4), 1359-1372. http://doi.org/10.31943/mathline.v8i4.532

PRELIMINARY

In recent years, there has been an exponential positive growth in the volume of data in the big data phenomenon. Apart from increasing volume, the variety and complexity of data is also experiencing rapid development. The impact of the big data phenomenon is very significant in various sectors, especially the business sector. Today's business competition is determined by the ability to process data to achieve optimal user solutions (Riahi & Riahi, 2018). According to (Romero et al., 2021), studying the current situation based on Business Intelligence (BI) in the economic and business fields can positively impact making effective and accurate decisions in companies. This includes acquiring analytical skills, IT capabilities, business knowledge, and communication skills. The goal is to enhance a company's market position with innovative solutions and gain a competitive edge in business.

COVID-19 emerged in Wuhan, China, in December 2019 and has devastated global health. It was declared a pandemic by the WHO on March 11, 2020. Lockdowns and quarantine measures have been implemented worldwide to contain its spread. Capital markets have been affected due to uncertainty around its impact on investments (Parwati et al., 2023). In Indonesia, the COVID-19 virus spread rapidly in 2020, leading to restrictions on community activities. This has caused many companies to reduce output capacity by decreasing working hours and stopping machine use. Some businesses were forced to stop operating due to regulatory factors. This has had a significant impact on multiple sectors and has slowed down the Indonesian economy (Badan Pusat Statistik, 2020).

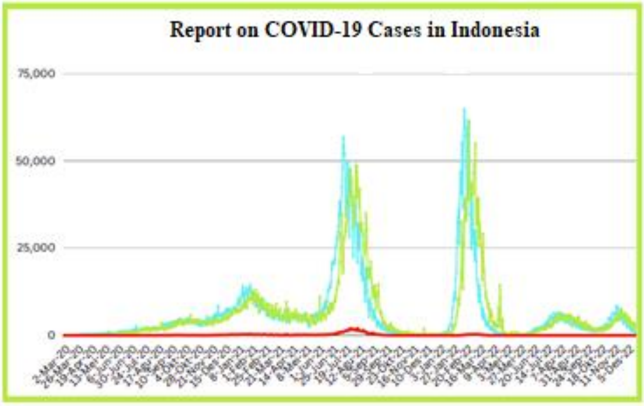

Figure 1. Covid-19 Daily Case Graph

Source: (Komite Penanganan Covid-19 & Pemulihan Ekonomi Nasional, 2023)

In 2021, COVID-19 cases decreased despite a rise in daily new cases in February-March 2022. Daily new cases gradually reduced until December 2022, as shown in Figure 1. This period marked Indonesia's transition from the pandemic, with some business sectors recovering. However, people's purchasing power is still weak, and sales transactions have yet to reach company targets. PT. XYZ is one such company that has started to improve during this transition period.

PT. XYZ operates in the food and beverage (F&B) sector and has 200 outlets throughout Indonesia. PT. XYZ grows along with digitalization, where more than 70% of sales come from online orders. Therefore, to increase people's purchasing power, it is necessary to offer attractive & targeted types of promotional vouchers. Sales transaction data at PT. XYZ has a large data volume with complexity, including menu variations, voucher types, channels, and many stores with varying customer purchasing power. So, it is necessary to use a data mining-based algorithm to determine the right promotional voucher offer by classifying the purchasing power level of PT customers. XYZ. Accurate analysis based on purchasing power and consumer behavior will increase the opportunity to purchase products offered by marketers (Rahim et al., 2021).

Classification is a data mining technique that predicts future trends based on historical data. It falls under the category of predictive mining, which is a type of supervised learning. There are various methods of classification, such as decision trees, the C4.5 algorithm, random forest, naïve Bayes, support vector machine, neural network, and more. Based on research findings by (Schonlau & Zou, 2020), random forest models have higher prediction accuracy than parametric models like linear regression and logistic regression. Multiclass outcomes and regressions yield greater performance improvements than binary outcomes. Additionally, it has a feature selection process that enables the model to work efficiently on complex parameters of big data (Pavlov, 2019).

According to a research study conducted by (Zaw et al., 2019), the Naïve Bayes method was able to accurately detect 81.25% of tumor images and 100% of non-tumor images, resulting in an overall accuracy rate of 94%. The study concluded that Naïve Bayes is a reliable and fast method for detecting brain tissue abnormalities. In addition, another research study conducted by (Putro et al., 2020) found that Naïve Bayes was effective in customer classification, achieving a precision value of 100%, a recall value of 91%, and an accuracy value of 92%. The Naive Bayes method can also be used for sentiment analysis and has good processing time complexity on big data during text classification, where the algorithm is used as a classification engine (Aprilia et al., 2021). These results highlight the effectiveness and efficiency of Naïve Bayes and Random Forest methods in various classification applications.

Based on the explanation above, this research aims to predict the most suitable promotional voucher offer by classifying the level of customer purchasing power using the Random Forest and Naïve Bayes methods. Furthermore, the model evaluation results will compare the performance of the Random Forest and Naïve Bayes methods to determine the best modeling approach. The benefit of this research is providing insights into the most effective method for classifying customer purchasing power levels, helping the marketing team to create customer segmentation and market analyses.

METHODS

Data mining is a big data analysis technique to find patterns and meaningful relationships hidden in data. Data mining generally processes data originating from observations with large data volumes. As a result, data mining is connected to various other scientific fields, including mathematics (particularly optimization), computer science, machine learning, artificial intelligence, image processing, text mining, and others (Durugkar et al., 2022).

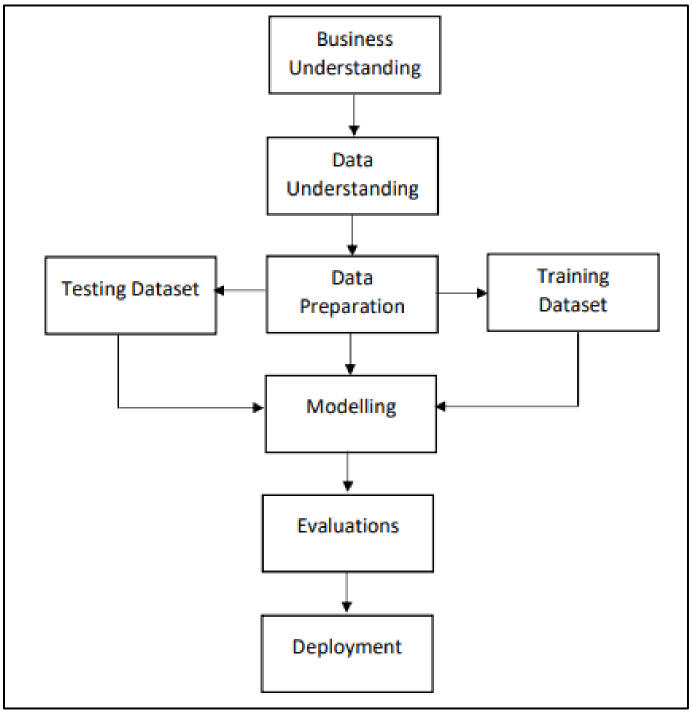

Figure 2. Flowchart Research

Cross-Industry Standard Process for Data Mining (Crisp-DM) is a data mining process model (data mining framework). CRISP-DM explicitly introduces business understanding and data understanding as the primary foundation for digging deeper insights to achieve business goals. The research flowchart is shown in Figure 2, and the model life cycle of the CRISP-DM process in this research is as follows (Schröer et al., 2021).

1. Business Understanding

The following are the stages of business understanding:

a. Determine business goals

The first step of this research is to identify the problem and further investigate the company's business objectives, products, and sales system by conducting interviews with relevant parties.

b. Conducting an assessment

Evaluating the data availability and assessing the situation to determine whether the data held follows the analysis needs.

c. Determining data mining objectives

In this research, the data mining method aims to obtain knowledge in predicting the suitable type of voucher offer based on customers' level of purchasing power from each outlet location.

