e-ISSN: 2614-8404
p-ISSN : 2776-3234

Genetic Algorithm Optimization on Naive Bayes for Airline Customer Satisfaction Classification

Donny Maulana[1 *)], Yoga Religia[2]
[1,2]Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University
email: [1]donny.maulana@pelitabangsa.ac.id , [2]yoga.religia@pelitabangsa.ac.id

Abstract −Airline companies need to provide satisfactory service quality so that people do not switch to using other airlines. The way that can be used to determine customer satisfaction is to use data mining techniques. Currently, the website www.kaggle.com has provided Airline Passenger Satisfaction data consisting of 22 attributes, 1 label and 25976 instances which are included in the supervised learning data category. Based on several previous studies, the Naïve Bayes algorithm can provide better classification performance than other classification algorithms. Several studies also state that the use of Naive Bayes can be optimized using Genetic Algorithm (GA) to obtain better performance. The use of Genetic Algorithm for Nave Bayes optimization in classifying Airline Passenger Satisfaction data requires further research to ensure the performance of the given classification. This study aims to compare the use of the Naive Bayes algorithm for the classification of Airline Passenger Satisfaction with and without GA optimization. The data validation process used in this study is to use split validation to divide the dataset into 95% training data and 5% testing data. The test results show that the use of GA on Naive Bayes can improve the classification performance of Airline Passenger Satisfaction data in terms of accuracy and recall with an accuracy value of 85.99% and a recall of 87.91%.

Keywords - data mining, classification, Naïve Bayes, Genetic Algorithm, Customer Satisfaction.

I. INTRODUCTION

Geographically, Indonesia, which is an archipelagic country, requires transportation facilities that make it easier for people to accommodate accommodation, one of which is by air. This is a great potential that can be taken by airline companies [1]. Airline companies need to provide satisfactory service quality so that people do not switch to using other airlines [2]. The service quality of an airline cannot be measured from the company's point of view, but must be seen from the point of view of customer satisfaction [3]. The method that can be used to determine customer satisfaction is to use data mining techniques [4].

One way that can be used to predict customer satisfaction with data mining techniques is by using a classification model. Classification models can be used on supervised learning data [5]. Currently on the websitewww.kaggle.com has provided Airline Passenger Satisfaction data consisting of 22 attributes, 1 label and 25976 instances included in the supervised learning data category [6], so that it can be used to create a classification model. It takes a good algorithm for making an optimal classification model, one of which uses the Naïve Bayes algorithm.

Based on several previous studies, the Naïve Bayes algorithm can provide better classification performance than other classification algorithms such as k-NN, C4.5, Decision Tree, and even Neural Networks. [7] [8] [9]. These studies try to compare the Naive Bayes algorithm with classification algorithms to predict various types of datasets to find out which algorithm has the best performance. Besides being able to provide good classification performance, the Naïve Bayes algorithm can also be used for imbalance data [10] [11], so it is suitable to be used to classify Airline Passenger Satisfaction data.

Although Nave Bayes has shown outstanding classification accuracy, currently independent assumptions are rarely discussed in the Nave Bayes classification. One way to try independent assumptions in the Naïve Bayes algorithm is by attribute weighting [12]. This is also supported by Liangxiao Jiang (2019) which states that it is necessary to propose an attribute weighting method to reduce independent assumptions [13]. Attribute weighting can be done using Genetic Algorithm (GA) through Feature Selection [14].

GA is one of the optimization algorithms created to mimic some of the processes observed in natural evolution [15]. The optimization carried out by GA is to predict the right number of iterations, so that there is no need to calculate the number of different iterations to get complete occurrences of independent paths. [16]. The most significant advantage of GA is its ability to search globally as well as adaptability to a wide spectrum of problems [17]. Based on several previous studies, it is stated that the use of GA can improve the classification performance of Naïve Bayes [18] [19].

Based on previous research, it shows that GA is able to improve classification performance on Naïve Bayes, but has not found the application of GA to Naïve Bayes for the classification of airline customer satisfaction. This study analyzes GA optimization on Naïve Bayes for the classification of Airline Passenger Satisfaction data.

