International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 8, No. 3, (2024), pp. 177-185

Customer Loyalty Classification with Comparison of Naive Bayes, C4.5, and KNN Methods

Embun Fajar Wati[1], Elvi Sunita Perangin-Angin[2], Luthfi Indriyani[3]
[1,2,3]Universitas Bina Sarana Informatika, Indonesia
Email: embun.efw@bsi.ac.id[1], elvi.evt@bsi.ac.id[2], luthfi.lfy@bsi.ac.id[3]

Abstract

Customer loyalty is indispensable for the survival of a company. Customer loyalty needs to be maintained in order to return to visit and transact with the Company. Customer data consisting of age, annual income, purchase amount, region, purchase frequency, and loyalty score features can produce new information, namely analyzing customers who have high loyalty. Data processing is carried out using three data mining algorithms, namely Naïve Bayes, C4.5 or Decision Tree, and KNN. The stages carried out in data processing consist of data selection, preprocessing, transformation, and modelling. The customer data used amounted to 238. Modelling is carried out using Rapid Miner Software. Customer loyalty classification can be easily done with the three algorithms, namely Naive Bayes, and C4.5 or Decision Tree, and KNN which is validated by the 10-fold cross-validation method so as to produce the highest percentage of accuracy and the similarity of the accuracy value of the Naive Bayes and C4.5 algorithms, which is 96.67%. In the AUC value, it can be seen that the Naive Bayes algorithm is superior to the C4.5 algorithm or Decision Tree and KNN. The result of the highest AUC value is 0.997, the highest precision percentage is 98.92% achieved by the Naive Bayes algorithm. The result of the highest recall percentage is C4.5 of 100%. The results of the AUC value and accuracy percentage on the three algorithms prove that the performance of the three algorithms is very good.

Keywords: *customer loyalty, naïve bayes, c4.5, knn, data mining*

1. Introduction

The rapid development of technology, information systems, and science has resulted in increasingly fierce competition in the business world. In the business world, customers are the main asset [1]. Companies must strive to survive, one way is to maintain customer loyalty. Data mining techniques can be used to predict customer loyalty [2]. Loyalty is a firmly held commitment to repurchase or subscribe to products and services in the future despite situational influences [3].

Consumption patterns or loyalty have several factors that affect it. To find out consumption patterns or loyalty, the variables used are income, shopping habits such as shopping frequency and the amount of shopping in a certain period of time [4]. Factors of consumer shopping habits can affect purchasing decisions or customer loyalty [5]. This information allows for a comprehensive analysis of how these factors affect consumer shopping decisions which can affect customer loyalty [6]. The study also analyzes the influence of factors such as age, revenue level, and transaction frequency on customer loyalty through feature interest analysis. The proposed machine learning approach offers valuable insights for companies to identify and target loyal customers, thereby enabling effective customer relationship management and improved business performance [7].

This research begins the search to estimate the likelihood of returning customers to transact through machine learning, i.e. data mining [8]. Data mining is a method to find knowledge in a large enough pile of data by the process of digging and analyzing a very large amount of data in obtaining something true, new and useful so that a pattern or pattern can be found in the data [9]. Data mining can be used to predict customer loyalty.

ISSN : 2580-7250

Customer loyalty predictions are useful for identifying customers who will switch to one of their customers' competitors [10]. A regular customer will buy from the same company, thus generating revenue for several years, and providing the company with a relatively stable position in the market [11].

Because loyal customers make more profits, this study aims to classify customers and their loyalty [12]. The process and results of this research are targeted to build various machine learning models to predict customer loyalty [13]. The purpose of this clustering is to get a segmentation of users who have different Customer Lifetime Values. Another purpose of this classification is to predict new users into the user segmentation obtained as a result of its grouping [14]. By using the data mining method, the results obtained from data processing are able to provide important information for companies to group customers that must be prioritized [15].

