Application of Discretization and Information Gain on Naïve Bayes to Diagnose Heart Disease

Taufik Fajar Mubaroq[1], Endang Sugiharti[1], Isa Akhlis[1]
[1] Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: fajartaufik005@gmail.com

ARTICLE INFO

Article history
Received 28 July 2019
Revised 18 September 2019
Accepted 2 October 2019

Keywords
Naïve Bayes
Discretization
Information gain
Data mining
Heart disease

ABSTRACT

In the health sector, there is a lot of data that can be processed and utilized. Current technology can be used to process data and produce predictions or diagnosis of disease. To diagnose the disease, it is necessary to have a patient medical record or health data which have collected in the past. In the process of processing the data requires a method that is called data mining. In data mining, some methods can be used for example classification. One of the algorithms found in the classification method is the Naïve Bayes algorithm. Naïve Bayes is an algorithm of classification method that is often used. The improvement of the accuracy of Naïve Bayes algorithms can be done by using discretization and information gain. The purpose of this study was to determine the application of discretization and information gain in heart disease datasets. The data used in this study are datasets of heart disease obtained from the UCI repository of machine learning consisting of 270 instances and 14 features. In this study, the mining process uses k-fold cross-validation with a value of k = 10. The results of the application of the Naïve Bayes algorithm classification obtained an accuracy of 85.1852% while the accuracy of the Naïve Bayes algorithm with discretization and information gain accuracy increased to 85.5556%. The enhancement of accuracy is obtained from the removal of scales performed using information gain and discretization techniques on Naïve Bayes algorithms with an increase of 0.3704% compared with the accuracy of the Naïve Bayes algorithm.

1. Introduction

Nowadays technology is developing very rapidly. Current technology can be used to predict or diagnose disease. To diagnose disease, it can use several classification algorithms, one of them is Naïve Bayes algorithm, Naïve Bayes is a classification algorithm in data mining for predictions with high accuracy and speed when applied to data, and can also be used to solve large dataset problems (Sudibyo, Astuti & Kurniawan, 2017). Data mining is the application of algorithms to extract patterns (models) from data sets (Mankar & Burange, 2014). Data mining also means the process of finding interesting patterns or information in selected data using certain techniques or methods (Muzakir & Wulandari, 2016). Collection of patient medical record data or health data that has been collected in the past can be used to diagnose a disease, one of which is heart disease. Heart disease is a disorder that occurs in the large blood vessel system, so it causes the heart and blood circulation to not function properly (Widiastuti, Santosa & Supriyanto, 2014). This study uses discretization and information gain on Naïve Bayes for the diagnosis of heart disease obtained from the UCI repository of machine learning datasets.

In the previous studies, the algorithm used to predict or diagnose disease was compared with the Particle Swarm Optimization (PSO) technique. The Particle Swarm Optimization (PSO) technique searches using the swarm of individuals (particles) to be updated from the iteration (Nurmalasari, Soesanto & Indriani, 2017). In addition, it can use the selection feature to compare the algorithms used. One of the feature selections is information gain (IG). IG is one of the algorithms used for attribute selection with a filtering method that uses entropy to determine the best attributes. The greater the IG value obtained from an attribute, the better for prediction. The use of IG will be used to evaluate the attributes in the data (Essra, Rahmadani & Safriadi, 2016). IG is defined as a measure of the effectiveness of an attribute in classifying data (Suyanto, 2017). In IG there is also a discretization technique for processing numerical data types. Discretization can reduce and simplify data from a database, and the use of discrete features is usually more concise and shorter than continuous data usage (Cheng, Chen & Wei, 2010).

The Naïve Bayes algorithm has been widely studied by many researchers. Widiastuti's et al., (2014) study used the Naïve Bayes algorithm and using particle swarm optimization (PSO) techniques to optimize accuracy using data from laboratory records. The study yielded an accuracy of 82.14%. Sabransyah, Nasution, and Amijaya (2017) compare the accuracy of the amount of data using the Naïve Bayes algorithm and the confusion matrix to determine the accuracy of the different data amounts. Research conducted by Agustin and Adi (2016) uses a Naïve Bayes algorithm with preprocessing data using k-fold validation to process training data before being applied to Naïve Bayes algorithms using tuberculosis disease data. Accuracy results obtained from Symptom, Lab, Symptom and Lab, Symptom and X-ray, Lab and X-ray, and Lab and X-ray Symptom methods respectively were 58.08%, 68.08%, 69.36%, 68.13%, 85.95%, 85.21%. The study conducted by Naufal, Wahono and Syukur (2015), which explains the comparison of the accuracy of the support vector machine algorithm using bootstrapping and information gain to determine predictions of employee loyalty.

