Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12182
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Customer Classification Using Naive Bayes Classifier With Genetic Algorithm Feature Selection

Juliansyah Putra Tanjung[1)*], Fenny Chintya Tampubolon[2)], Ari Wahyuda Panggabean[3)], M. Anjas Asmara Nandrawan[4)]
[1,2,3,4)] Universitas Prima Indonesia, Indonesia
[1]juliansyahputratanjung@unprimdn.ac.id, [2)] fennych19@gmail.com, [3)]aripanggabean5@gmail.com, [4)]anjasasmaran72@gmail.com

Submitted : Jan 31, 2023 | **Accepted** : Feb 6, 2023 | **Published** : Feb 12, 2023

Abstract: There is a tendency to decrease the number of speedy customers in the operational area of North Sumatra due to customer dissatisfaction. Termination of employment is carried out by the customer against PT. Telekomunikasi Indonesia, Tbk in North Sumatra. There is no management of customer data classification so that classification information based on certain product purchases cannot be known. Naïve Bayes is a classification algorithm that is easy to use but has weaknesses which result in poor performance, therefore feature selection is needed, the genetic algorithm is an algorithm that is able to select attributes in research, will be selected based on the highest weight so that the accuracy of the prediction results is more optimal. The steps taken in the measurement model using the Naive Bayes Classifier (NBC) approach and the model using the GA-NBC approach obtained accurate results from cross validation measurements, Confusion Matrix, ROC curves for the classification of existing and speedy telephone subscribers. The stages of the Naive Bayes process are: data collection, data preprocessing, processing of the Naive Bayes Classifier algorithm. Then the results are validated and evaluated using the Text Mining Algorithm, and calculating the parameters based on the genetic algorithm. The accuracy produced by the Naive Bayes Classifier model is 85.08%. The accuracy produced by the Naive Bayes Classifier model with the selection of Genetic Algorithm features increased to 89.31%.

Keywords: customer classification, Naive Bayes Classifier, Genetic Algorithm Feature Selection

INTRODUCTION

Classification is the process of planning objects into a class, group, or category based on predetermined characteristics (Muhamad et al., 2017). Customer classification needs to be done to find out how customer demographics use the product which can be seen from how many customers and the level of transactions made (Devi et al., 2020). Effective customer classification is the main problem that must be solved by companies in management relationships with customers so that it can affect the interests of large companies (Shi et al., 2020). Companies need to provide satisfactory service quality so that customers do not switch to using other brands in the products being marketed. Using data mining techniques is a method that can be used to determine customer satisfaction (Religia & Maulana, 2021).

There is a tendency to decrease the number of speedy customers in the operational area of North Sumatra due to customer dissatisfaction. Termination of employment is carried out by the customer against PT. Telekomunikasi Indonesia, Tbk in North Sumatra. There is no management of customer data classification so that classification information based on certain product purchases cannot be known.

Classification technique is an important analytical mechanism in the prediction of various levels of accuracy. Classification is one of the methods in data mining to categorize certain group items into target groups. The goal is to predict the nature of an item or data based on the class of goods available. The construction of a classification model is always determined by the available training data set (Oluwaseun & Chaubey, 2019). Data Mining provides a set of techniques for finding hidden patterns from data (Arasa & Setiawanb, 2022).

One of the algorithms that is often used to classify data using probability calculations is the Naïve Bayes Classifier Algorithm. From this calculation, each word in the document will result in a positive, negative, and

*juliansyahputratanjung@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12182
e-ISSN : 2541-2019
p-ISSN : 2541-044X

neutral category classification obtained from the previous calculation process, namely tf-idf weighting (Chatrina, Siregar et al., 2020). Naïve Bayes is a classification algorithm that is easy to use but has weaknesses that result in poor performance, because of that feature selection is needed, genetic algorithms are algorithms that are capable of selecting attributes in research, will be selected based on the highest weight so that the accuracy of more optimal prediction results (Cahya Putri Buani, 2021). Genetic Algorithm is a technique for finding optimal solutions to problems that have many solutions. This technique will search from several solutions obtained to get the best solution according to predetermined criteria or what is known as the fitness function. This algorithm is included in the group of evolutionary algorithms using the Darwinian evolutionary approach in the field of biology such as inheritance, natural selection, gene mutation and combination (crossover). Because it is an optimal search technique in the field of computer science, this algorithm is also included in the group of metaheuristic algorithms.