II. RESEARCH METHODOLOGY

A. Data used

This study uses Airline Passenger Satisfaction data taken from the sitewww.kaggle.com on April 24, 2021 [6]. Airline Passenger Satisfaction data is data that contains a survey of airline passenger satisfaction in the world. Airline Passenger Satisfaction data is still a new dataset that has not been widely used for research because the data has been uploaded to the site www.kaggle.com since May 2020. This data has 1 label with a boolean data type consisting of 22 attributes and 25976 instances. The purpose of using this data is to find out what factors are most correlated with airline passenger satisfaction, so that this data is suitable to be used to create a classification model. Each attribute and label contained in the Airline Passenger Satisfaction data can be seen in Table 1.

Table1. Airline Passenger Satisfaction Attributes and Labels

Content	Information	Ket
Gender	Passenger gender (Female, Male)	Attribute
Customer Type	Type of customer (Loyal customers, disloyal customers)	Attribute
age	Actual passenger age	Attribute
Type of Travel	Passenger flight destinations (Private Travel, Business Trip)	Attribute
Class	Class of travel on passenger aircraft (Business, Eco, Eco Plus)	Attribute
flight distance	Flight distance of this trip	Attribute
Inflight wifi service	Satisfaction level of inflight wifi service (1-5)	Attribute
Arrival time convenient	Satisfaction level Departure / Arrival time comfortable (1-5)	Attribute
Ease of Online booking	Online order satisfaction level (1-5)	Attribute
Gate location	Gate location satisfaction level (1-5)	Attribute
Food and drink	Food and beverage satisfaction level (1-5)	Attribute
Online boarding	Online boarding satisfaction level (1-5)	Attribute
Seat comfort	Seat comfort level of satisfaction (1-5)	Attribute
Inflight entertainment	Satisfaction level of inflight entertainment (1-5)	Attribute
On-board service	On-board service satisfaction level (1-5)	Attribute
Leg room service	Room service satisfaction level (1-5)	Attribute
Baggage handling	Baggage handling satisfaction level (1-5)	Attribute
Check-in service	Check-in service satisfaction level (1-5)	Attribute
Inflight service	In-flight service satisfaction level (1-5)	Attribute
Cleanliness	Cleanliness satisfaction level Tingkat (1-5)	Attribute
Departure Delay	Minutes delayed on departure	Attribute
Arrival Delay	Minutes delayed on Arrival	Attribute
Satisfaction	Airline satisfaction level (Satisfied, Dissatisfied)	Label

Airline Passenger Satisfaction Data does not have a missing value, so it can be directly used for the classification process without the need to go through preprocessing data.

B. Research Model

Airline Passenger Satisfaction data is used to form a classification model. The label used is the attribute "Satisfaction" with a value of "Satisfied" and "Unsatisfied". From all data used, 66% are instances labeled "Not Satisfied" while the rest are instances labeled "Satisfied". This research carried out the test twice which later will be analyzed the results obtained. The first test is done using GA optimization, while the second test is done without GA optimization.

The classification model built in this study uses the spit validation process to divide the data into training data and testing data. The training data used in this study is 95% of all Airline Passenger Satisfaction data, while the remaining 5% is used for testing data. The training data obtained from the validation process will be used for classification modeling using the Naïve Bayes algorithm. The resulting model is then used as an apply model for use in testing data. After the classification has been carried out, then the performance of the classification model is measured based on the values of accuracy, precision, and recall.

