Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 4 No. 2, December 2019, pp. 39 – 47

Diagnosis Of Heart Disease Using K-Nearest Neighbor Method Based On Forward Selection

Junta Zeniarja*[1], **Anisatawalanita Ukhifahdhina**[2], **Abu Salam**[3]
Faculty of Computer Science, Universitas Dian Nuswantoro
E-mail : junta@dsn.dinus.ac.id[1], anisafadhina@gmail.com[2], abu.salam@dsn.dinus.ac.id[3]*
**Corresponding author*

Abstract - Heart is one of the essential organs that assume a significant part in the human body. However, heart can also cause diseases that affect the death. World Health Organization (WHO) data from 2012 showed that all deaths from cardiovascular disease (vascular) 7.4 million (42.3%) were caused by heart disease. Increased cases of heart disease require a step as an early prevention and prevention efforts by making early diagnosis of heart disease. In this research will be done early diagnosis of heart disease by using data mining process in the form of classification. The algorithm used is K-Nearest Neighbor algorithm with Forward Selection method. The K-Nearest Neighbor algorithm is used for classification in order to obtain a decision result from the diagnosis of heart disease, while the forward selection is used as a feature selection whose purpose is to increase the accuracy value. Forward selection works by removing some attributes that are irrelevant to the classification process. In this research the result of accuracy of heart disease diagnosis with K-Nearest Neighbor algorithm is 73,44%, while result of K-Nearest Neighbor algorithm accuracy with feature selection method 78,66%. It is clear that the incorporation of the K-Nearest Neighbor algorithm with the forward selection method has improved the accuracy result.

Keywords - K-Nearest Neighbor, Classification, Heart Disease, Forward Selection, Data Mining

1. INTRODUCTION

Heart disease is a disorder that occurs in the large blood vessel system, causing heart and blood circulation to not work properly [1]. Data from the World Health Organization (WHO) in 2012 showed that 17.5 million people in the world died from cardiovascular disease or 31% of 56.5 million deaths worldwide and are expected to continue to increase to 23.3 million by 2030 [2], [3]. Of all deaths from cardiovascular disease (blood vessels) 7.4 million (42.3%) of which were caused by heart disease and 6.7 million (38.3%) caused by stroke. Whereas according to economic status, heart disease occurs most often at the lower economic level, which is around 2.1% and at the lower middle economic level, which is around 1.6% [2]. This proves that heart disease is the number one deadliest disease in various countries including Indonesia, because it has a high death rate. Based on these conditions, prevention and early treatment of heart disease is the most important thing to reduce mortality from heart disease.

One method that can be used in making early diagnosis of heart disease is to use data mining disciplines. There are several diagnostic classification studies on heart disease that have been carried out, including research conducted by Purushottam, Kanak Saxena and Richa

Sharma in a 2016 journal entitled "Heart Disease Prediction System Evaluation Using C4.5 Rules and Partial Tree" [4] who use C4.5 method and combined with Partial Tree which produces an accuracy of 70.93%. Then there was the research conducted by Febri Maspiyanti and Jullend Gatc in a 2015 journal entitled "Diagnosis of Heart Disease in Cellphones Using Decision Tree" [5], which uses the C4.5 method, which produces an accuracy of 81.29%. In addition to these methods, there are several studies that use other methods such as Naïve Bayes, Artificial Neural Network and K-Nearest Neighbor [1], [6], [7], [8], [9]. The advantages of the Naïve Bayes method are very simple, efficient, and has good performance with a lot of domain coverage [10]. It does not rule out the possibility of the Naive Bayes method that has weaknesses, namely in this method there are many gaps to reduce effectiveness, for example passing data into a particular class that is clearly passed data is not feasible to enter the class [11]. While the advantages possessed by the Artificial Neural Network method is that it can handle nonlinear data, the ability to tolerate noise in the system, and tend to produce low prediction errors, but the method of Artificial Neural Network has a disadvantage that requires a long processing time [12].