The purpose of this study was to determine the application of discretization and IG on Naïve Bayes algorithms in diagnosing heart disease and knowing the accuracy of Naïve Bayes algorithms after and before applying discretization and IG in diagnosing heart disease.

2. Methods

2.1 Discretization

The application of discretization in the data mining pre-processing process as follows:

1. Analyze attributes from the heart disease dataset by grouping attributes of numerical type and attributes of nominal type. The application of discretization only applies to attributes of numerical type.
2. Look for entropy values from heart disease data or commonly formulated with entropy (S).

$$Entropy(S) = \sum_{i}^{c} - p_i log_2 p_i \quad (1)$$

In which:
c : number of values contained in the attribute
i : class
pi : ratio between the number of samples in class i and the number of all samples

Calculate the entropy value of the attributes of the numerical type, usually formulated with entropy (Sv).

$$Entropy(S_v) = \sum_{i}^{c} - p_v log_2 p_v \quad (2)$$

3. Look for the gain value of each data in the numeric attribute.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in values\,(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3)$$

In which:

A	: attribute
V	: value for attribute A
$Values(A)$	: set of values for attribute A
$\| Sv \|$	: number of samples for the value of v

$|S|$: number of all data samples
Entropy (*Sv*): entropy for the sample of value v

4. Look for the highest gain value after getting the gain value from each data. After getting the highest gain value from the numeric attribute, make the process on each numeric type attribute.

2.2 Feature Selection Information Gain

The following is the application of the feature selection information gain.

1. Analyze attributes from the heart disease dataset by classifying attributes of the nominal type.
2. Calculate the amount of data in the attribute according to class and character. As with gender attributes with two characters, male and female, count how many are included in the present and absent classes.
3. Look for entropy values from heart disease data or commonly formulated with entropy (S).

$$Entropy(S) = \sum_i^c - p_i log_2 p_i \tag{4}$$

In which:
c : number of values contained in the attribute
i : class
p_i : ratio between the number of samples in class i and the number of all samples

4. Calculate the entropy value of each character in each attribute, usually formulated with entropy (Sv).

$$Entropy(S_v) = \sum_i^c - p_v log_2 p_v \tag{5}$$

5. Calculate the gain value on attributes of nominal type.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in values\ (A)} \frac{|S_v|}{|S|}\ Entropy(S_v) \tag{6}$$

In which:
A : attribute
V : value for attribute A
$Values(A)$: set of values for attribute A
$|S_v|$: number of samples for the value of v
$|S|$: number of all data samples
$Entropy(S_v)$: entropy for sample value v

6. Sort all gain values obtained on all attributes from the smallest to the largest.
7. Eliminate the attribute with the smallest gain= value, then to the Naïve Bayes algorithm to calculate the value of accuracy.
8. Do it until you get a higher accuracy from the calculation process that only uses Naïve Bayes algorithms.

2.3 Application of the Naïve Bayes Algorithm

The following is the application of the research Naïve Bayes algorithm.

1. Prepare heart disease training data.
2. Pre-processing the data using the discretization and information gain method.
3. Look for the gain value of each attribute and remove the attribute with the lowest gain value.
4. Distribution of datasets into training data and test data using k-fold cross-validation.
5. Classifying using Naïve Bayes based on k-fold cross-validation.