2. Data Understanding

The process of data understanding begins with collecting initial data and the results of activities as a first step in getting to know the data well. This helps to identify initial insights and interesting subsets of the data, which can be used to form hypotheses about valuable information. The data used in this research is primary data with 416,603 sales transaction data, including holidays. This data is daily sales transaction data from 200 outlets across Indonesia.

3. Data Preparation

The data preparation stage involves all activities in forming the dataset that will be used in modeling. At this stage, attribute selection, table formation, transformation, and data cleaning are carried out to remove missing values and outliers until the data is ready for modeling. The data variables used in this research are outlet location, transaction time, channel, and average ticket (AT), which have been divided into several classes according to the level of purchasing power.

4. Modeling

In this stage, we first visualize the data and then select and implement data mining techniques to solve the problem. For this research, we utilized the Random Forest and Naïve Bayes algorithms. The data for classification was divided into two groups- training data and testing data with the assistance of R Studio Statistical Computing software.

5. Evaluations

Evaluation is used to determine the level of accuracy or error rate in a model that has been created. This helps to assess the model's effectiveness in solving a particular problem. Additionally, evaluation can be used to compare the performance of different algorithms that are used to solve the same problem.

6. Deployment

After evaluating the modeling results, the Decision Support System (DSS). is implemented to predict suitable vouchers based on customer's purchasing power. It is important to monitor the models for accuracy in case of operational changes. For this research, primary data was sourced from PT. XYZ covers the period between January 1, 2022, and December 31, 2022, which is the transition period for the COVID-19 pandemic, where the Indonesian economy is starting to recover from the impact of the pandemic. The classification analysis was conducted using the Random Forest & Naïve Bayes method using R statistical computing software.

Random forest was designed by (Breiman, 2001), which is a supervised learning method developed from the work of (Amit & Geman, 1997; Ho, 1998; Dietterich, 2000). This method significantly improves performance compared to single tree classifiers such as C4.5. Random forest is a powerful tool for prediction modeling because it can handle datasets with a large number of predictor variables. However, it is often beneficial to minimize the number of predictors needed to obtain accurate outcome predictions to improve efficiency. Random forest is a combination of several predictor trees called decision trees, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Random forest prediction results are obtained through the majority of results from each individual decision tree, namely through voting for classification and averaging for regression. A Random Forest consisting of N trees can be formulated in the equation (1) below

$$l(y) = argmax_c\left(\sum_{n=1}^{N} I_{h_n(y)=c}\right) \tag{1}$$

where I is the indicator function and h_n is the n-th tree (Speiser et al., 2019).

The random forest algorithm can be divided into two stages, namely, forming a random forest and then making predictions from the random forest classifier formed in the first stage. The formation of a random forest can be briefly explained through the following pseudocode in Figure 3 (Robin & Jean-Michel, 2020):

1. Randomly select k features from a total of m features, where k << m (k is much smaller than m)
2. Among k features, calculate node d using the best-split point
3. Split nodes into daughter nodes using the best split
4. Repeat steps 1 to 3 until l nodes are reached
5. Create a forest by repeating steps 1 to 4 n times to form n trees

Figure 3. Pseudocode of Random Forest

The Naïve Bayes method is a classic classification technique that utilizes statistical calculations, specifically Bayes' Theorem. The main feature of Naïve Bayes classification is the strong assumption that all parameters are independent. Thomas Bayes, a British statistician, proposed this method, which involves predicting future possibilities based on previous experiences. Bayes' theorem can be written in equation (2)

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{2}$$

where $P(C|X)$ is a posterior, $P(X|C)$ is a likelihood, $P(C)$ is class prior probability, and $P(X)$ is predictor prior probability(Kubat, 2021). Naïve Bayes classification estimates the probability equation in (3) & (4) as follows :

$$P(y) = \frac{n_y}{n} \tag{3}$$

$$P(x_i|y) = \frac{n_y \cap x_i}{n_y} \tag{4}$$

where n is the total number of data points in the training data set; n_y is the number of target data points of class y; $n_y \cap x_i$ is the number of data points with target class y; and i is an attribute variable of x_i (Makruf et al., 2021).

By using the maximum likelihood estimation principle, Naive Bayes classification determines the most probable category for a given sample (Larose & Larose, 2019).

$$P(C_i|X) = Max\{P(C_1|X), P(C_2|X), \ldots, P(C_n|X)\} \tag{5}$$

Suppose the sample $X = (A_1, A_2, \ldots, A_k)$ is an attribute vector, A_j is the jth attribute which may have several different value x_j. In Naïve Bayes classification, it is assumed that the attributes are independent of each other, so:

$$P(X|C_i) = \prod_{j=1}^{k} P(A_j = x_j|C_i) \tag{6}$$

$$P(C_i|X) = \frac{\prod_{j=1}^{k} P(A_j = x_j|C_i) P(C_i)}{P(X)} \tag{7}$$

Let $\frac{1}{P(X)} = \alpha\ (> 0)$, that is $P(C_i|X) = \alpha \prod_{j=1}^{k} P(A_j = x_j|C_i)\, P(C_i)$ (8)

Bayesian decision theory, a fundamental approach to decision-making under the probability framework, makes optimal classification decisions based on probabilities and costs of misclassification when all relevant probabilities are known. The pseudocode for the Naive Bayes classifier algorithm is briefly shown in Figure 4 (Zhou, 2021).

1. Read the training dataset T
2. Calculate the mean and standard deviation of the predictor variables in each class
3. Repeat then calculate the probability of f_i using the gauss density equation in each class until the probability of all predictor variables $(f_1, f_2, f_3, \dots, f_n)$ has been calculated
4. Calculate the likelihood for each class
5. Get the greatest likelihood

Figure 4. Pseudocode of Naïve Bayes

Table 1 shows the confusion matrix for the multi-class classification case with k classes. Furthermore, from the confusion matrix, an evaluation of the algorithm's performance is calculated based on accuracy, precision, recall (sensitivity), and specificity sequentially using the formula shown in equations (9), (10), (11), and (12) where TP refers to truly identified as a positive result, TN refers to truly identified as negative, FP refers to falsely identified as positive result and FN refers to falsely identified as negative result (Markoulidakis et al., 2021).

Table 1. Confusion Matrix Form for Multi-Class Classification

		Actual			
		Class 1	Class 2	⋯	Class k
Prediction	Class 1	f_{11}	f_{12}	⋯	f_{1k}
	Class 2	f_{21}	f_{22}	⋯	f_{2k}
	⋮	⋮	⋮	⋮	⋮
	Class k	f_{k1}	f_{k2}	⋯	f_{kk}

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}} \tag{9}$$

$$\text{Precision of Class } \mathrm{C_i} = \frac{TP(C_i)}{TP(C_i)+FP(C_i)} \tag{10}$$

$$\text{Recall of Class } \mathrm{C_i} = \frac{TP(C_i)}{TP(C_i)+FN(C_i)} \tag{11}$$

$$\text{Spcificity of Class } \mathrm{C_j} = \frac{\sum_{\substack{i=1 \\ i\neq j}}^{N}\sum_{\substack{k=1 \\ k\neq j}}^{N} C_{i,k}}{\sum_{i=1}^{N}\sum_{\substack{k=1 \\ k\neq j}}^{N} C_{i,k}} \tag{12}$$

RESULT AND DISCUSSION

Data Description

At PT. XYZ, the sales transaction app, provides valuable insights into overall sales volume and product performance. However, some limitations need to be addressed:

1. It is necessary to determine the appropriate type of voucher promo based on the level of purchasing power of customers.
2. The type of voucher offered does not consider transaction times during high and low sales volume periods on specific dates.
3. The app does not provide additional information, such as sales predictions and analysis of factors influencing sales. Therefore, supporting data is needed for informed decision-making by the Business Director.