For the C4.5 method, the advantages possessed are being able to process data continuously with a faster process and produce decision trees that describe the rules so that they are easier to understand and implement, while the weaknesses that are owned are if the class and criteria are used too much there will be overlap which results in increasing time in making decisions and the amount of memory needed [5]. For the K-Nearest Neighbor method, some of the advantages possessed are the K-Nearest Neighbor method, which is a very simple method, strong for training noisy data and fast and effective processing time even with large training data [7], [13], [14]. However, the K-Nearest Neighbor method also has several disadvantages which include the need to determine the value of the parameter k (the number of closest neighbors), training based on unclear distance about what type of distance to use and which attributes to use for get the best results [7]. From the strengths and weaknesses of several methods mentioned, the K-Nearest Neighbor method can cover the weaknesses of the Naïve Bayes method, Artificial Neural Network and C4.5. Therefore, the present study will focus on the K-Nearest Neighbor method, which will be combined with the Feature Selection method, namely Forward Selection so that the obtained accuracy value is higher.

In this study, one type of feature selection method is Forward Selection that aims to improve the accuracy of the K-Nearest Neighbor algorithm by removing irrelevant attributes. The purpose of this study was to improve the performance of the K-Nearest Neighbor algorithm in diagnosing heart disease by adding the Forward Selection feature selection method. The benefit of this research is that people can diagnose heart disease easily and quickly, can be used by medical experts as a decision support system in diagnosing heart disease and the diagnosis can be used to treat and prevent early heart disease.

2. RESEARCH METHOD

Preprocessing of datasets is divided into data cleaning (removing code number and missing value), discretization and selection of forward selection features. The dataset will be selected using the Forward Selection method as a feature selection method for attributes that are less influential or irrelevant in the dataset with the aim of increasing accuracy. Then classification is done using the K-Nearest Neighbor algorithm as an algorithm that classifies the diagnosis of heart disease. The following are steps in applying the proposed method:

2.1. Pre-processing Data

Preprocessing data is the first step in diagnosing heart disease in this study. Selection of the forward selection feature enters the data processing stage in order to obtain the influential attributes in the classification process later.

2.2. Attribute Selection

One technique in preprocessing data is to maximize an algorithm's work by removing attributes that are not related in the classification process sequentially using Forward Selection.

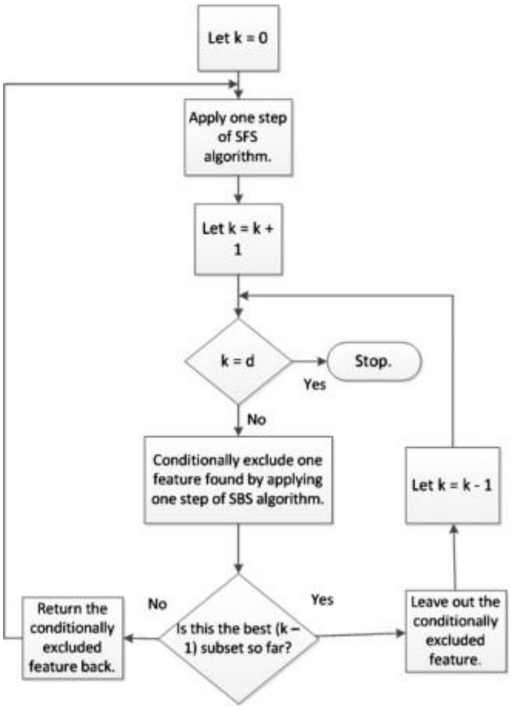
Let k = 0
Apply one step of SFS algorithm.
Let k = k + 1
k = d
Yes
Stop.
No
Conditionally exclude one feature found by applying one step of SBS algorithm.
Let k = k - 1
No
Yes
Is this the best (k – 1) subset so far?
Return the conditionally excluded feature back.
Leave out the conditionally excluded feature.

Figure 1. Flowchart of Forward Selection

Feature selection commonly referred to as feature selection is one of the important methods used in data preprocessing in data mining to optimize performance and speed up the process of an algorithm. The technique can choose a subset of features with a sufficiently large number, which leads to a reduction or removal of features that are less influential or irrelevant to classification. The main concept of feature selection is to choose a subset of existing features, because not all features have an effect or are relevant to the problem being raised. On the other side of some features that can cause interference and reduce the level of accuracy, therefore features that have no effect should be removed to increase the value of accuracy [15].