6. In the Naïve Bayes classification, there are two ways to calculate data attributes, namely data with numeric and nominal types. In data attributes of nominal type, Naïve Bayes uses the equation:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (7)$$

In which:
X : data or tuple object (class C)
H : hypothesis
P(H|X) : probability that hypothesis H is in condition
P(H) : prior probability that the H hypothesis is valid (true)
P(X) : prior probability of tuple X

7. In data attributes of numeric type, use the equation:

$$P(X_k|C_i) = \frac{1}{\sigma_{C_i}\sqrt{2\pi}} e^{-\frac{(x-\mu_{C_i})}{2\sigma_{C_i}{}^2}} \qquad (8)$$

8. Calculate the accuracy using confusion matrix, which is done by classification in two classes, where the values of the two classes are 0 and 1 as shown in Table 1. If f_{ij} denotes the number of records from class i, the prediction results go to class j during testing. For example, cell f11 is the amount of data in class 1 that is correctly mapped to class 1, and f10 is data in class 1 which is mapped incorrectly to class 0.

Table 1. Confusion matrix for two classes classification

	f_{ij}	Prediction Class (j)	
		Class = 1	Class = 0
Real Class (*i*)	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

The flowchart of the Naïve Bayes algorithm classification is shown in Figure 1. While Figure 2 shows the flowchart of the Naïve Bayes algorithm classification by applying discretization and information gain.

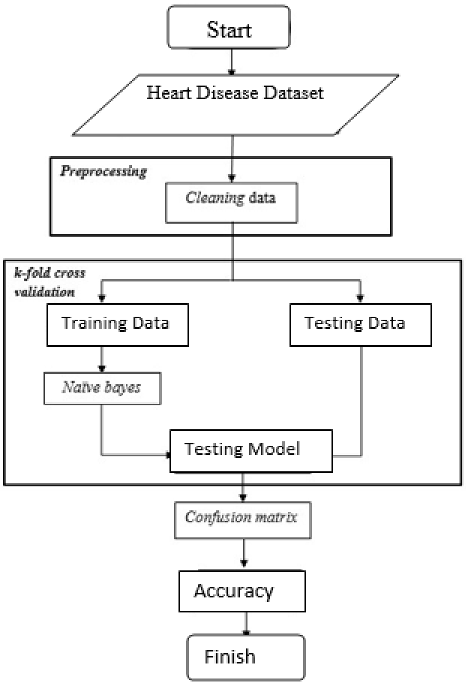


Figure 1. The flowchart of the Naïve Bayes algorithm

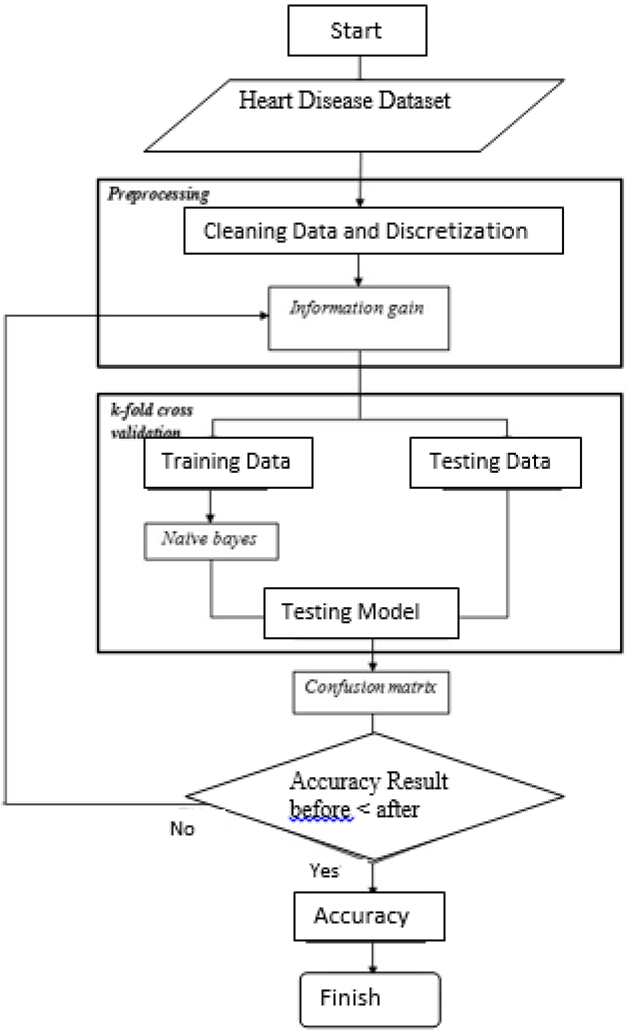
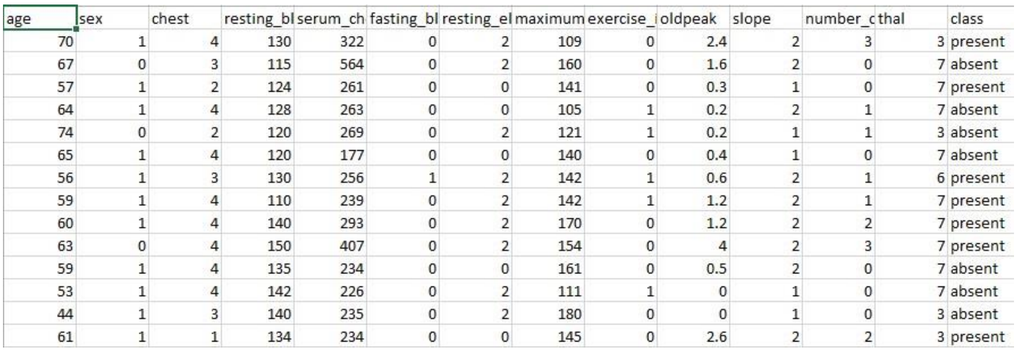