Data preparation is proposed in this research after carrying out business understanding and data understanding. In data preparation, data cleaning, feature selection, data transformation, viewing data dimensions, reviewing the structure of the input dataset, and checking for missing data from certain customers are carried out. In detail, the features of the dataset are shown in Table 2 to provide a more comprehensive understanding.

Table 2. Dataset After Pre-Processing

No	Column	Type	Details
1	Location	Factor *(Class Object)*	Location of outlets spread across Indonesia, such as Bali, Jakarta, West Java, Central Java, South Sumatera, North Sumatera, and others.
2	Month	Factor *(Class Object)*	Month of sales transaction such as January, February, March, etc.
3	Week in year	Factor *(Class Object)*	Period per week where sales transactions occur such as week1, week2, week3, …, and week 52
4	Quartal	Factor *(Class Object)*	Sales transaction quarter period such as Q1, Q2, Q3 and Q4.
5	Channel	Factor *(Class Object)*	Types of transaction services such as Gojek, Grab, Shopeefood, Traveloka, Dine In, Carry Out and others.
6	AT	Integer	Average Ticket is the average price paid per customer in one visit
7	Decision	Factor *(Class Object)*	Class types on customer purchasing power are divided into 6 classes

Figure 5 shows the distribution of location and channel variables through histogram visualization. Five locations have the highest outlets, including Jakarta, Bodetabek (Bogor, Depok, Tangerang, Bekasi), West Java, East Java, and Central Java, which show the highest sales of PT. XYZ in meeting higher raw material inventories. Moreover, for channels, the five highest types of service are Carry Out, Grab, Gojek, Applications, and Shopeefood, which show that the highest service interest is purchasing in stores via Carry Out. However, the influence of e-commerce such as Grab, Gojek, and Shoppefood is enormous in online purchases of PT. XYZ.

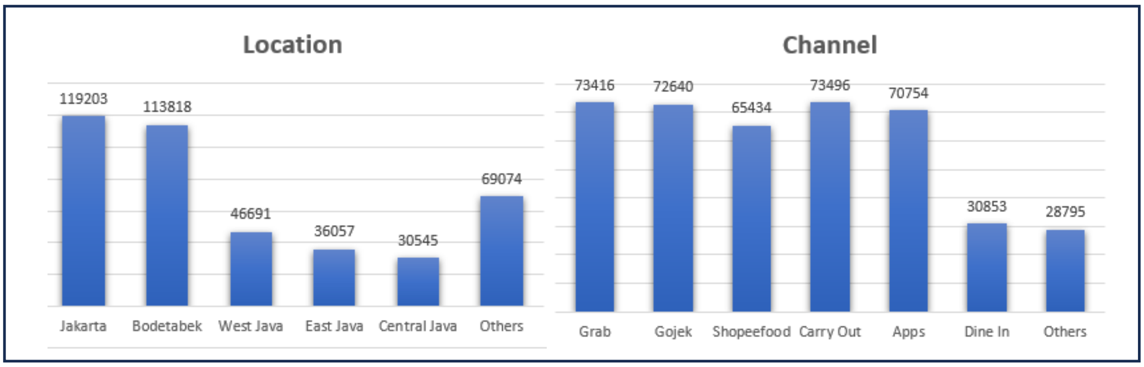

Figure 5. Variable Distribution

Random Forest Classification Analysis Results

The classification of offering the suitable voucher type to customers uses the Random Forest and Naïve Bayes algorithms. This research divides the data into training and testing data with proportions of 75% and 25%, respectively. This division is the best parameter in experiments that have been carried out previously. The analysis results with random forest used a number of trees of 500 to obtain a minimum error of 0.01% (see Figure 6). The more trees used, the smaller the error obtained, but in PT. XYZ transaction data will have the smallest error by using 500 trees.

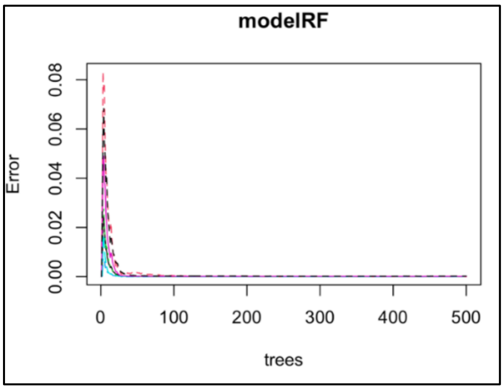

Figure 6. Plot of Error Rate Against Number of Trees

Based on the evaluation of the Random Forest algorithm using a confusion matrix for multi-class classification of test data, it was found that the classification accuracy was 99.99%. Table 3 shows that the random forest model was able to accurately classify 195,698 out of 195,715 data according to their actual class. Specifically, it correctly predicted 4,491 data for class 1, 27,799 data for class 2, 62,186 data for class 3, 62,980 data for class 4, 25,651 data for class 5, and 12,591 data for class 6.

Table 3. Confusion Matrix Random Forest on Testing Data

		Actual					
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Prediction	Class 1	4491	6	0	0	0	0
	Class 2	0	27799	0	0	0	0
	Class 3	0	0	62186	5	0	0
	Class 4	0	0	0	62980	4	0
	Class 5	0	0	0	0	25651	2
	Class 6	0	0	0	0	0	12591

Table 4. Random Forest Algorithm Evaluation

	Precision	Recall	Specificity
Class 1	99,867%	100%	99,99%
Class 2	100%	99,98%	100%
Class 3	99,99%	99,99%	100%
Class 4	99,99%	99,99%	100%
Class 5	99,99%	99,98%	100%
Class 6	100%	99,98%	100%

Table 4 displays the precision, recall, and specificity values used to evaluate the random forest model. The accuracy, precision, and recall values produced from the evaluation showed high values, indicating that the random forest algorithm performs well and effectively classifies data. The algorithm determines suitable promotional vouchers based on customer purchasing power level data, which helps handle large volumes of data.

Naïve Bayes Classification Analysis Results

Applying the Naïve Bayes classification algorithm to test data produces a multi-class confusion matrix, as seen in Table 5. The model achieved an accuracy of 92.99% and successfully classified 182,002 out of 195,715 customer purchasing power level data according to their actual class. The algorithm accurately predicted 4,083 data for class 1, 26,275 data for class 2, 57,613 data for class 3, 59,398 data for class 4, 24,116 data for class 5, and 10,517 data for class 6.

In addition, the naïve Bayes model has demonstrated high precision and recall results, as indicated in Table 6. Hence, based on the accuracy, precision, recall, and specificity results, the Naïve Bayes algorithm proves to be highly effective in classifying large volume customer purchasing power level data.

Table 5. Confusion Matrix Naïve Bayes on Testing Data

		Actual					
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Prediction	Class 1	4083	314	0	0	0	8
	Class 2	309	26275	1739	0	0	273
	Class 3	0	1192	57613	2219	0	174
	Class 4	0	0	2834	59398	1191	365
	Class 5	0	0	0	1368	24116	1256
	Class 6	99	24	0	0	348	10517

Table 6. Naïve Bayes Algorithm Evaluation

	Precision	Recall	Specificity
Class 1	92,69%	90,91%	99,83%
Class 2	91,88%	94,49%	98,62%
Class 3	94,14%	92,65%	97,32%
Class 4	93,12%	94,30%	96,69%
Class 5	90,19%	94%	98,46%
Class 6	95,71%	83,5%	99,74%

Comparison of Random Forest & Naïve Bayes Algorithms

After analyzing the results of both the random forest and naïve Bayes algorithms, it can be concluded that the random forest method performs better than the naïve Bayes method in predicting the level of customer purchasing power in the case of multi-class classification of PT. XYZ data, involving large data volumes. The accuracy, precision, and recall values produced by the random forest method are more significant than the naïve Bayes method (see Table 7). Therefore, determining the prediction of the type of promo voucher based on the level of customer purchasing power is recommended using a random forest model for more accurate multi-class classification.