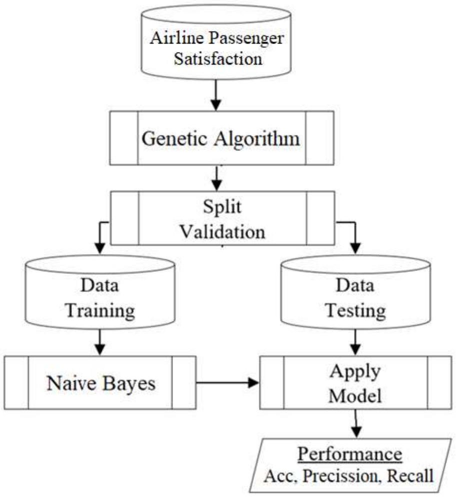

Figure 1. First Test of Naïve Bayes Classification Using Genetic Algorithm

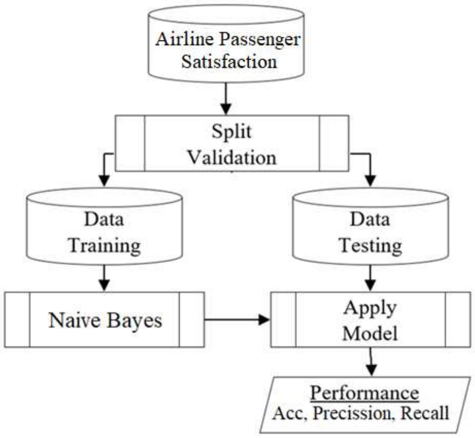

Figure 2. First Test of Naïve Bayes Classification Without Using Genetic Algorithm

In Figure 1 and Figure 2 shows that in this study the test was carried out 2 times, namely: (1) Classification of Airline Passenger Satisfaction data using Naïve Bayes with optimization of Genetic Algorithm, (2) Classification of Airline Passenger Satisfaction data using Naïve Bayes without optimization of Genetic Algorithm . The performance results of the two tests will be compared and then analyzed to show the research findings.

C. Classification with Naïve Bayes

Naïve Bayes is widely used to solve classification problems in real-world applications because of its ease of building and interpreting data, and its good performance. [13]. The Naïve Bayes algorithm is a supervised learning algorithm based on the Bayes theorem with the assumption of independence between predictors. This means that the features in the class are independent of other features. The Naive Bayes classifier can be used for both continuous and categorical variables [12]. It is based on the Bayes formula which is the probability of event A given proof of B which can be seen in the following equation [7]:

$$P(A,B) = P(A)P(B) \quad (1)$$

Through equation (1) and using the concept of the Bayes theorem, the final equation of the Naïve Bayes algorithm is obtained as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (2)$$

Based on equation (2), it is known that A is a class and B is an instance. A represents the dependent event which means the predicted variable and B represents the previous event which means the predictor attribute. The final step of the Naive Bayes algorithm is to find the maximum probability that will serve as a predictor class.

D. Optiomation with Genetic Algorithm

Genetic Algorithm (GA) was discovered by John Holland in 1960 who was inspired by the process of evolution in nature [20]. GA is an optimization method developed based on the mechanism of natural selection by imitating the genetics of living things in solving difficult problems with high complexity and undesirable structures. [21]. The optimization process in GA is carried out based on the sample population by developing a population candidate solution towards a better solution [22].

The first step of GA is the formation of chromosomes. Each chromosome yields one answer to one problem. New answers are generated after applying the crossover, mutation, and selection operations. The fitness function evaluates the benefits of chromosomes. GA then finds the most feasible chromosome with the maximum fitness function value from generation to generation. Many circumstances such as initial population size, number of generations, crossover operator, mutation operator and fitness function determine the performance of the genetic algorithm [23]. Fewer generations are required to reach the optimal answer in order to produce a more accurate fitness function.

E. Evaluation with Cross Validation

The cross validation method or also known as k-fold cross validation is a validation method that involves splitting a random sample set into a series of equal-sized folds (groups), where k indicates the number of partitions, or folds, the data set is broken down. [24]. For example, if the k value of ten is used, the data set is divided into ten partitions. In this case, nine partitions are used for training data, while the other partitions are used for data testing. The training is repeated ten times, each time using a different partition as the test set, then the other nine partitions are used as training data. The results are then averaged for reporting [25].

F. Confution Matrix for Performance Testing

In a binary confution matrix, observations that are correctly classified into a positive class are called true positives (TP) and observations that are correctly classified into a negative class are called true negatives (TN). Instances of a positive class that are classified incorrectly as negative are called false negatives (FN) and instances of a negative class that are classified incorrectly as positive are called false positives (FP). Based on the values of TP, FP, TN and TP, classification performance indicators can be calculated that reflect how the classifier performs in detecting a given class. The most commonly used indicators are accuracy, precision, recall (sensitivity) which can be written in the following equation [26]:

$$Akurasi = \frac{TP+TN}{(TP+TN+FP+F\ \ )} \quad (3)$$

$$Presisi = \frac{TP}{(TP+FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP+F\ \ )} \quad (5)$$

Accuracy is the simplest and most widely used metric for measuring the performance of a classification model. In addition to using accuracy, this study also considers classification performance measures in terms of precision and recall. According to Brendan Juba and Hai S. Le (2019), classification performance measures using accuracy, precision and recall are recommended because they are suitable for classification of imbalance data. [27].