LITERATURE REVIEW

Naive Bayes models are a group of extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets. Because they are so fast and have so few tunable parameters, they end up being very useful as a quick-and-dirty baseline for a classification problem (VanderPlas & Vanderplas, 2016). The results of a comparative classification using the naïve Bayes algorithm and C4.5 state that the naive Bayes algorithm is the best classification model that can be used for IndiHome customer classification to determine potential and non-potential customers (Syafii et al., 2022). Genetic algorithm to optimize feature selection in the Naive Bayes algorithm so that the accuracy increases. Able to increase accuracy in Naive Bayes by optimizing the feature selection process in the form of words in the abstract of each study. The results of the experiments in this study show that the accuracy value increases by 26.06% by using the Genetic Algorithm in the feature selection process (Bakhtiar, 2020). Previous research using Naïve Bayes had an accuracy rate of only up to 69.60% after feature selection was carried out with a genetic algorithm the accuracy rate increased to 96.67%, the difference in the difference in accuracy increase was 27.07% (Cahya Putri Buani, 2021).

The genetic algorithm learning pattern informs new arrivals or new classifications with a faster time. Configuration difficulties in naïve Bayes can be helped by genetic algorithms and providing adequate modeling to describe the system. The research resulted in an accuracy value of 81.19%, it is proven that the Naïve Bayes algorithm can improve its performance with the genetic algorithm in bacterial classification (Priyanti, 2021). The implementation of the naïve Bayes method and genetic algorithm-based feature selection for online SAMBAT document classification resulted in an average accuracy of 86.12% with the highest accuracy value of 89.79% in 5 trials using 49 test data and with the best parameter value, namely many generations of 70 , a population size of 20, a crossover rate of 0.8 and a mutation rate of 0.2 (Prayogi et al., 2019).

METHOD

The steps taken in the measurement of the model with the Naive Bayes Classifier (NBC) approach and the model with the GA-NBC approach obtained accurate results from measurements of cross validation, confusion matrix, ROC curve for the classification of existing and speedy telephone subscribers.

(1) Data Collection

The data obtained is primary data from the Business Planning and Performance division of PT. Telecommunication. Tbk Sumatera Utara which consists of 878 existing telephone and speedy customer data with 14 attributes including the class label attribute (output attribute), namely the Package attribute. Of the 878 existing telephone and speedy customer data, there are 587 Indihome (3P) customers and 291 telephone and speedy customers who rejected the Indihome (Decline) offer.

(2) Product Purchase Classification Data Process

The data used is Speedy customer data and IndiHome customers as much as 200 data. Validation data is data that is used as a reference for determining classes in data testing. While data testing is data that researchers make as test material to find out the results of the classification. This data can be observed in Table 1 below, which is the result of product purchase classification data of PT Telekomunikasi.Tbk.

Table 1. Product Purchase Classification Data

No	Existing Package	Number of Cases	2p	Decline	3p	Decline
1	Inet L15h	6	6	0,01	0	0
2	Inet L50h	3	1	2	0	0,01
3	Inet U1m	15	10	5	0,02	0,02
4	Inet U2m	5	4	1	0,01	0

*juliansyahputratanjung@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12182
e-ISSN : 2541-2019
p-ISSN : 2541-044X

5	Inet U384k	25	20	5	0,03	0,02
6	Inet U384r	6	3	3	0,01	0,01
7	Inet U3m	1	1	0	0	
8	Inet U512k	4	4	0,01	0	
9	Inet U512r	2	2	0	0	
10	Inet_1mb_H	6	5	1	0,01	0
…	……	….	..	…	…	….
31	Giptv	1	1		0	0

(3) Process with Naïve Bayes Classifier Algorithm

Calculation steps using the Bayes method are as follows:

(a) Calculation of the prior probability of the Naive Bayes algorithm

Probability is the likely outcome in a given data set. For example, in the case of a mortgage, P(Y) is the default rate p(3p), which is 0.66856492. P(Y|X) is called the conditional probability, which gives the probability of the outcome given the proof, i.e. when the value of X is known below the result of the prior probability calculation.