Table 7. Comparison of Random Forest & Naïve Bayes Evaluation

	Random Forest				Naïve Bayes			
	Acc	Precision	Recall	Specificity	Acc	Precision	Recall	Specificity
Class 1	99,99%	99,87%	100%	99,99%	92,99%	92,69%	90,91%	99,83%
Class 2		100%	99,98%	100%		91,88%	94,49%	98,62%
Class 3		99,99%	99,99%	100%		94,14%	92,65%	97,32%
Class 4		99,99%	99,99%	100%		93,12%	94,30%	96,69%
Class 5		99,99%	99,98%	100%		90,19%	94%	98,46%
Class 6		100%	99,98%	100%		95,71%	83,5%	99,74%

CONCLUSION

This research aims to classify the level of purchasing power of PT customers. XYZ using random forest and naïve Bayes methods in multi-class classification cases. This classification will determine the type of promotional voucher that will be offered to customers according to the level of purchasing power and time. The data used is sales transaction data per day from January 1, 2022, to December 31, 2022, where this period is the transition era of the COVID-19 pandemic. The data consists of 416,603 row data and 7-column data, where the data is divided into training data (75%) and testing data (25%). This division is the best parameter in the experiment. Random forest and naïve Bayes methods methods are effective for large data volumes. Evaluation using the random forest method produces an accuracy of 99.99%, while the performance of the Naïve Bayes algorithm has an accuracy of 92.99%. The random forest method's precision, recall, and specificity values are also higher than those of the naïve Bayes algorithm. Therefore, it can be concluded that the performance of the random forest method is better than the naïve Bayes method in the case of multi-class classification in predicting the level of customer purchasing power at PT. XYZ. This means that in determining the type of promotional voucher based on the customer's purchasing power level, it is recommended that PT. XYZ uses a random forest model for more accurate multi-class classification.

REFERENCES

Aprilia, N. P., Pratiwi, D., & Ariwibowo, A. B. (2021). Sentiment Visualization of Covid-19 Vaccine Based On Naive Bayes Analysis. *Journal of Information Technology and Computer Science*, *6*(2), 195-208. https://doi.org/10.25126/jitecs.202162353

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, *9*(7). https://doi.org/10.1162/neco.1997.9.7.1545

Badan Pusat Statistik. (2020). Analisis Hasil Survei Dampak COVID-19 Jilid 2. *Analisis Hasil Survei Dampak COVID-19 Terhadap Pelaku Usaha*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dietterich, T. G. (2000). Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139-157. https://doi.org/10.1023/A:1007607513941

Durugkar, S. R., Raja, R., Nagwanshi, K. K., & Kumar, S. (2022). Introduction to data mining. In *Data Mining and Machine Learning Applications* (pp. 1–19). wiley. https://doi.org/10.1002/9781119792529.ch1

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832-844. https://doi.org/10.1109/34.709601

Komite Penanganan Covid-19 dan Pemulihan Ekonomi Nasional. (2023, February 3). *Kasus Covid-19 di Indonesia*. Https://Covid19.Go.Id/Id.

Kubat, M. (2021). An Introduction to Machine Learning. In *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-81935-4

Larose, C. D., & Larose, D. T. (2019). Data science using python and R. In *Data Science Using Python and R*. Wiley Blackwell. https://doi.org/10.1002/9781119526865

Makruf, M., Bramantoro, A., Alyamani, H. J., Alesawi, S., & Alturki, R. (2021). Classification methods comparison for customer churn prediction in the telecommunication industry. *International Journal of Advanced and Applied Sciences*, *8*(12), 1-8. https://doi.org/10.21833/ijaas.2021.12.001

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-class confusion matrix reduction method and its application on net promoter score classification problem. *The 14th Pervasive Technologies Related to Assistive Environments Conference*, 412–419. https://dl.acm.org/doi/abs/10.1145/3453892.3461323

Parwati, L. S., Nugrahani, E. H., & Budiarti, R. (2023). Forecasting Stock Price Using Armax-Garchx Model During The Covid-19 Pandemic. *Mathline: Jurnal Matematika Dan Pendidikan Matematika*, *8*(2), 489–502. https://doi.org/10.31943/mathline.v8i2.413

Pavlov, Y. L. (2019). Random forests. In *Random Forests*. De Gruyter Mouton. https://doi.org/10.4324/9781003109396-5

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), 19-24. https://doi.org/10.30646/tikomsin.v8i2.500

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, 102566. https://doi.org/10.1016/j.jretconser.2021.102566

Riahi, Y., & Riahi, S. (2018). Big data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524–528. http://dx.doi.org/10.21276/ijre.2018.5.9.5

Robin, G & Jean-Michel, P. (2020). Random Forests with R. In *Use R*. Springer Cham. http://www.springer.com/series/6991

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. In *Expert Systems with Applications* (Vol. 134, pp. 93–101). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2019.05.028

Romero, C. A. T, Ortiz, J. H., Khalaf, O. I., & Prado, A. R. (2021). Business intelligence: business evolution after industry 4.0. In *Sustainability (Switzerland)*, *13*(18), 10026. https://doi.org/10.3390/su131810026

Zaw, H. T., Maneerat, N., & Win, K. Y. (2019). Brain tumor detection based on Naïve Bayes classification. *Proceeding - 5th International Conference on Engineering, Applied Sciences and Technology, ICEAST 2019*, 1-4, https://doi.org/10.1109/ICEAST.2019.8802562

Zhou, Z. H. (2021). Machine Learning. In *Machine Learning*. Springer Nature. https://doi.org/10.1007/978-981-15-1967-3

CUSTOMER SEGMENTATION ANALYSIS USING RANDOM FOREST & NAÏVE BAYES METHOD IN THE CASE OF MULTI- CLASS CLASSIFICATION AT PT. XYZ

by Sofia Debi Puspa Debi Puspa

M A T H L I N E
JURNAL MATEMATIKA DAN PENDIDIKAN MATEMATIKA

ISSN 2502-5872 (Print)
ISSN 2622-3627 (Elektronik)

Volume 8 Nomor 4, November 2023, 1359-1372

CUSTOMER SEGMENTATION ANALYSIS USING RANDOM FOREST & NAÏVE BAYES METHOD IN THE CASE OF MULTI-CLASS CLASSIFICATION AT PT. XYZ

Sofia Debi Puspa[1*], Fani Puspitasari[2], Joko Riyono[1], Christina Eni Pujiastuti[1], David Leon Bijlsma[1], Joseph Andrew Leo[3]
[1]Departement of Mechanical Engineering, Universitas Trisakti, Jakarta Province, Indonesia
[2]Departement of Industrial Engineering, Universitas Trisakti, Jakarta Province, Indonesia
[3]Departement of Computer Science & Business Administration, University of Southern California, United States

*Correspondence: sofia.debi.puspa@trisakti.ac.id

ABSTRACT

Cases of the COVID-19 pandemic are gradually decreasing every day in Indonesia, but the impact of the COVID-19 pandemic has greatly affected various sectors, especially the economy and business. Sales transactions have not yet reached the company's target due to weak public purchasing power. The accuracy of customer segmentation analysis and attractive promo voucher offers are needed to increase the opportunity for people's purchasing power for a product. This study aimed to predict the level of customer purchasing power using the random forest and naïve Bayes methods in the case of multi-class data classification at PT. XYZ. The classification is carried out to determine the type of promo voucher suitable to be offered to customers according to the level of customer purchasing power. The data used is historical daily transaction data from January 1, 2022, to December 31, 2022, which is the transition period for the COVID-19 pandemic. Evaluation using the random forest method produces an accuracy of 99.99%, while the naïve Bayes method produces an accuracy of 92.99%. The random forest and naïve Bayes methods can work very well on large data volumes. However, from the comparison results, it can be concluded that the performance of the random forest method is better than the naïve Bayes method in the multi-class classification case in predicting the level of customer purchasing power at PT. XYZ.
Keywords: Classification, Random Forest, Naïve Bayes, Multi-Class, Customer Segmentation