III. RESULTS AND DISCUSSION

A. Testing Step

The Rapid Miner version 5.0 tools were used in this study to conduct testing. Rapid Miner can be used for research, rapid prototyping, and supports all steps of the data mining process such as data preparation, result visualization, validation and optimization. [28], so it is considered suitable for use in this study. The first stage in making a research model is to call the data *Airline Passenger Satisfaction*Rapid Miner tools, then the multiply function is performed to perform two tests at once, namely testing using GA and testing without using GA. The data validation process is carried out using split validation to divide the data into 95% training data and 5% testing data. In more detail about the data calling and validation process can be seen in Figure 3.

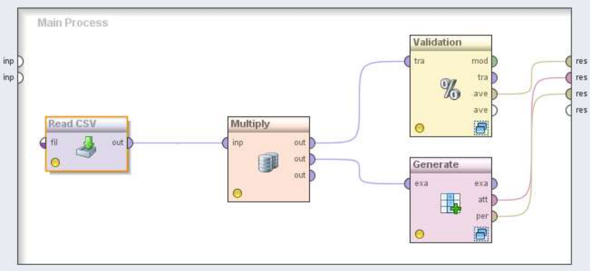

Figure 3. Data Calling and Validation Process

In each validation process shown in Figure 3, it contains a learning process with the Naïve Bayes algorithm which is then applied to the apply model to measure the performance of accuracy, precision and recall. The learning process in this study can be seen in Figure 4.

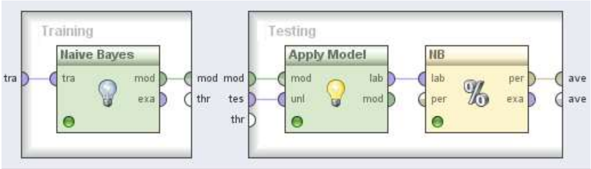

Figure 4. The Learning Process of Naïve Bayes and Apply Model

The next step after all research models have been formed is to run the model that has been built on Rapid Miner, then the results of accuracy, precision and recall will be obtained for analysis of the results.

B. Test result

After 2 tests, the accuracy, precision, and recall values of the two models were obtained. More complete test results can be seen in Table 2.

Table2. Test result

No.	Algorithm	Accuracy	Precision	Recall
1	Naive Bayes	84.53%	88.47%	84.90%
2	GA + Naive Bayes	85.99%	87.43%	87.91%

Based on Table 2, it can be seen that GA is able to improve the accuracy and recall of Naïve Bayes, but GA has not been able to increase the precision value of Naïve Bayes. The test results show that with an accuracy of 85.99%, GA optimization gives Naïve Bayes an increase in accuracy value of 1.46% and an increase in recall value of 3.01% for Airline Passenger Satisfaction data classification.

Table3. Genetic Algorithm Weighting Results on Data*Airline Passenger Satisfaction*

Attribute	Weighting
Gender	0
Customer Type	0
age	0
Type of Travel	0
Class	1
Flight Distance	0
Inflight wifi service	1
Departure / Arrival time convenient	0
Ease of Online booking	0
Gate location	0
Food and drink	0
Online boarding	0
Seat comfort	0
Inflight entertainment	0
On-board service	1
Leg room service	0
Baggage handling	0
Checkin service	1
Inflight service	0
Cleanliness	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	0

However, it turns out that the use of GA also reduces the precision value by 1.04% from the use of Naïve Bayes for the classification of Airline Passenger Satisfaction data. This is presumably because of the 22 attributes in the Airline Passenger Satisfaction data, it turns out that only 3 attributes are weighted by GA. This weighting result also explains why the increase in accuracy and recall provided by GA is not too large. In Table 3, it can be seen that there are only 4 attributes that are weighted by GA. This shows that, based on the 4th GA, these attributes are the most important to consider when classifying Airline Passenger Satisfaction data. The attributes are: Class, Inflight wifi service, On-board service and Check-in service.