The methods used in customer loyalty classification are naïve bayes, C4.5, and KNN. The naïve Bayes method has advantages, including being able to predict based on concrete data, so that the results obtained can be accounted for and used for future predictions [16]. This C4.5 algorithm is able to handle categorical and numerical data, and has the ability to select relevant attributes for decision-making [17]. KNN (K-Nearest Neighbor) is a powerful data classification technique that involves searching for similar cases by calculating the proximity between new and old cases using matching weights [18].

The features used in this study are age, annual income, purchase amount, region, purchase frequency, and loyalty score. These features will be processed by the KNN (K-Nearest Neighbor) algorithm, Naive Bayes, C4.5 [19]. The performance comparison of the three algorithms aims to measure the accuracy and execution time of each algorithm to get the best algorithm to be applied [20].

2. Research Methodology

In this study, 3 data mining algorithms will be used, namely KNN (K-Nearest Neighbor), Naive Bayes, and C4.5 or commonly known as Decision Tree. The three algorithms will be evaluated by comparing the results with the confusion matrix values, namely accuracy, precision, recall, and AUC values.

a) Naive Bayes is the most excellent search for probability values and is one of the algorithms of the classification type. This method was chosen because it is easy to apply to work independently, that is, a feature in a data is not related to the presence or absence of other features in the same data [19].
b) C4.5 or Decision Tree is one of the algorithms of the classification type that can produce an easy-to-understand decision tree model (pattern) [19].
c) KNN (K-Nearest Neighbor) is a classification of objects that are very close to each other and require training information called supervised procedures [19].

Final data processing using Rapid Miner software. The stages in processing raw data into data that has new information commonly called data mining can be described below.

a) Data Selection
   Data is collected from https://www.kaggle.com/ website. This data totals 238 with five features, namely age, annual income, purchase amount, region, purchase frequency, and loyalty score.
b) Preprocessing
   This stage is the process of cleaning the data by eliminating duplicates, errors, and ensuring proper validation rules. Incomplete or incorrect data will be replaced or deleted. The data totaling 238 was obtained completely and there were no duplications or errors, so that the data was used by all.
c) Transformation
   Features that are still numerical are transformed into 2 options in table 1.

Table 1. Data Transformation

Age	Teenager	<=25
	Mature	>25
Annual income	Big	>=50.000
	Small	<50.000
Purchase amount	Much	>=500
	Less	<500
Purchase frequency	Often	>=15
	Seldom	<15
Loyalty score	Loyal	>=5
	Not loyal	<5

d) Modelling

The data used in this modeling process is data training and data testing that has gone through a data transformation process so that it is ready for the data mining process. After going through the data selection process, the data must then be filtered to find out which attributes can affect customer loyalty, which is referred to as target data. The target data contains relevant attributes and supports the data mining process for customer loyalty classification.

3. Results and Discussion

The data used in this study was 238 customers taken from https://www.kaggle.com/. The data used for training and testing with 10-fold cross-validation is divided into 10 parts of the data or subsets of the same size. It was repeated 10 times in the training and testing process.

The data processing in this study uses Rapid Miner software version 10.3.001. The input variable or feature consists of 5 variables, including: age, annual income, purchase amount, region, and purchase frequency. Meanwhile, for labels, the loyalty score feature is used. Customer data can be seen in the following table 2.

Table 2. Customer Data

Age	Annual Income	Purchase Amount	Region	Purchase Frequency	Loyalty Score
teenager	small	less	North	seldom	not loyal
mature	big	less	South	often	loyal
mature	big	much	West	often	loyal
teenager	small	less	East	seldom	not loyal
mature	small	less	North	seldom	not loyal
…	…	…	…	…	…
mature	big	less	West	often	loyal
mature	big	less	North	often	loyal
mature	big	much	South	often	loyal
mature	big	less	West	often	loyal
mature	big	less	North	often	loyal

The proposed process model based on the customer loyalty dataset in table 2 using Naive Bayes, C4.5 or Decision Tree, and KNN algorithms can be seen in the following figure 1.