How to Cite: Puspa, S. D., Puspitasari, F., Riyono, J., Pujiastuti., Bijlsma, D. L., & Leo, J. A. (2023). Customer Segmentation Analysis Using Random Forest & Naïve Bayes Method In The Case of Multi-Class Classification at PT. XYZ. *Mathline: Jurnal Matematika dan Pendidikan Matematika, 8*(4), 1359-1372. http://doi.org/10.31943/mathline.v8i4.532

PRELIMINARY

In recent years, there has been an exponential positive growth in the volume of data in the big data phenomenon. Apart from increasing volume, the variety and complexity of data is also experiencing rapid development. The impact of the big data phenomenon is very significant in various sectors, especially the business sector. Today's business competition is determined by the ability to process data to achieve optimal user solutions (Riahi & Riahi, 2018). According to (Romero et al., 2021), studying the current situation based on Business Intelligence (BI) in the economic and business fields can positively impact making effective and accurate decisions in companies. This includes acquiring analytical skills, IT capabilities, business knowledge, and communication skills. The goal is to enhance a company's market position with innovative solutions and gain a competitive edge in business.

COVID-19 emerged in Wuhan, China, in December 2019 and has devastated global health. It was declared a pandemic by the WHO on March 11, 2020. Lockdowns and quarantine measures have been implemented worldwide to contain its spread. Capital markets have been affected due to uncertainty around its impact on investments (Parwati et al., 2023). In Indonesia, the COVID-19 virus spread rapidly in 2020, leading to restrictions on community activities. This has caused many companies to reduce output capacity by decreasing working hours and stopping machine use. Some businesses were forced to stop operating due to regulatory factors. This has had a significant impact on multiple sectors and has slowed down the Indonesian economy (Badan Pusat Statistik, 2020).

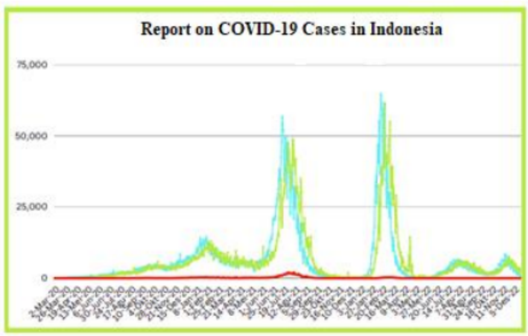

Figure 1. Covid-19 Daily Case Graph

Source: (Komite Penanganan Covid-19 & Pemulihan Ekonomi Nasional, 2023)

In 2021, COVID-19 cases decreased despite a rise in daily new cases in February-March 2022. Daily new cases gradually reduced until December 2022, as shown in Figure 1. This period marked Indonesia's transition from the pandemic, with some business sectors recovering. However, people's purchasing power is still weak, and sales transactions have yet to reach company targets. PT. XYZ is one such company that has started to improve during this transition period.

PT. XYZ operates in the food and beverage (F&B) sector and has 200 outlets throughout Indonesia. PT. XYZ grows along with digitalization, where more than 70% of sales come from online orders. Therefore, to increase people's purchasing power, it is necessary to offer attractive & targeted types of promotional vouchers. Sales transaction data at PT. XYZ has a large data volume with complexity, including menu variations, voucher types, channels, and many stores with varying customer purchasing power. So, it is necessary to use a data mining-based algorithm to determine the right promotional voucher offer by classifying the purchasing power level of PT customers. XYZ. Accurate analysis based on purchasing power and consumer behavior will increase the opportunity to purchase products offered by marketers (Rahim et al., 2021).

Classification is a data mining technique that predicts future trends based on historical data. It falls under the category of predictive mining, which is a type of supervised learning. There are various methods of classification, such as decision trees, the C4.5 algorithm, random forest, naïve Bayes, support vector machine, neural network, and more. Based on research findings by (Schonlau & Zou, 2020), random forest models have higher prediction accuracy than parametric models like linear regression and logistic regression. Multiclass outcomes and regressions yield greater performance improvements than binary outcomes. Additionally, it has a feature selection process that enables the model to work efficiently on complex parameters of big data (Pavlov, 2019).

According to a research study conducted by (Zaw et al., 2019), the Naïve Bayes method was able to accurately detect 81.25% of tumor images and 100% of non-tumor images, resulting in an overall accuracy rate of 94%. The study concluded that Naïve Bayes is a reliable and fast method for detecting brain tissue abnormalities. In addition, another research study conducted by (Putro et al., 2020) found that Naïve Bayes was effective in customer classification, achieving a precision value of 100%, a recall value of 91%, and an accuracy value of 92%. The Naive Bayes method can also be used for sentiment analysis and has good processing time complexity on big data during text classification, where the algorithm is used as a classification engine (Aprilia et al., 2021). These results highlight the effectiveness and efficiency of Naïve Bayes and Random Forest methods in various classification applications.

Based on the explanation above, this research aims to predict the most suitable promotional voucher offer by classifying the level of customer purchasing power using the Random Forest and Naïve Bayes methods. Furthermore, the model evaluation results will compare the performance of the Random Forest and Naïve Bayes methods to determine the best modeling approach. The benefit of this research is providing insights into the most effective method for classifying customer purchasing power levels, helping the marketing team to create customer segmentation and market analyses.

METHODS

Data mining is a big data analysis technique to find patterns and meaningful relationships hidden in data. Data mining generally processes data originating from observations with large data volumes. As a result, data mining is connected to various other scientific fields, including mathematics (particularly optimization), computer science, machine learning, artificial intelligence, image processing, text mining, and others (Durugkar et al., 2022).

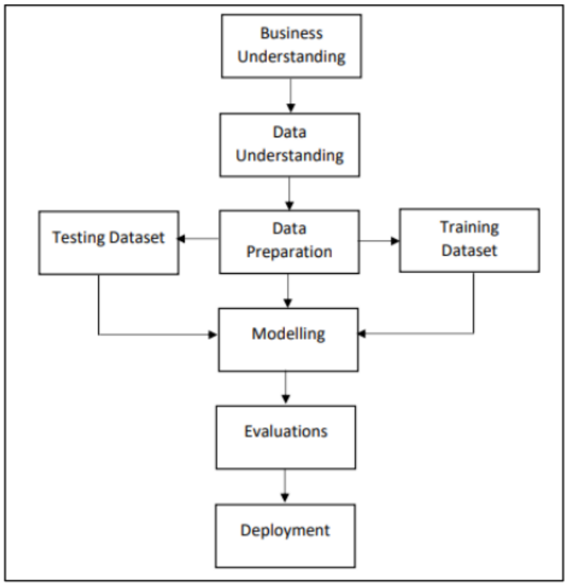

Figure 2. Flowchart Research

Cross-Industry Standard Process for Data Mining (Crisp-DM) is a data mining process model (data mining framework). CRISP-DM explicitly introduces business understanding and data understanding as the primary foundation for digging deeper insights to achieve business goals. The research flowchart is shown in Figure 2, and the model life cycle of the CRISP-DM process in this research is as follows (Schröer et al., 2021).

1. Business Understanding

The following are the stages of business understanding:

a. Determine business goals

The first step of this research is to identify the problem and further investigate the company's business objectives, products, and sales system by conducting interviews with relevant parties.

b. Conducting an assessment

Evaluating the data availability and assessing the situation to determine whether the data held follows the analysis needs.

c. Determining data mining objectives

In this research, the data mining method aims to obtain knowledge in predicting the suitable type of voucher offer based on customers' level of purchasing power from each outlet location.