Figure 2. The flowchart of the Naïve Bayes algorithm classification by applying discretization and information gain

3. Results and Discussion

3.1 Cleaning Data

Cleaning data is checked on the dataset, if there is a missing value in the dataset, treatment must be given to the data. In the dataset used for this study, there are no missing values as shown in Figure 3 which show a dataset of heart disease, because there are no missing values then enter the next stage.



age	sex	chest	resting_bl	serum_ch	fasting_bl	resting_el	maximum	exercise_i	oldpeak	slope	number_c	thal	class
70	1	4	130	322	0	2	109	0	2.4	2	3	3	present
67	0	3	115	564	0	2	160	0	1.6	2	0	7	absent
57	1	2	124	261	0	0	141	0	0.3	1	0	7	present
64	1	4	128	263	0	0	105	1	0.2	2	1	7	absent
74	0	2	120	269	0	2	121	1	0.2	1	1	3	absent
65	1	4	120	177	0	0	140	0	0.4	1	0	7	absent
56	1	3	130	256	1	2	142	1	0.6	2	1	6	present
59	1	4	110	239	0	2	142	1	1.2	2	1	7	present
60	1	4	140	293	0	2	170	0	1.2	2	2	7	present
63	0	4	150	407	0	2	154	0	4	2	3	7	present
59	1	4	135	234	0	0	161	0	0.5	2	0	7	absent
53	1	4	142	226	0	2	111	1	0	1	0	7	absent
44	1	3	140	235	0	2	180	0	0	1	0	3	absent
61	1	1	134	234	0	0	145	0	2.6	2	2	3	present

Figure 3. The dataset in Excel format

3.2 Feature Selection of Information Gain

In the process of information gain, if there are attributes with a numerical type, the processing is done by discretization technique.

The process gets the entropy value for each attribute, by finding the entropy value for the entire data (class = "1" and "0") first. After getting the entropy, class value calculate the entropy value and gain value on each attribute. Information gain will be removed from the data attribute to optimize calculations that will affect the accuracy of the Naïve Bayes method. Removal of attributes is done one by one from attributes that have the smallest gain. The following are the results of the gain values of each attribute in the heart disease data after being sorted from the lowest to highest gain values shown in Table 2 as follows.

Table 2. The Results of the gain values of each attributes heart disease

No.	Attribute	Gain Value
1.	Fasting blood sugar	0.0002
2.	Resting blood pressure	0.0164
3.	Resting electrocardiographic results	0.0242
4.	Serum cholesterol	0.0267
5.	Age	0.0567
6.	Gender	0.0669
7.	Slope	0.1112
8.	Oldpeak	0.1196
9.	Maximum heart rate achieved	0.1203
10.	Exercise-induced angina	0.1299
11.	Number of major vessels	0.1752
12.	Chest	0.1922
13.	Thal	0.2086

3.3 Results of Algorithm Implementation

3.3.1 Stage of Naïve Bayes classification

The distribution of training data and data testing on the Naïve Bayes algorithm classification uses k-fold cross-validation. After getting the Naïve Bayes algorithm classification model, calculate the accuracy using a confusion matrix. Naïve Bayes Classification algorithm will produce better results if using more training data (Sugiharti, Firmansyah & Devi, 2017). The results of the accuracy in the Naïve Bayes classification are shown in Table 3.