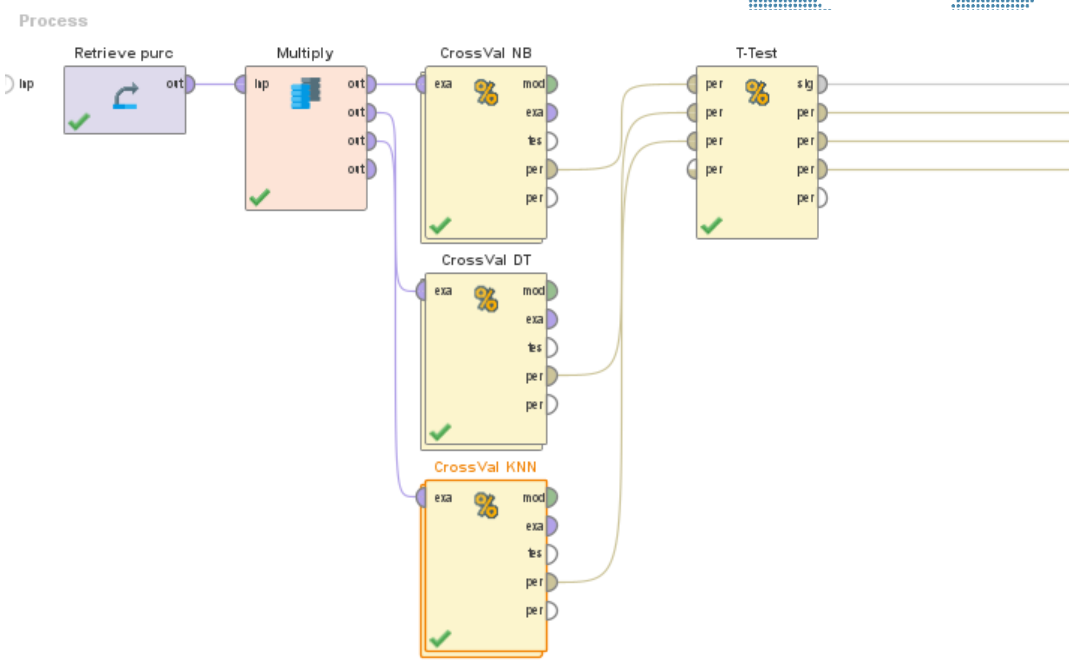

Figure 1. Process Model

In the process model, there is a training and testing model that is carried out. The training and testing model of customer loyalty data processing using the naïve bayes, C4.5, and KNN methods can be seen in figure 2.