2. Data Understanding

The process of data understanding begins with collecting initial data and the results of activities as a first step in getting to know the data well. This helps to identify initial insights and interesting subsets of the data, which can be used to form hypotheses about valuable information. The data used in this research is primary data with 416,603 sales transaction data, including holidays. This data is daily sales transaction data from 200 outlets across Indonesia.

3. Data Preparation

The data preparation stage involves all activities in forming the dataset that will be used in modeling. At this stage, attribute selection, table formation, transformation, and data cleaning are carried out to remove missing values and outliers until the data is ready for modeling. The data variables used in this research are outlet location, transaction time, channel, and average ticket (AT), which have been divided into several classes according to the level of purchasing power.

4. Modeling

In this stage, we first visualize the data and then select and implement data mining techniques to solve the problem. For this research, we utilized the Random Forest and Naïve Bayes algorithms. The data for classification was divided into two groups- training data and testing data with the assistance of R Studio Statistical Computing software.

5. Evaluations

Evaluation is used to determine the level of accuracy or error rate in a model that has been created. This helps to assess the model's effectiveness in solving a particular problem. Additionally, evaluation can be used to compare the performance of different algorithms that are used to solve the same problem.

6. Deployment

After evaluating the modeling results, the Decision Support System (DSS). is implemented to predict suitable vouchers based on customer's purchasing power. It is important to monitor the models for accuracy in case of operational changes. For this research, primary data was sourced from PT. XYZ covers the period between January 1, 2022, and December 31, 2022, which is the transition period for the COVID-19 pandemic, where the Indonesian economy is starting to recover from the impact of the pandemic. The classification analysis was conducted using the Random Forest & Naïve Bayes method using R statistical computing software.

Random forest was designed by (Breiman, 2001), which is a supervised learning method developed from the work of (Amit & Geman, 1997; Ho, 1998; Dietterich, 2000). This method significantly improves performance compared to single tree classifiers such as C4.5. Random forest is a powerful tool for prediction modeling because it can handle datasets with a large number of predictor variables. However, it is often beneficial to minimize the number of predictors needed to obtain accurate outcome predictions to improve efficiency. Random forest is a combination of several predictor trees called decision trees, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Random forest prediction results are obtained through the majority of results from each individual decision tree, namely through voting for classification and averaging for regression. A Random Forest consisting of N trees can be formulated in the equation (1) below

$$l(y) = argmax_c\left(\sum_{n=1}^{N} I_{h_n(y)=c}\right) \quad (1)$$

where I is the indicator function and h_n is the n-th tree (Speiser et al., 2019).

The random forest algorithm can be divided into two stages, namely, forming a random forest and then making predictions from the random forest classifier formed in the first stage. The formation of a random forest can be briefly explained through the following pseudocode in Figure 3 (Robin & Jean-Michel, 2020):

1. Randomly select k features from a total of m features, where $k << m$ (k is much smaller than m)
2. Among k features, calculate node d using the best-split point
3. Split nodes into daughter nodes using the best split
4. Repeat steps 1 to 3 until l nodes are reached
5. Create a forest by repeating steps 1 to 4 n times to form n trees

Figure 3. Pseudocode of Random Forest

The Naïve Bayes method is a classic classification technique that utilizes statistical calculations, specifically Bayes' Theorem. The main feature of Naïve Bayes classification is the strong assumption that all parameters are independent. Thomas Bayes, a British statistician, proposed this method, which involves predicting future possibilities based on previous experiences. Bayes' theorem can be written in equation (2)

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (2)$$

where $P(C|X)$ is a posterior, $P(X|C)$ is a likelihood, $P(C)$ is class prior probability, and $P(X)$ is predictor prior probability(Kubat, 2021). Naïve Bayes classification estimates the probability equation in (3) & (4) as follows :

$$P(y) = \frac{n_y}{n} \quad (3)$$

$$P(x_i|y) = \frac{n_y \cap x_i}{n_y} \quad (4)$$

where n is the total number of data points in the training data set; n_y is the number of target data points of class y; $n_y \cap x_i$ is the number of data points with target class y; and i is an attribute variable of x_i (Makruf et al., 2021).

By using the maximum likelihood estimation principle, Naive Bayes classification determines the most probable category for a given sample (Larose & Larose, 2019).

$$P(C_i|X) = Max\{P(C_1|X), P(C_2|X), \dots, P(C_n|X)\} \quad (5)$$

Suppose the sample $X = (A_1, A_2, \dots, A_k)$ is an attribute vector, A_j is the jth attribute which may have several different value x_j. In Naïve Bayes classification, it is assumed that the attributes are independent of each other, so:

$$P(X|C_i) = \prod_{j=1}^{k} P(A_j = x_j|C_i) \quad (6)$$

$$P(C_i|X) = \frac{\prod_{j=1}^{k} P(A_j = x_j|C_i) P(C_i)}{P(X)} \quad (7)$$

Let $\frac{1}{P(X)} = \alpha\ (> 0)$, that is $P(C_i|X) = \alpha \prod_{j=1}^{k} P(A_j = x_j|C_i)\, P(C_i)$ (8)

Bayesian decision theory, a fundamental approach to decision-making under the probability framework, makes optimal classification decisions based on probabilities and costs of misclassification when all relevant probabilities are known. The pseudocode for the Naive Bayes classifier algorithm is briefly shown in Figure 4 (Zhou, 2021).

1. Read the training dataset T
2. Calculate the mean and standard deviation of the predictor variables in each class
3. Repeat then calculate the probability of f_i using the gauss density equation in each class until the probability of all predictor variables $(f_1, f_2, f_3, \dots, f_n)$ has been calculated
4. Calculate the likelihood for each class
5. Get the greatest likelihood

Figure 4. Pseudocode of Naïve Bayes

Table 1 shows the confusion matrix for the multi-class classification case with k classes. Furthermore, from the confusion matrix, an evaluation of the algorithm's performance is calculated based on accuracy, precision, recall (sensitivity), and specificity sequentially using the formula shown in equations (9), (10), (11), and (12) where TP refers to truly identified as a positive result, TN refers to truly identified as negative, FP refers to falsely identified as positive result and FN refers to falsely identified as negative result (Markoulidakis et al., 2021).

Table 1. Confusion Matrix Form for Multi-Class Classification

		Actual			
		Class 1	Class 2	$\cdots$	Class k
Prediction	Class 1	f_{11}	f_{12}	$\cdots$	f_{1k}
	Class 2	f_{21}	f_{22}	$\cdots$	f_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Class k	f_{k1}	f_{k2}	$\cdots$	f_{kk}

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} TP(C_i)}{\sum_{i=1}^{N}\sum_{j=1}^{N} C_{i,j}} \tag{9}$$

$$\text{Precision of Class } \mathrm{C_i} = \frac{TP(C_i)}{TP(C_i)+FP(C_i)} \tag{10}$$

$$\text{Recall of Class } \mathrm{C_i} = \frac{TP(C_i)}{TP(C_i)+FN(C_i)} \tag{11}$$

$$\text{Spcificity of Class } \mathrm{C_j} = \frac{\sum_{\substack{i=1 \\ i\neq j}}^{N} \sum_{\substack{k=1 \\ k\neq j}}^{N} C_{i,k}}{\sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq j}}^{N} C_{i,k}} \tag{12}$$

RESULT AND DISCUSSION

Data Description

At PT. XYZ, the sales transaction app, provides valuable insights into overall sales volume and product performance. However, some limitations need to be addressed:

1. It is necessary to determine the appropriate type of voucher promo based on the level of purchasing power of customers.
2. The type of voucher offered does not consider transaction times during high and low sales volume periods on specific dates.
3. The app does not provide additional information, such as sales predictions and analysis of factors influencing sales. Therefore, supporting data is needed for informed decision-making by the Business Director.