C. Discussion of Results

Based on the results of the tests that have been carried out, classification Airline Passenger Satisfaction data has shown that the use of GA optimization can improve the accuracy and recall performance of the Naïve Bayes algorithm, although not too large. The small increase in performance given is thought to be because the attributes given weighting by GA are less than 25% of all the attributes in the Airline Passenger Satisfaction data. This makes the probability calculation process in Naïve Bayes less influential. Even in terms of precision, it turns out that the use of GA actually decreases the performance of Naïve Bayes.

Although the optimization of GA does not give maximum results, by using GA it turns out which attributes can be obtained which can be used as evaluation priorities to see the satisfaction of airline customers. By looking at the attributes given weighting by GA, it can be used as a reference to consider these attributes as the main focus for service improvement. The attributes that are given weighting by GA include: Class, Inflight wifi service, On-board service and Checkin service. This finding is expected to provide a practical contribution to the future services that will be provided by airlines to their customers.

VI. CONCLUSION

This study has tested the use of the Naïve Bayes algorithm to classify Airline Passenger Satisfaction data and compared it with the Nave Bayes classification using GA optimization. Based on the tests that have been carried out, it shows several results, namely:

1. The highest accuracy and recall of Airline Passenger Satisfaction data classification is using the Naïve Bayes algorithm with GA optimization. The maximum accuracy obtained is 85.99% and the maximum recall is 87.91%.
2. The maximum precision value from the classification of Airline Passenger Satisfaction data is to use the Naïve Bayes algorithm without GA optimization with a precision value of 88.47%.
3. The GA algorithm has not been able to provide maximum performance addition to the Naïve Bayes algorithm to classify Airline Passenger Satisfaction data.
4. Attributes Class, Inflight wifi service, On-board service and Checkin service are attributes that need to be considered by airlines to maximize customer satisfaction.

The results of this study are still not able to provide a good enough performance for Airline Passenger Satisfaction data classification, because neither accuracy, precision nor recall has a score of more than 90%. This requires further research to obtain a better Airline

Passenger Satisfaction data classification model in the future. Based on the findings of this study, it is suggested that future research can apply other optimization methods to further optimize the performance of the Naïve Bayes algorithm, for example the Particle swarm optimization (PSO) algorithm or boostraping.

REFERENCES

[1] E. L. Widjaja, A. Aprilia dan A. Harianto, “Analisa Pengaruh Kualitas Layanan Terhadap Kepuasan Penumpang Maskapai Penerbangan Batik Air,” *Jurnal Hospitality dan Manajemen Jasa,* vol. 5, no. 2, pp. 118-132, 2017.

[2] W. Ardhia, “Tingkat Kepuasan Penumpang Terhadap Layanan Maskapai Penerbangan PT. Lion Air Rute Menuju Jakarta,” *Jurnal Perhubungan Udara,* vol. 41, no. 1, pp. 19-28, 2015.

[3] M. D. Darus dan K. Mahalli, “Analisis Tingkat Kepuasan Penumpang Terhadap Kualitas Pelayanan di Bandar Udara Internasional Kualanamu,” *Jurnal Ekonomi dan Keuangan,* vol. 3, no. 6, pp. 408-420, 2015.

[4] M. S. Garver, “Using Data Mining for Customer Satisfaction Research,” *Marketing Research,* vol. 14, no. 1, pp. 8-17, 2002.

[5] V. Gopalakrishnan dan C. Ramaswamy, “Patient Opinion mining to Analyze Drugs Satisfaction Using Supervised Learning,” *Journal of Applied Research and Technology,* vol. 15, no. 1, pp. 311-319, 2017.

[6] Kaggle, “Kaggle.com,” Mei 2020. [Online]. Available: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction. [Diakses 24 Maret 2021].

[7] I. A. A. Amra dan A. Y. A. Maghari, “Students Performance Prediction Using KNN and Naïve Bayesian,” dalam *8th International Conference on Information Technology (ICIT)*, Al-Zaytoonah University of Jordan, Jordan, 2017.

[8] F. Osisanwo, J. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi dan J. Akinjobi, “Supervised Machine Learning Algorithms: Classification and Comparison,” *International Journal of Computer Trends and Technology (IJCTT),* vol. 48, no. 3, pp. 128-138, 2017.