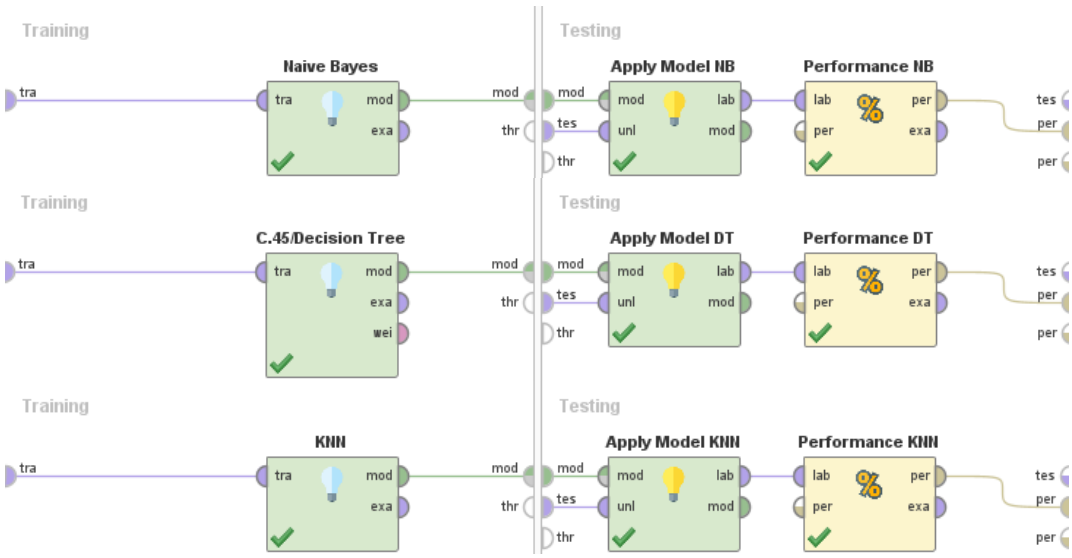

Figure 2. Model Training and Testing

From the results of the final test of the model, the results of the confusion matrix for the naïve bayes algorithm in table 3, C4.5 or Decision Tree in table 4 and KNN in table 5 were obtained.

Table 3. Naive Bayes Confusion Matrix

	true not loyal	**true loyal**	**class precision**
pred. not loyal	49	6	89.09%
pred. loyal	2	181	98.91%
class recall	96.08%	96.79%	

In table 3, pred not loyal and true not loyal amount to 49 and pred loyal and true loyal amount to 181. Class recall true not loyal as much as 96.08% and true loyal as much as 96.79%. Class precision pred not loyal 89.09% and pred loyal 98.91%.

Table 4. C4.5 Confusion Matrix

	true not loyal	**true loyal**	**class precision**
pred. not loyal	43	0	100.00%
pred. loyal	8	187	95.90%
class recall	84.31%	100.00%	

In table 4, pred not loyal and true not loyal amount to 43 and pred loyal and true loyal amount to 187. Class recall true not loyal as much as 84.31% and true loyal as much as 100%. Class precision pred not loyal 100% and pred loyal 95.90%.

Table 5. KNN Confusion Matrix

	true not loyal	**true loyal**	**class precision**
pred. not loyal	46	6	88.46%
pred. loyal	5	181	97.31%
class recall	90.20%	96.79%	

In table 5, pred not loyal and true not loyal amount to 46 and pred loyal and true loyal amount to 181. Class recall true not loyal as much as 90.20% and true loyal as much as 96.79%. Class precision pred not loyal 88.46% and pred loyal 97.31%.

The results of the ROC curve of the naïve bayes algorithm can be seen in figure 3, C4.5 or decision tree in figure 4, and KNN in figure 5.