Data preparation is proposed in this research after carrying out business understanding and data understanding. In data preparation, data cleaning, feature selection, data transformation, viewing data dimensions, reviewing the structure of the input dataset, and checking for missing data from certain customers are carried out. In detail, the features of the dataset are shown in Table 2 to provide a more comprehensive understanding.

Table 2. Dataset After Pre-Processing

No	Column	Type	Details
1	Location	Factor *(Class Object)*	Location of outlets spread across Indonesia, such as Bali, Jakarta, West Java, Central Java, South Sumatera, North Sumatera, and others.
2	Month	Factor *(Class Object)*	Month of sales transaction such as January, February, March, etc.
3	Week in year	Factor *(Class Object)*	Period per week where sales transactions occur such as week1, week2, week3, …, and week 52
4	Quartal	Factor *(Class Object)*	Sales transaction quarter period such as Q1, Q2, Q3 and Q4.
5	Channel	Factor *(Class Object)*	Types of transaction services such as Gojek, Grab, Shopeefood, Traveloka, Dine In, Carry Out and others.
6	AT	Integer	Average Ticket is the average price paid per customer in one visit
7	Decision	Factor *(Class Object)*	Class types on customer purchasing power are divided into 6 classes

Figure 5 shows the distribution of location and channel variables through histogram visualization. Five locations have the highest outlets, including Jakarta, Bodetabek (Bogor, Depok, Tangerang, Bekasi), West Java, East Java, and Central Java, which show the highest sales of PT. XYZ in meeting higher raw material inventories. Moreover, for channels, the five highest types of service are Carry Out, Grab, Gojek, Applications, and Shopeefood, which show that the highest service interest is purchasing in stores via Carry Out. However, the influence of e-commerce such as Grab, Gojek, and Shoppefood is enormous in online purchases of PT. XYZ.

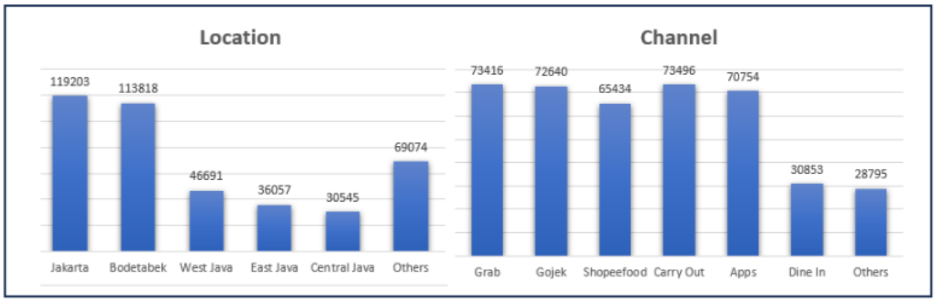

Figure 5. Variable Distribution

Random Forest Classification Analysis Results

The classification of offering the suitable voucher type to customers uses the Random Forest and Naïve Bayes algorithms. This research divides the data into training and testing data with proportions of 75% and 25%, respectively. This division is the best parameter in experiments that have been carried out previously. The analysis results with random forest used a number of trees of 500 to obtain a minimum error of 0.01% (see Figure 6). The more trees used, the smaller the error obtained, but in PT. XYZ transaction data will have the smallest error by using 500 trees.

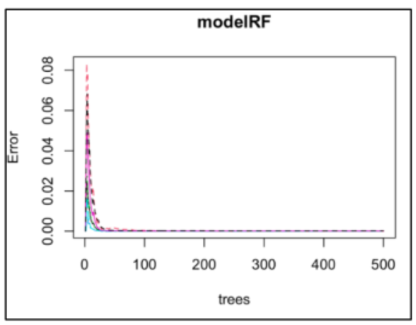

Figure 6. Plot of Error Rate Against Number of Trees

Based on the evaluation of the Random Forest algorithm using a confusion matrix for multi-class classification of test data, it was found that the classification accuracy was 99.99%. Table 3 shows that the random forest model was able to accurately classify 195,698 out of 195,715 data according to their actual class. Specifically, it correctly predicted 4,491 data for class 1, 27,799 data for class 2, 62,186 data for class 3, 62,980 data for class 4, 25,651 data for class 5, and 12,591 data for class 6.

Table 3. Confusion Matrix Random Forest on Testing Data

		Actual					
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Prediction	Class 1	4491	6	0	0	0	0
	Class 2	0	27799	0	0	0	0
	Class 3	0	0	62186	5	0	0
	Class 4	0	0	0	62980	4	0
	Class 5	0	0	0	0	25651	2
	Class 6	0	0	0	0	0	12591

Table 4. Random Forest Algorithm Evaluation

	Precision	Recall	Specificity
Class 1	99,867%	100%	99,99%
Class 2	100%	99,98%	100%
Class 3	99,99%	99,99%	100%
Class 4	99,99%	99,99%	100%
Class 5	99,99%	99,98%	100%
Class 6	100%	99,98%	100%

Table 4 displays the precision, recall, and specificity values used to evaluate the random forest model. The accuracy, precision, and recall values produced from the evaluation showed high values, indicating that the random forest algorithm performs well and effectively classifies data. The algorithm determines suitable promotional vouchers based on customer purchasing power level data, which helps handle large volumes of data.

Naïve Bayes Classification Analysis Results

Applying the Naïve Bayes classification algorithm to test data produces a multi-class confusion matrix, as seen in Table 5. The model achieved an accuracy of 92.99% and successfully classified 182,002 out of 195,715 customer purchasing power level data according to their actual class. The algorithm accurately predicted 4,083 data for class 1, 26,275 data for class 2, 57,613 data for class 3, 59,398 data for class 4, 24,116 data for class 5, and 10,517 data for class 6.

In addition, the naïve Bayes model has demonstrated high precision and recall results, as indicated in Table 6. Hence, based on the accuracy, precision, recall, and specificity results, the Naïve Bayes algorithm proves to be highly effective in classifying large volume customer purchasing power level data.

Table 5. Confusion Matrix Naïve Bayes on Testing Data

		Actual					
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Prediction	Class 1	4083	314	0	0	0	8
	Class 2	309	26275	1739	0	0	273
	Class 3	0	1192	57613	2219	0	174
	Class 4	0	0	2834	59398	1191	365
	Class 5	0	0	0	1368	24116	1256
	Class 6	99	24	0	0	348	10517

Table 6. Naïve Bayes Algorithm Evaluation

	Precision	Recall	Specificity
Class 1	92,69%	90,91%	99,83%
Class 2	91,88%	94,49%	98,62%
Class 3	94,14%	92,65%	97,32%
Class 4	93,12%	94,30%	96,69%
Class 5	90,19%	94%	98,46%
Class 6	95,71%	83,5%	99,74%

Comparison of Random Forest & Naïve Bayes Algorithms

After analyzing the results of both the random forest and naïve Bayes algorithms, it can be concluded that the random forest method performs better than the naïve Bayes method in predicting the level of customer purchasing power in the case of multi-class classification of PT. XYZ data, involving large data volumes. The accuracy, precision, and recall values produced by the random forest method are more significant than the naïve Bayes method (see Table 7). Therefore, determining the prediction of the type of promo voucher based on the level of customer purchasing power is recommended using a random forest model for more accurate multi-class classification.