[9] E. N. Azizah, U. Pujianto, E. Nugraha dan Darusalam, “Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student Academic Performance in A Virtual Learning Environment,” dalam *4th International Conference on Education and Technology (ICET)*, Malang, Indonesia, 2018.

[10] K. Madasamy dan M. Ramaswami, “Data Imbalance and Classifiers: Impact and Solutions from A Big Data Perspective,” *International Journal of Computational Intelligence Research,* vol. 13, no. 9, pp. 2267-2281, 2017.

[11] E. M. Hassib, A. I. El-Desouky, E.-S. M. El-Kenawy dan S. M. El-Ghamrawy, “An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance,” *Journal & Magazines,* vol. 7, no. 1, pp. 170774-170795, 2019.

[12] S. Chen, G. I. Webb, L. Liu dan X. Ma, “A Novel Selective Naïve Bayes Algorithm,” *Knowledge-Based Systems,* vol. 192, pp. 1-15, 2020.

[13] L. Jiang, L. Zhang, L. Yu dan D. Wang, “Class-Specific Attribute Weighted Naive Bayes,” *Pattern Recognition,* vol. 88, no. 1, pp. 321-330, 2019.

[14] S. Ernawati, R. Wati, N. Nuris, L. S. Marita dan E. R. Yulia, “Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application,” *Journal of Physics: Conference Series,* vol. 1641, pp. 1-7, 2020.

[15] S. Ernawati, E. R. Yulia, Frieyadie dan Samudi, “Implementation of The Naïve Bayes Algorithm with Feature Selection using Genetic Algorithm for Sentiment Review Analysis of Fashion Online Companies,” dalam *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, Medan, Indonesia, 2018.

[16] A. Arwan dan D. S. Rusdianto, “Optimization of Genetic Algorithm Performance Using Naïve Bayes for Basis Path Generation,” *Kinetik,* vol. 2, no. 4, pp. 273-282, 2017.

[17] E. Stripling, S. v. Broucke, K. Antonio, B. Baesens dan M. Snoecka, “Profit Maximizing Logistic Model for Customer Churn Prediction Using Genetic Algorithms,” *Swarm and Evolutionary Computation,* vol. 40, no. 1, pp. 116-130, 2018.

[18] D. K. Choubey, S. Paul, S. Kumar dan S. Kumar, “Classification of Pima Indian Diabetes Dataset Using Naive Bayes With Genetic Algorithm As An Attribute Selection,” dalam *The International Conference on Communication and Computing Systems (ICCCS)*, Ranchi, India, 2016.

[19] L. G. P. Suardani, I. M. A. Bhaskara dan M. Sudarma, “Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients,” *International Journal of Engineering and Emerging Technology,* vol. 3, no. 1, pp. 66-70, 2018.

[20] M. Melanie, An Introduction to Genetic Algorithms, London, England: First MIT Press, 1999.

[21] Y. Religia, A. Nugroho dan W. Hadikristanto, “Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing,” *Jurnal Rekayasa Sistem dan Teknologi Informasi,* vol. 5, no. 1, pp. 187-192, 2021.

[22] E. Habibi, M. Salehi, G. Yadegarfar dan A. Taheri, “Optimization of ANFIS Using A Genetic Algorithm for Physical Work Rate Classification,” *International Journal of Occupational Safety and Ergonomics,* vol. 26, no. 3, pp. 436-443, 2020.

[23] H. Motieghader, A. Najafi, B. Sadeghi dan A. Masoudi-Neja, “A Hybrid Gene Selection Algorithm for Microarray Cancer Classification Using Genetic Algorithm and Learning Automat,” *Informatics in*

Medicine Unlocked, vol. 9, no. 1, pp. 246-254, 2017.

[24] C. A. Ramezan, T. A. Warner dan A. E. Maxwell, “Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification,” *Remote Sensing,* vol. 11, no. 2, pp. 2-21, 2019.

[25] M. Stone, “Cross-Validatory Choice and Assessment of Statistical Predictions,” *Journal of the Royal Statistical Society,* vol. 36, no. 2, pp. 111-147, 1974.

[26] S. Ruuskaa, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen dan J. Mononen, “Evaluation of The Confusion Matrix Method in The Validation of An Automated System for Measuring Feeding Behaviour of Cattle,” *Behavioural Processes,* vol. 148, no. 1, pp. 56-62, 2018.