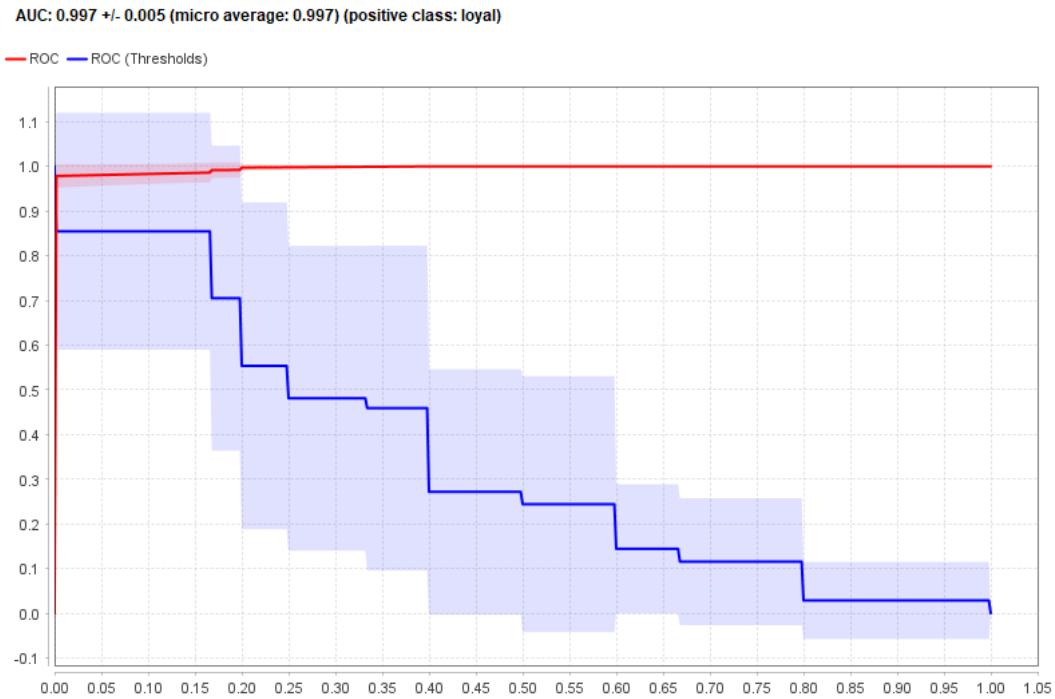

Figure 3. Naïve Bayes ROC Curve

Inside the ROC curve is the AUC value. The AUC value in figure 3 with the naïve bayes algorithm has the largest value of 0.997. AUC (optimistic) is 0.998 and AUC (pessimistic) is 0.995.

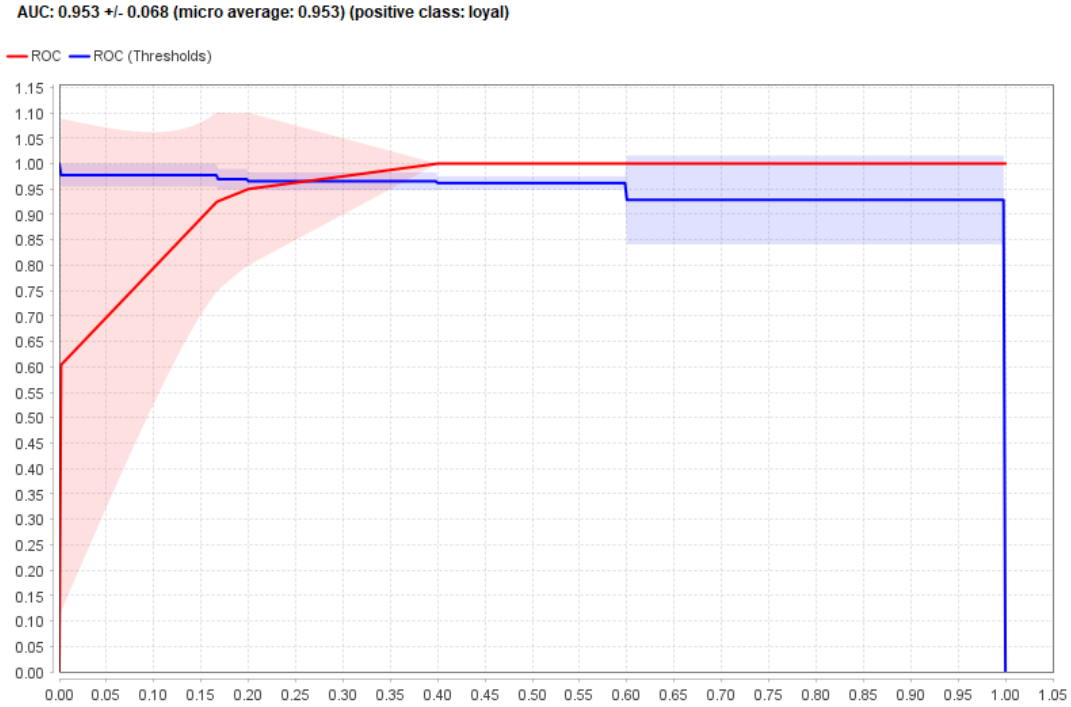

Figure 4. C4.5 ROC Curve

The AUC value in figure 4 with the C4.5 algorithm or decision tree is 0.953. AUC (optimistic) is valued at 1 and AUC (pessimistic) is valued at 0.907.