P (3P) = 587:878 = 0,66856492

P (Decline) = 291:878 = 0,33143508

Tabel. 2 Hasil Probabilitas Prior

Atribut		number of cases	P(H|X)			
			3p	Decline	3p	Decline
Total		878	587	291	0,67	0,33
Regional	2	878	587	291	1	1
Witel	Kabupaten Asahan		12	12	0,02	0
	Kabupaten Batu Bara	55	1	54	0	0,19
	Kabupaten Dairi	89	12	77	0,02	0,26
	……	……..	….	…	…	…..
Kwadran	4	878	587	291	1	1
Type Jaringan	Copper	509	386	123	0,66	0,42
	Ftth	231	110	121	0,19	0,42
	Msan	138	91	47	0,16	0,16
Zona	Attack_Promo	16	10	6	0,02	0,02
	Super_Winning	745	493	252	0,84	0,87
	Winning	117	84	33	0,14	0,11
Migrasi	2p	878	587	291	1	1
R2bb	Giptv	228	107	121	0,18	0,42
	Gsp5	1	1	0	0	0
	……	…	..	..	..	….
Paket Eksisting	Inet L15h	6	6	0,01	0	
	Inet L50h	3	1	2	0	0,01
	Inet U1m	15	10	5	0,02	0,02
	……	…	….	…	…	….
	Giptv	1	1		0	0

(b) Posterior Probability Results

*juliansyahputratanjung@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12182
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Posterior probability in Naive Bayes statistics is the revised or updated probability of an event occurring after considering new information. The posterior probability is calculated by updating the previous probability using Bayes' theorem. If there are cases like the following:

Table 3 Probabilitas Posterior Result

No	Atribut	Value	P(H|X)	
			3p	Dec line
1.	Regional	2	1	1
2.	Witel	Sumut	0,2	0,04
3.	Kwadran	4	1	1
4.	Paket Eksisting	Inetr1m1	0,34	0,34
5.	Type Jaringan	Sp7	0,2	0,17
6.	Zona	*Winning*	0,14	0,11
7.	Migrasi	2p	1	1
8.	R2bb	Msan	0,16	0,16

1. P(H|3P) = 1 x 0,20 x 1 x 0,34 x 0,20 x 0,16 x 1 x 0,14 = 0,000304
2. P(H|Decline) = 1 x 0,04 x 1 x 0,34 x 0,17 x 0,16 x 1 x 0,11 = 0,000040
3. P (H|paket = 3P) P(3P) = 0,000304 x 0,66856492 = 0,000203
4. P(H|paket = Decline) P(Decline) = 0,000040 x 0,33143508= 0,000013
   so P (H|paket = 3P) P (3P) > P(H|paket = Decline) P(Decline)
5. From the results of these calculations it is known the value P (H|3P) bigger than value P (H|Decline), so it can be concluded that for the case entered into the 3P classification.

From the results of these calculations it is known that the P value (H|3P) is greater than the P value (H|Decline), so it can be concluded that this case is included in the 3P classification.
(c) Measuring the Accuracy Level of the Naïve Bayes Classifier Model
The parameters used in the performance operator are accuracy, classification error, Area Under Curve (AUC) to display the level of accuracy of the NBC model and the GA-NBC model. Based on the analysis of the results of the evaluation of data mining for classification with the Naive Bayes algorithm, it can be summarized as follows:

Table 4. Comparison of Performance Results

Measurement	Result	
	NBC	GA - NBC
Accuracy	85,08%	89,31%
Classification Error	14,92%	10,69%
AUC	0.841	0.843

The resulting accuracy value based on the confusion matrix is 89.31%. The AUC result is 0.843 so it can be concluded that the Naive Bayes model with the Genetic Algorithm feature selection is the right model to be used as a customer classification with an accuracy of 0.80 – 0.90 = Good classification.
From the data generated from the Genetic Algorithm Process-Naive Bayes Classifier which is used to determine the class in new cases, the classification of existing customers is carried out using Rapidminer 5.3 software. The GA-NBC model will select the existing predictive attributes. The parameters used in the performance operator are accuracy, classification error, accuracy, classification error is the ratio of correct predictions (positive and negative) to the entire data.
The parameter used in operator performance is Area Under Curve (AUC). AUC is the area under the curve (Area under the Curve of) ROC (Receiver Operating Characteristics), a curve that describes probabilities with sensitivity and specificity variables with a limit value between 0 and 1 to display the level of accuracy of the NBC model and the GA-NBC model. Based on the analysis of the results of the evaluation of data mining for classification with the Naive Bayes algorithm, it can be seen as follows:

*juliansyahputratanjung@unprimdn.ac.id

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7 Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12182
e-ISSN : 2541-2019
p-ISSN : 2541-044X

PerformanceVector
PerformanceVector:
accuracy: 85.06% +/- 3.47% (mikro: 85.06%)
ConfusionMatrix:
True: DECLINE 3P
DECLINE: 213 53
3P: 78 534
classification_error: 14.92% +/- 3.47% (mikro: 14.92%)
ConfusionMatrix:
True: DECLINE 3P
DECLINE: 213 53
3P: 78 534
AUC (optimistic): 0.842 +/- 0.041 (mikro: 0.842) (positive class: 3P)
AUC: 0.841 +/- 0.041 (mikro: 0.841) (positive class: 3P)
AUC (optimistic): 0.840 +/- 0.041 (mikro: 0.840) (positive class: 3P)

Competitive performance of Naïve Bayes in the classification process even using the assumption of attribute independence. The performance of Naïve Bayes classification when numerical attributes are discretized rather than assumed with the distribution approach as above [Dougherty]. Numerical values will be mapped to nominal values in the form of fixed intervals.

RESULT

From the Posterior Probability results data used to determine the class in new cases then the Genetic Algorithm-Naive Bayes Classifier is calculated. The GA-NBC model will select the existing predictive attributes. The following is the result of the attribute selection:

Table 5 Genetic Algorithm

NO	ATRIBUT	WEIGHT
1.	Regional	0
2.	Witel	1
3.	Kwadran	1
4.	Paket Existing	0
5.	R2bb	0
6.	Type Jaringan	1
7.	Zona	0
8.	Migrasi	0

The parameters used in the performance operator are accuracy, classification error, Area Under Curve (AUC) to display the level of accuracy of the NBC model and the GA-NBC model. Based on the analysis of the results of the evaluation of data mining for classification with the Naive Bayes algorithm, it can be summarized as Fig. 2 follows, where Naïve Bayes performance is competitive in the classification process even though it uses the assumption of attribute independence, Naive Bayes classification performance and obtains Accuracy results of 89.31%, Classification Error 10.69%, AUC 0.843:

PerformanceVector
PerformanceVector:
accuracy: 89.31% +/- 3.65% (mikro 89.29%)
ConfusionMatrix:
True: DECLINE 3P
DECLINE: 216 19
3P: 75 568
classification_error: 10.69% +/- 3.65% (mikro: 10.71%)
ConfusionMatrix:
True: DECLINE 3P
DECLINE: 216 19
3P: 75 568
AUC: 0.843 +/- 0.062 (mikro: 0.843) (positive class: 3P)

*juliansyahputratanjung@unprimdn.ac.id

The accuracy value generated by 1. Confusion Matrix is 89.31%. The AUC result is 0.843 so it can be concluded that the Naive Bayes model with the Genetic Algorithm feature selection is the right model to be used as a customer classification with an accuracy of 0.80 – 0.90 = Good classification.

DISCUSSIONS

Based on the results of applying the NBC Model genetic algorithm, the parameter used in the performance operator is Area Under Curve (AUC). AUC is the area under the curve (Area under the Curve of) ROC (Receiver Operating Characteristics), a curve that describes probabilities with sensitivity and specificity variables with a limit value between 0 to 1 to display the level of accuracy of the NBC model and the GA-NBC model the result of Confusion Matrix is 85.08%.

Then in the ROC Curve process, the AUC result is 0.841 so that it can be concluded that this Naive Bayes model is a good model to be used as a customer classification that accuracy 0.80 – 0.90 = Good classification, Performance of Naive Bayes classification and get Accuracy results of 89.31%, Classification Error 10.69%, AUC 0.843 so it can be concluded that the Naive Bayes model with feature selection of the Genetic Algorithm is a good model to be used as a customer classification with an accuracy of 0.80 – 0.90 = Good classification.