Feature selection, also known as feature, subset selection, attribute selection or variable selection, can also be interpreted as the process of selecting the right features to be used in the classification or clustering process. The purpose of this feature selection is to reduce the complexity of a classification algorithm, improve the accuracy of the classification algorithm, and be able to know the most influential features of the accuracy level. The forward selection method is modeling starting from the zero variable (empty model), then one by one the variables are entered until certain criteria are fulfilled. The steps of the forward selection method are as follows [16]:

a) Create a model by regressing the Y response variable with each predictor variable. Then the model that has the highest R2 value is selected. For example, the model is what makes the predictor X_a, namely:
$$\hat{Y} = b_0 + b_a X_a \tag{1}$$

b) Regressing the Y response variable, with predictor X_a, plus each predictor other than X_a and other predictors. Then the model that has the highest R2 value is chosen, for example containing an additional predictor X_b, that is the model as follows:
$$\hat{Y} = b_0 + b_a X_a + b_b X_b \tag{2}$$

c) The selected X_b predictor means having a higher expansion. The $F_{sequential}$ formula for X_b is as follows:
$$F_{seq} = R(\beta_b|\beta_{0,}\beta_a)/MSE/db \tag{3}$$

d) The $F_{sequential}$ value for X_b can also be obtained by squaring the value of the T test predictor X_b.

e) Sort the distance and determine which neighbor is closest based on the minimum distance k.

f) Determine the category of the nearest neighbor.

g) Use the majority category from the nearest neighbor as the new data predictive value.

2.3. *Proposed Method*

All of these stages are the processing stage, where the diagnosis of heart disease is done by classification using the K-Nearest Neighbor algorithm.

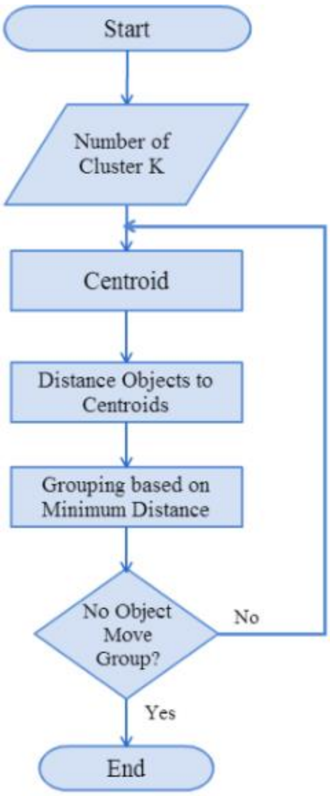
Start
Number of Cluster K
Centroid
Distance Objects to Centroids
Grouping based on Minimum Distance
No Object Move Group?
No
Yes
End

Figure 2. Flowchart of K-Nearest Neighbor Algorithm

K-Nearest Neighbor Algorithm is a method that is often used for the classification of text and data. This algorithm works by grouping new data based on the distance of the new data into some of the closest data / neighbors [17].
The steps taken in the K-Nearest Neighbor algorithm are as follows [7]:

a) Determine the parameter k = number of closest neighbors.
b) Calculate the distance between data that will be evaluated with training data.
c) Sort the distance and determine which neighbor is closest based on the minimum distance k.
d) Determine the category of the nearest neighbor.
e) Use the majority category from the nearest neighbor called the new data prediction value.

The distance search method used in this study is Euclidean Distance, which is the calculation of the closest distance. The closest distance calculation is needed to determine the number of similarities calculated from the similarity of the appearance of the text that belongs to a data. After that, the appearance of the text being tested is compared to each sample of the original data [18]. The following is a formula for determining the Euclidean Distance equation:

$$d_{(i,j)} = \sqrt{\left(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{in} - x_{jn}\right|^2\right)} \quad (4)$$

Notes :
d(i, j) : distance from i-data to j-data
xin : nth data in i-data

xjn : nth data in j-data

While for the formula of the K-Nearest Neighbor algorithm are as follows [7]:

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2} \quad (5)$$

Notes:
x1 : sample data
x2 : testing data
i : data variable
d : distance
p : data dimensions

3. RESULTS AND DISCUSSION

This study provides results in the form of accuracy obtained from the tests that have been carried out, with the aim of testing the accuracy and performance of the K-Nearest Neighbor algorithm based on the forward selection feature selection in classifying the diagnosis of heart disease. The forward selection method is used with the aim of removing attributes that have no effect in the classification process. To select attributes can be done by finding the correlation relationship closest to the target so that it can increase the accuracy of the classification using the K-Nearest Neighbor algorithm.

3.1. K-Nearest Neighbor Algorithm

The K-Nearest Neighbor algorithm in this study was used as a classification method in diagnosing heart disease. The dataset used is a valid dataset of 211 records with the initial number of attributes as many as 15 attributes and 1 label, because one attribute is primary key, in the calculation process the attribute is removed so that in calculating the attributes used are 14 attributes and 1 label.