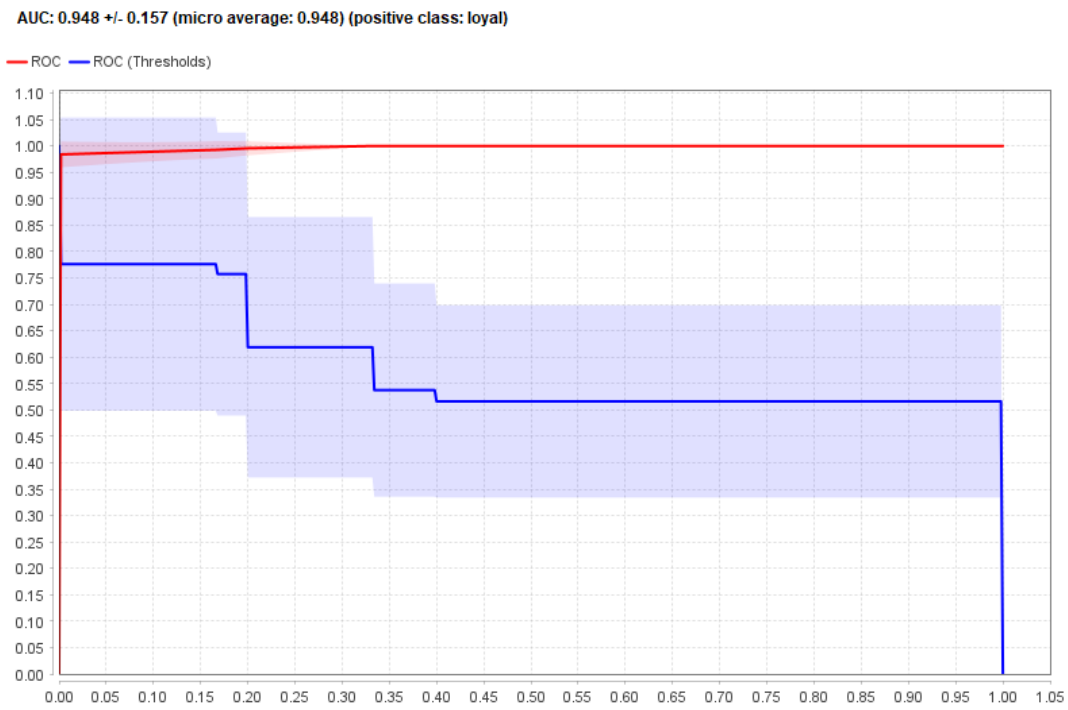

Figure 5. KNN ROC Curve

The AUC value in figure 5 with the KNN algorithm is 0.948. AUC (optimistic) is 0.999 and AUC (pessimistic) is 0.996. The results of the confusion matrix are shown by comparing the percentages of accuracy, precision, recall, and AUC values of the ROC curve in the three algorithms (table 6).

Table 6. Comparison of Confusion Matrix

Algoritms	Accuracy	Precision	Recall	AUC
Naïve Bayes	96.67%	98.92%	96.78%	0.997
C4.5	96.67%	96.13%	100%	0.953
KNN	95.40%	97.49%	96.78%	0.948

In table 6 of the confusion matrix comparison, it can be concluded that the Naive Bayes algorithm has an accuracy of 96.67% and an AUC of 0.997 higher than the C4.5 algorithm which ranks second with an accuracy of 96.67% and an AUC of 0.953. While the third place is the KNN algorithm with an accuracy of 95.40% and an AUC of 0.948.

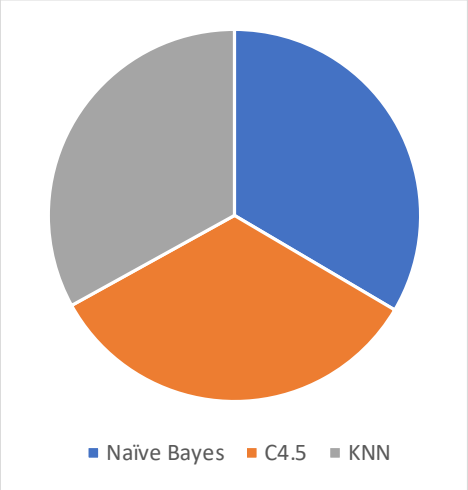

Figure 6. Comparison Accuracy

Naive Bayes' algorithm has better performance compared to C4.5 or Decision Tree and KNN which can be seen in table 6. The three algorithms or methods used have a high percentage of accuracy and AUC value, so they have high performance in processing data. The accuracy percentage in figure 6 for the Naïve Bayes and C4.5 algorithms is the same at 96.67%, while for the KNN algorithm there is only a slight difference with Naïve Bayes and C4.5 at 95.40%. The difference between Naïve Bayes and C4.5 is that Naïve Bayes' AUC value is greater at 0.997.