CONCLUSION

The results showed that with the feature selection genetic algorithm, the naive Bayes classifier model becomes a more precise and accurate classification model to be applied to the classification of PT. Telekomunikasi Indonesia, Tbk in North Sumatra in marketing its new product, namely Indihome. The accuracy produced by the Naive Bayes Classifier model is 85.08%. The accuracy produced by the Naive Bayes Classifier model with the Genetic Algorithm feature selection increased to 89.31%.

REFERENCES

Arasa, R. A., & Setiawanb, N. A. (2022). Comparison Of Data Mining Classification Techniques For Heart Disease Prediction System. *Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *2*(2), 85–90.

Bakhtiar, M. Y. (2020). Klasifikasi Penelitian Dosen Menggunakan Naïve Bayes Classifier dan Algoritma Genetika. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *5*(2), 134. https://doi.org/10.30998/string.v5i2.6912

Cahya Putri Buani, D. (2021). Penerapan Algoritma Naïve Bayes dengan Seleksi Fitur Algoritma Genetika Untuk Prediksi Gagal Jantung. *EVOLUSI : Jurnal Sains Dan Manajemen*, *9*(2). https://doi.org/10.31294/evolusi.v9i2.11141

Chatrina, Siregar, N., Ruli, A, Siregar, R., & Yoga, Distra, Sudirman, M. (2020). Implementasi Metode Naive Bayes Classifier (NBC) Pada Komentar Warga Sekolah Mengenai Pelaksanaan Pembelajaran Jarak Jauh (PJJ). *Jurnal Teknologia*, *34*(1), 102–110. https://aperti.e-journal.id/teknologia/article/view/67

Devi, C., Soleman, O., Pramaita, N., & Sudarma, M. (2020). Classification Of Loyality Customer Using K-Means Clustering, Studi Case : PT. Sucofindo (Persero) Denpasar Branch. *International Journal of Engineering and Emerging Technology*, *5*(2), 160–167.

Muhamad, H., Prasojo, C. A., Sugianto, N. A., Surtiningsih, L., & Cholissodin, I. (2017). Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *4*(3), 180. https://doi.org/10.25126/jtiik.201743251

Oluwaseun, A., & Chaubey, M. S. (2019). Data Mining Classification Techniques on the analysis of student performance. *Global Scientific Journal*, *7*(April), 79–95.

Prayogi, T. F., Cholissodin, I., & Santoso, E. (2019). Klasifikasi Dokumen SAMBAT Online Menggunakan Metode Naive Bayes dan Seleksi Fitur Berbasis Algoritme Genetika. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(3), 2173–2179. http://j-ptiik.ub.ac.id

Priyanti, E. (2021). Peningkatan Algoritma Naïve Bayes Menggunakan Algoritma Genetika Pada Klasifikasi Bakteri. *Swabumi*, *9*(2), 78–81. https://doi.org/10.31294/swabumi.v9i2.11217

Religia, Y., & Maulana, D. (2021). Genetic Algorithm Optimization on Nave Bayes for Airline Customer Satisfaction Classification. *JISA(Jurnal Informatika Dan Sains)*, *4*(2), 121–126. https://doi.org/10.31326/jisa.v4i2.925

Shi, X., Li, G., Li, K., Liu, J., & Wang, R. (2020). Customer Classification Method of Logistics Enterprises Based on BP-AdaBoost. *Journal of Physics: Conference Series*, *1670*(1). https://doi.org/10.1088/1742-6596/1670/1/012018

Syafii, A., Dwilestari, G., & Ajiz, A. (2022). Komparasi Algoritma Naïve Bayes Dan Algoritma C4.5 Dalam Klasifikasi Pelanggan Produk Indihome. *Jurnal Sistem Informasi Dan Manajemen(JURISMA)*, *10*(2), 60–70. https://ejournal.stmikgici.ac.id/

VanderPlas, J., & Vanderplas, J. T. (2016). *Python Data Science Handbook // Python data science handbook*. O'Reilly Media, Inc.

*juliansyahputratanjung@unprimdn.ac.id