To find out the performance of the K-Nearest Neighbor algorithm in classifying the diagnosis of heart disease, testing is done using confusion matrix so that the results of calculations as below are obtained:

Table 1. Confusion Matrix of K-Nearest Neighbor Algorithm

	TRUE POSITIVE	TRUE NEGATIVE	PRECISION CLASS
POSITIVE PREDICTION	145	41	77,96%
NEGATIVE PREDICTION	15	10	40,00%
RECALL CLASS	90,62%	19,61%	

The results of the confusion matrix table above can be calculated as follows:
Accuracy = (145 + 10) / (145 + 10 + 15 + 41) * 100%
= 0.7345971564 * 100%
= 73.44%

3.2. K-Nearest Neighbour Based On Forward Selection Algorithm

From a total of 211 data used, initialization of the attributes in the data will be carried out, the first attribute being X1 until the 14th attribute becomes X14.

Of all the attributes starting from X1 to X14 the calculation of correlation with the target variable (label / class) is initialized with Y, where the value of the label or class itself is assumed to be NEGATIVE = 0 and POSITIVE = 1. The correlation of the correlation coefficient in the regression is calculated by formula simple regression correlation like the following:

$$r = \frac{n\Sigma X_{1i}Y_i - (\Sigma_{X1i})(\Sigma Y_i)}{\sqrt{\{n\Sigma X_1^2 - (\Sigma X_i)^2\}\{n\Sigma Y_i^2 - (\Sigma Y_i)^2\}}} \qquad (6)$$

Calculations using a simple regression correlation formula are applied to all attributes, the point is to find the highest R value.

Then after getting the first correlation coefficient, then the correlation coefficient calculation between the selected attributes, Y and other attributes besides the selected attribute itself uses multiple linear regression with the following formula:

$$ry.x1.x2 = \sqrt{\frac{r_{yx1}^2 + r_{yx2}^2 - 2r_{yx1}r_{yx2}r_{x1x2}}{1 - r_{x1x2}^2}} \qquad (7)$$

Then the calculation is continued using multiple regression correlation formulas to obtain attributes that have a strong and influential correlation.

From the attribute attributes that have been selected, classification will be done using the K-Nearest Neighbor algorithm to determine the diagnosis of heart disease in patients with the decision results in the form of probabilities of the diagnostic classification itself. Based on the K value that has been set, namely 2, the value of the distance taken is the smallest 2.

3.3. Evaluation and Validation

This stage of validation and evaluation aims to test the algorithms that have been applied using the k-fold cross validation method, namely by forming k subset of the existing dataset. This validation method begins by dividing the data as much as n-fold as desired. In the initial process, the data will be divided into n data in the same proportion, then the training and testing process is carried out n times.

In this study, testing was carried out with a k value of 10 fold, where the aim was to determine the accuracy of the performance of the algorithm used, namely the forward selection K-Nearest Neighbor algorithm applied to the diagnosis of heart disease if tested using different data training and testing. The test uses the K-Fold Cross Validation method with a K value of 10 fold which is the best fold selection in the validity test [19]. The 10-Fold Cross Validation method will work by doing a test that is repeated as many K-fold, in this case, the K-fold is worth 10. The measurement results from the test repeated 10 times are in the form of an average accuracy value of 10 tests.

The initial stage in the 10-fold cross validation method is to divide the dataset, where in this study the initial data is 211 records, because the amount of data is odd, the data is rounded to 210 data, the remaining data will be added to testing data, then the data is divided into 10 subset of data (parts). Then each 1 subset contains 21 data for each iteration. The first fold contains combination data from 9 different sub-sets that have been combined and act as training data. While the remaining 1 subset is used for data testing, then the data training and testing, process is repeated until the 10th fold. The description of the distribution of training data and testing data using the 10-fold cross validation method is as follows:

Table 2. Distribution of Training Data and Testing Data

Fold	*Training Data*		*Testing Data*		Accuracy
	Subset	Data	Subset	Data	
1	$S_2,S_3,S_4,S_5,S_6,S_7,S_8,S_9,S_{10}$	189	S_1	22	66,55%
2	$S_1,S_3,S_4,S_5,S_6,S_7,S_8,S_9,S_{10}$	189	S_2	22	78,33%
3	$S_1,S_2,S_4,S_5,S_6,S_7,S_8,S_9,S_{10}$	189	S_3	22	79,88%
4	$S_1,S_2,S_3,S_5,S_6,S_7,S_8,S_9,S_{10}$	189	S_4	22	79,91%
5	$S_1,S_2,S_3,S_4,S_6,S_7,S_8,S_9,S_{10}$	189	S_5	22	78,86%
6	$S_1,S_2,S_3,S_4,S_5,S_7,S_8,S_9,S_{10}$	189	S_6	22	81,86%
7	$S_1,S_2,S_3,S_4,S_5,S_6,S_8,S_9,S_{10}$	189	S_7	22	80,41%
8	$S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_9,S_{10}$	189	S_8	22	82,08%
9	$S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_8,S_{10}$	189	S_9	22	79,91%
10	$S_1,S_2,S_3,S_4,S_5,S_6,S_7,S_8,S_9$	189	S_{10}	22	78,86%
Average of Accuracy Value					78,66%

Based on the table 2 above, it can be seen that the merger of the K-Nearest Neighbor algorithm and the forward selection method is good for use in the classification process because it produces an average accuracy value of 78.66% tested using 10-fold cross validation on 189 training data and 22 testing data.

3.4. Analysis and Discussion

Evident from the results of testing with the K-Nearest Neighbor algorithm, what was done on all data in the dataset were 211 accuracy records of 73.44%. While the accuracy of the experiments using the K-Nearest Neighbor algorithm with the forward selection method was 78.66%. This means that testing using the forward selection method can improve the results of accuracy. The results of increasing accuracy can be seen in the figure 3 below :

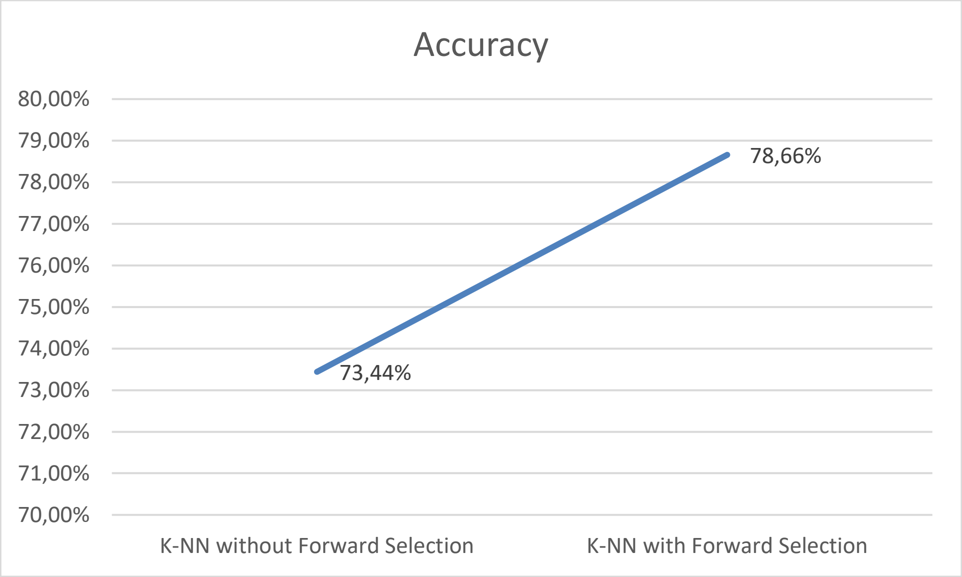

Figure 3. Comparison of Accuracy Results

4. CONCLUSION

In this study, conclusions were obtained based on the use of the K-Nearest Neighbor algorithm with the forward selection feature in the diagnosis of heart disease. The use of the forward selection method as proven feature selection to increase the accuracy value of the K-Nearest Neighbor algorithm, where there was a reduction in attributes from the number 14 to 8 attributes.

Then, the result of testing the K-Nearest Neighbor algorithm without using feature selection is 73.44%, while for the performance of the K-Nearest Neighbor algorithm using the forward selection method has increased to 78.66%. So, it is proven that the performance of the K-Nearest Neighbor algorithm with the forward selection method in the case of a diagnosis of heart disease is superior to just applying the K-Nearest Neighbor algorithm.

REFERENCES

[1] N. A. Widiastuti, S. Santosa, and C. Supriyanto, “Algoritma Klasifikasi Data Mining Naive Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung,” *J. Pseudocode*, vol. 1, no. 1, pp. 11–14, 2014.