4. Conclusion

The conclusion of this study is that customer loyalty classification can be easily done with three algorithms, namely Naive Bayes, and C4.5 or Decision Tree, and KNN which is validated by the 10-fold cross-validation method so as to produce the highest percentage of accuracy and the similarity of accuracy values from the Naive Bayes and C4.5 algorithms, which is 96.67%. In the AUC value, it can be seen that the Naive Bayes algorithm is superior to the C4.5 algorithm or Decision Tree and KNN. The result of the highest AUC value is 0.997, the highest precision percentage is 98.92% achieved by the Naive Bayes algorithm. The result of the highest recall percentage is C4.5 of 100%. C.45 or Decision Tree also has a good AUC value of 0.953. The KNN algorithm also has a good AUC value of 0.948. The results of the AUC value and accuracy percentage on the three algorithms prove that the performance of the three algorithms is very good.

In the next research with a similar case, it is hoped that it can be developed by comparing several other classification algorithms other than KNN, Naive Bayes, or C4.5 or Decision Tree.

References

[1] Afifah Ratna Safitri And Much Aziz Muslim, "Improved Accuracy Of Naive Bayes Classifier For Determination Of Customer Churn Uses Smote And Genetic Algorithms," *Journal Of Soft Computing Exploration*, Vol. 1, No. 1, Pp. 70–75, 2020.

[2] Bayu Satrio, "Analisis Perbandingan Metode Pemilihan Atribusi Untuk Memprediksi Loyalitas Pelanggan," *Portaldata*, Vol. 2, No. 8, Pp. 1–12, 2022.

[3] Ridlo Muttaqien, Musthofa Galih P, And Andri Pramuntadi, "Implementation Of Data Mining Using C4.5 Algorithm For Predicting Customer Loyalty Of Pt. Pegadaian (Persero) Pati Area Office," *International Journal Of Computer And Information System (Ijcis)*, Vol. 2, No. 3, Pp. 64–68, 2021.

[4] Dina Maulida Rahmi And Nurman Setiawan Fadjar, "Pengaruh Pendapatan, Kesesuaian Harga Kebutuhan Pokok, Kebiasaan Berbelanja Dan Kesadaran Kesehatan Terhadap Pola Konsumsi," *Journal Of Development Economic And Social Studies*, Vol. 1, No. 4, Pp. 539–549, 2022.

[5] Ressa Artanovelia, Zulfahmi Sengaji, Egi Radiansyah, And Edison Ginting, "Analisis Pengaruh Kebiasaan, Gaya Hidup Dan Pendapatan Terhadap Keputusan Pembelian Motor Honda Vario Di Kalianda," *Kalianda Halok Gagas*, Vol. 7, No. 1, Pp. 45–55, 2024.

[6] Andi Diah Kuswanto, Said Imam Puro, Jodi Hariyan, And Ridho Rafliansyah, "Analisa Data Shopping Trends Menggunakan Algoritma Klasifikasi Dengan Metode Naive Bayes," *Publikasi Teknik Informatika Dan Jaringan*, Vol. 2, No. 3, Pp. 119–134, 2024.

[7] Sulaiman O. Abdulsalam, Ismaila Yusuf, Samson O. Ojerinde, And Abdullahi Yahaya, "Detection Of Banks' Customers Loyalty Using Naive Bayes And Support Vector Machine Classifiers: A Machine Learning Approach," *Journal Of Science And Logics In Ict Research*, Vol. 12, No. 1, Pp. 45–54, 2024.