Table 7. Comparison of Random Forest & Naïve Bayes Evaluation

	Random Forest				Naïve Bayes			
	Acc	Precision	Recall	Specificity	Acc	Precision	Recall	Specificity
Class 1	99,99%	99,87%	100%	99,99%	92,99%	92,69%	90,91%	99,83%
Class 2		100%	99,98%	100%		91,88%	94,49%	98,62%
Class 3		99,99%	99,99%	100%		94,14%	92,65%	97,32%
Class 4		99,99%	99,99%	100%		93,12%	94,30%	96,69%
Class 5		99,99%	99,98%	100%		90,19%	94%	98,46%
Class 6		100%	99,98%	100%		95,71%	83,5%	99,74%

CONCLUSION

This research aims to classify the level of purchasing power of PT customers. XYZ using random forest and naïve Bayes methods in multi-class classification cases. This classification will determine the type of promotional voucher that will be offered to customers according to the level of purchasing power and time. The data used is sales transaction data per day from January 1, 2022, to December 31, 2022, where this period is the transition era of the COVID-19 pandemic. The data consists of 416,603 row data and 7-column data, where the data is divided into training data (75%) and testing data (25%). This division is the best parameter in the experiment. Random forest and naïve Bayes methods methods are effective for large data volumes. Evaluation using the random forest method produces an accuracy of 99.99%, while the performance of the Naïve Bayes algorithm has an accuracy of 92.99%. The random forest method's precision, recall, and specificity values are also higher than those of the naïve Bayes algorithm. Therefore, it can be concluded that the performance of the random forest method is better than the naïve Bayes method in the case of multi-class classification in predicting the level of customer purchasing power at PT. XYZ. This means that in determining the type of promotional voucher based on the customer's purchasing power level, it is recommended that PT. XYZ uses a random forest model for more accurate multi-class classification.

REFERENCES

Aprilia, N. P., Pratiwi, D., & Ariwibowo, A. B. (2021). Sentiment Visualization of Covid-19 Vaccine Based On Naive Bayes Analysis. *Journal of Information Technology and Computer Science*, *6*(2), 195-208. https://doi.org/10.25126/jitecs.202162353

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, *9*(7). https://doi.org/10.1162/neco.1997.9.7.1545

Badan Pusat Statistik. (2020). Analisis Hasil Survei Dampak COVID-19 Jilid 2. *Analisis Hasil Survei Dampak COVID-19 Terhadap Pelaku Usaha*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dietterich, T. G. (2000). Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Machine Learning*, *40*(2), 139-157. https://doi.org/10.1023/A:1007607513941

Durugkar, S. R., Raja, R., Nagwanshi, K. K., & Kumar, S. (2022). Introduction to data mining. In *Data Mining and Machine Learning Applications* (pp. 1–19). wiley. https://doi.org/10.1002/9781119792529.ch1

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832-844. https://doi.org/10.1109/34.709601

Komite Penanganan Covid-19 dan Pemulihan Ekonomi Nasional. (2023, February 3). *Kasus Covid-19 di Indonesia*. Https://Covid19.Go.Id/Id.

Kubat, M. (2021). An Introduction to Machine Learning. In *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-81935-4

Larose, C. D., & Larose, D. T. (2019). Data science using python and R. In *Data Science Using Python and R*. Wiley Blackwell. https://doi.org/10.1002/9781119526865

Makruf, M., Bramantoro, A., Alyamani, H. J., Alesawi, S., & Alturki, R. (2021). Classification methods comparison for customer churn prediction in the telecommunication industry. *International Journal of Advanced and Applied Sciences*, *8*(12), 1-8. https://doi.org/10.21833/ijaas.2021.12.001

Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021). Multi-class confusion matrix reduction method and its application on net promoter score classification problem. *The 14th Pervasive Technologies Related to Assistive Environments Conference*, 412–419. https://dl.acm.org/doi/abs/10.1145/3453892.3461323

Parwati, L. S., Nugrahani, E. H., & Budiarti, R. (2023). Forecasting Stock Price Using Armax-Garchx Model During The Covid-19 Pandemic. *Mathline: Jurnal Matematika Dan Pendidikan Matematika*, *8*(2), 489–502. https://doi.org/10.31943/mathline.v8i2.413

Pavlov, Y. L. (2019). Random forests. In *Random Forests*. De Gruyter Mouton. https://doi.org/10.4324/9781003109396-5

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), 19-24. https://doi.org/10.30646/tikomsin.v8i2.500

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, 102566. https://doi.org/10.1016/j.jretconser.2021.102566

Riahi, Y., & Riahi, S. (2018). Big data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524–528. http://dx.doi.org/10.21276/ijre.2018.5.9.5

Robin, G & Jean-Michel, P. (2020). Random Forests with R. In *Use R*. Springer Cham. http://www.springer.com/series/6991

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. In *Expert Systems with Applications* (Vol. 134, pp. 93–101). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2019.05.028

Romero, C. A. T, Ortiz, J. H., Khalaf, O. I., & Prado, A. R. (2021). Business intelligence: business evolution after industry 4.0. In *Sustainability (Switzerland)*, *13*(18), 10026. https://doi.org/10.3390/su131810026

Zaw, H. T., Maneerat, N., & Win, K. Y. (2019). Brain tumor detection based on Naïve Bayes classification. *Proceeding - 5th International Conference on Engineering, Applied Sciences and Technology, ICEAST 2019*, 1-4, https://doi.org/10.1109/ICEAST.2019.8802562

Zhou, Z. H. (2021). Machine Learning. In *Machine Learning*. Springer Nature. https://doi.org/10.1007/978-981-15-1967-3

CUSTOMER SEGMENTATION ANALYSIS USING RANDOM FOREST & NAÏVE BAYES METHOD IN THE CASE OF MULTI-CLASS CLASSIFICATION AT PT. XYZ

ORIGINALITY REPORT

11%	9%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ojs.istp-press.com Internet Source	1%
2	Submitted to Universiti Teknologi MARA Student Paper	1%
3	etda.libraries.psu.edu Internet Source	1%
4	www.sciencepubco.com Internet Source	1%
5	Khalid Shaikh, Sabitha Krishnan, Rohit Thanki. "Chapter 3 Artificial Intelligence and Learning Algorithms", Springer Science and Business Media LLC, 2021 Publication	1%
6	"International Conference on Information Systems and Intelligent Applications", Springer Science and Business Media LLC, 2023 Publication	1%

7	files.library.northwestern.edu Internet Source	1 %
8	Karisma Trinanda Putra, Heri Wijayanto, Shazana Dhiya, Heri Kurniawan Asshiddiq, Zeyad Albatati, Rania Riyadhini Khairuddin. "Electrical Load Forecasting on A 500/150 kV Substation Based on Historical Daily Records Using Multivariate LSTM Network", 2022 2nd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), 2022 Publication	1 %
9	Peddle, D.R.. "Large area forest classification and biophysical parameter estimation using the 5-Scale canopy reflectance model in Multiple-Forward-Mode", Remote Sensing of Environment, 20040130 Publication	1 %
10	ebin.pub Internet Source	<1 %
11	Submitted to The University of Wolverhampton Student Paper	<1 %
12	real-j.mtak.hu Internet Source	<1 %
13	Jaime Lynn Speiser, Michael E. Miller, Janet Tooze, Edward Ip. "A comparison of random	<1 %

forest variable selection methods for classification prediction modeling", Expert Systems with Applications, 2019
Publication

14	pure.ulster.ac.uk Internet Source	<1 %
15	Submitted to Academic Library Consortium Student Paper	<1 %
16	Mauricio A. Valle, Gonzalo A. Ruz. "Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms", Applied Artificial Intelligence, 2015 Publication	<1 %
17	ejournals.umn.ac.id Internet Source	<1 %
18	www.idouba.net Internet Source	<1 %
19	www.tandfonline.com Internet Source	<1 %
20	Aditya Sai Srinivas T., Ramasubbareddy Somula, Govinda K., Akriti Saxena, Pramod Reddy A.. "Estimating rainfall using machine learning strategies based on weather radar data", International Journal of Communication Systems, 2019 Publication	<1 %

21	enoumen.com Internet Source		<1 %
22	idss.iocspublisher.org Internet Source		<1 %
23	www.zhende.com Internet Source		<1 %

Exclude quotes On
Exclude bibliography On

Exclude matches < 10 words

CUSTOMER SEGMENTATION ANALYSIS USING RANDOM FOREST & NAÏVE BAYES METHOD IN THE CASE OF MULTI-CLASS CLASSIFICATION AT PT. XYZ

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14