[2] DEPKES, “Penyakit Jantung Penyebab Kematian Tertinggi, Kemenkes Ingatkan CERDIK,” *29 Juli 2017*, 2017. [Online]. Available: http://www.depkes.go.id/article/view/17073100005/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-.html. [Accessed: 14-Feb-2020].

[3] L. Ghani, M. D. Susilawati, and H. Novriani, “Faktor Risiko Dominan Penyakit Jantung Koroner di Indonesia,” *Bul. Penelit. Kesehat.*, vol. 44, no. 3, pp. 153–164, 2016.

[4] P. Sharma, K. Saxena, and R. Sharma, “Heart disease prediction system evaluation using C4.5 rules And Partial Tree,” *Adv. Intell. Syst. Comput.*, vol. 411, no. Cvd, pp. 285–294, 2016.

[5] F. Maspiyanti and J. Gatc, “Diagnosa Penyakit Jantung Pada Ponsel Menggunakan Pohon Keputusan,” *J. Teknol. Terpadu*, vol. 1, no. 1, pp. 13–20, 2015.

[6] B. Rifai, “Algoritma Neural Network Untuk Prediksi Penyakit Jantung,” *Techno Nusa Mandiri*, vol. IX, no. 1, pp. 1–9, 2013.

[7] M. Lestari, “Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk Mendeteksi Penyakit Jantung,” *Fakt. Exacta*, vol. 7, no. September 2010, pp. 366–371, 2014.

[8] R. Rakhmat Sani, J. Zeniarja, and A. Luthfiarta, “Penerapan Algoritma K-Nearest Neighbor pada Information Retrieval dalam Penentuan Topik Referensi Tugas Akhir,” *J. Appl. Intell. Syst.*, vol. 1, no. 2, pp. 123–133, 2016.

[9] R. R. Sani, J. Zeniarja, and A. Luthfiarta, “Pengembangan Aplikasi Penentuan Tema Tugas Akhir Berdasarkan Data Abstrak Menggunakan Algoritma K-Nearest Neighbor,” in *Proceeding SENDI_U*, 2016, vol. 2, pp. 103–111.

[10] L. D. Utami and R. S. Wahono, “Integrasi Metode Information Gain untuk Seleksi Fitur dan AdaBoost untuk Mengurangi Bias pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naive Bayes,” *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[11] S. Natalius, “Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen,” *J. Sist. Inf. Sekol. Tinggi Elektro dan Inform. Inst. Teknol. Bandung*, no. 3, pp. 1–5, 2011.

[12] E. Saleh, E. Noor, and T. Djatna, “Prediksi Masa Kedaluwarsa Wafer Dengan Artificial Neural Network (ANN) Berdasarkan Parameter Nilai Kapasitansi,” vol. 33, no. 4, pp. 450–457, 2013.

[13] J. Zeniarja, A. Luthfiarta, and C. Supriyanto, “Aplikasi Pencarian Perguruan Tinggi Dengan Algoritma K-Nearest Neighbor Berbasis Information Retrieval Dan Geographic Information System,” in *Prosiding SINTAK 2017*, 2017, pp. 137–146.

[14] A. Salam, J. Zeniarja, and R. S. U. Khasanah, “Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa,” in *Prosiding SINTAK*, 2018, pp. 480–486.

[15] S. Tabakhi, P. Moradi, and F. Akhlaghian, “An unsupervised feature selection algorithm based on ant colony optimization,” *Eng. Appl. Artif. Intell.*, vol. 32, pp. 112–123, 2014.

[16] W. Supriyanti, Kusrini, and A. Amborowati, “Perbandingan Kinerja Algoritma C4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa,” *J. Inf. Politek. Indonusa Surakarta*, vol. 1, no. 3, pp. 61–67, 2016.

[17] A. Rohman, “Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa,” *Neo Tek.*, vol. 1, no. 1, pp. 1–9, 2015.

[18] E. Hardiyanto and F. Rahutomo, “Studi Awal Klasifikasi Artikel Wikipedia Bahasa Indonesia Dengan Menggunakan Metoda K-Nearest Neighbor,” in *Seminar Nasional Terapan Riset Inovatif*, 2016, vol. 01, pp. 158–165.

[19] A. M. Zamani, B. Amaliah, and A. Munif, “Implementasi Algoritma Genetika pada Struktur Backpropagation Neural Network untuk Klasifikasi Kanker Payudara,” *J. Tek. ITS*, vol. 1, pp. 222–227, 2012.