[8] Rumana Shahid *Et Al.*, "Predicting Customer Loyalty In The Airline Industry: A Machine Learning Approach Integrating Sentiment Analysis And User Experience," *International Journal On Computational Engineering*, Vol. 1, No. 2, Pp. 50–54, 2024.

[9] Rudi Wardana Nasution, Suhada, Ika Okta Kirana, Indra Gunawan, And Ika Purnama Sari, "Penerapan Data Mining Untuk Pengelompokan Minat Konsumen Terhadap Pengguna Jasa Pengiriman Pada Pt. Jalur Nugraha Ekakurir (Jne) Pematangsiantar," *Resolusi : Rekayasa Teknik Informatika Dan Informasi*, Vol. 1, No. 4, Pp. 274–281, 2021.

[10] Andri Wijaya And Abba Suganda Girsang, "The Use Of Data Mining For Prediction Of Customer Loyalty," *Commit (Communication & Information Technology) Journal*, Vol. 10, No. 1, Pp. 41–47, 2016.

[11] Tynchenko Vadim Sergeevich, Kukartsev Vladislav Viktorovich, Boyko Andrei Anatolievich, And Danilchenko Yuriy Vitalievich, "Optimization Of Customer Loyalty Evaluation Algorithm For Retail Company," In *International Conference Economy In The Modern World (Icemw)*, 2018, Pp. 177–182.

[12] Hossein Alizadeh And Behrouz Minaei-Bidgoli, "Introducing A Hybrid Data Mining Model To Evaluate Customer Loyalty," *Engineering, Technology & Applied Science Research*, Vol. 6, No. 6, Pp. 1235–1240, 2016.

[13] Guido Van Der Heijden, Harry Collins, And Suhaib Aslam, "Predicting Customer Loyalty Using Various Regression Models," *Advanced Machine Learning*, 2019.

[14] Ngurah Agus Sanjaya Er And I Gusti Agung Gede Arya Kadyanan, "User Loyalty Prediction Using Naïve Bayes Method In 'Udatari' An Art Performance Marketplace," *Jurnal Ilmu Komputer*, Vol. 14, No. 2, Pp. 60–65, 2021.

[15] Sari Indah, Fati Gratianus Nafiri Larosa, And Yolanda Y. P. Rumapea, "Penerapan Data Mining Menggunakan Metode K-Means Untuk Penentuan Reward Pelanggan

(Studi Kasus: Ud. Penyubur Tani)," *Methomika: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, Vol. 7, No. 2, Pp. 201–207, 2023.

[16] Embun Fajar Wati, Elvi Sunita Perangin-Angin, And Anggi Puspita Sari, "Improved Naive Bayes Algorithm With Particle Swarm Optimization To Predict Student Graduation," *International Journal Of Information System & Technology*, Vol. 7, No. 6, Pp. 386–391, 2024.

[17] Embun Fajar Wati, Budi Sudrajat, And Raudah Nasution, "Modelling Of C4.5 Algorithm For Graduation Classification," *International Journal Of Information System & Technology*, Vol. 8, No. 1, Pp. 40–46, 2024.

[18] Embun Fajar Wati, Elvi Sunita Perangin-Angin, And Anggi Puspita Sari, "Prediction Of Student Graduation Using The K-Nearest Neighbors Method," *International Journal Of Information System & Technology*, Vol. 7, No. 3, Pp. 211–216, 2023.

[19] Embun Fajar Wati And Biktra Rudianto, "Penerapan Algoritma Knn, Naive Bayes Dan C4.5 Dalam Memprediksi Kelulusan Mahasiswa," *Jurnal Format*, Vol. 11, No. 2, Pp. 168–175, 2022.

[20] E. F. Wati, A. P. Sari, E. T. Alawiah, M. H. Siregar, And B. Rudianto, "Particle Swarm Optimization Comparison On Decision Tree And Naive Bayes For Pandemic Graduation Classification," In *2nd International Conference On Advanced Information Scientific Development (Icaisd)*, 2021, Pp. 1–11.