Tabel 3. The result of the accuracy *Naïve Bayes*

Method	Accuracy
Naïve Bayes	85.1852 %

3.3.2 Classification of Naïve Bayes with discretization and information gain

The Naïve Bayes classification with information gain starts from eliminating attributes that have the smallest gain value and the process stops when the accuracy results are greater than the results of the accuracy in the Naïve Bayes classification. The accuracy results generated from the Naïve Bayes method by applying discretization and information gain as a feature selection are presented in Table 4.

Table 4. The results of eliminating attribute and accuracy method Naïve Bayes with discretization and information gain

Eliminating Attribute	Attribute	Accuracy
6	*Fasting blood sugar*	84.8148 %
4	*Resting blood pressure*	84.0741 %
7	*Resting electrocardiographic result*	84.4444 %
5	*Age*	84.0741 %
1	*Gender*	85.5556 %

The application of the Naïve Bayes method with discretization and information gain gets the highest accuracy results on eliminating attribute 1 which is 85.5556%. From the results of accuracy 85.1852% gain increasing by 0.3704% compared with the results of the accuracy of the algorithm Naïve Bayes. The result of the comparison of the accuracy Naïve Bayes classification algorithm testing using k-fold cross-validation is shown in Table 5.

Table 5. The comparison of the accuracy of Naïve Bayes method

Naïve Bayes	Naïve Bayes with discretization and Information Gain
85.1852%	85.5556%

4. Conclusion

The application of discretization and IG on Naïve Bayes to diagnose heart disease will be carried out at the pre-processing stage, they are the cleaning data stage and the feature selection stage. In the phase of cleaning data, check on the dataset, if there is a missing value in the dataset, treatment of the data must be given. Missing values can be filled manually, and the missing value can also be filled based on the average model by replacing empty values with an average value based on the available values for the feature. In the dataset used for this study, there were no missing values, then they continued the next stage. In the feature selection stage, eliminating attributes from the data will be carried out to optimize calculations that will affect the accuracy of the Naïve Bayes method.

Eliminating is performed using information gain and discretization techniques. Furthermore, the data is classified by the Naïve Bayes method, wherein this process the eliminating of attributes is done one by one from attributes that have the smallest information gain value and will be classified. The process of eliminating attributes will stop when the results of accuracy are greater than the results of accuracy using only Naïve Bayes algorithms and that accuracy decreases after the eliminating of the next attribute.

The application of discretization and information gain on Naïve Bayes has an increase in accuracy of 0.3704% from 85.1852% to 85.5556% with the elimination of five attributes because this process will stop when the accuracy results are greater than accuracy which only uses Naïve Bayes methods and decreases after eliminating of the next attribute. Based on the results in this study it can be concluded that by applying discretization and information gain the accuracy of the Naïve Bayes algorithm is better than just using a Naïve Bayes algorithm in diagnosing heart disease.

References

Cheng, C. H., Chen, T. L., & Wei, L. Y. (2010). A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Information Sciences*, 180(9), 1610-1629. doi:10.1016/j.ins.2010.01.014

Essra, A., Rahmadani, & Safriadi. (2016). Analisis information gain attribute evaluation untuk klasifikasi serangan intrusi [Analysis of attribute information obtained evaluation for the classification of intrusion attacks]. *Information System Development*, 2(2), 9-14.

Mankar, A. B., & Burange, M. S. (2014). Data mining - An evolutionary view of agriculture. *International Journal of Application or Innovation in Engineering & Management*, 3(3), 102-105.

Muzakir, A., & Wulandari, R. A. (2016). Model data mining sebagai prediksi penyakit hipertensi kehamilan dengan teknik decision tree [Data mining model as a prediction of pregnancy hypertension with a decision tree technique]. *Scientific Journal of Informatics,* 3(1), 19-26. doi:10.15294/sji.v3i1.4610

Naufal, A. R., Wahono, R. S., & Syukur, A. (2015). Penerapan bootstrapping untuk ketidakseimbangan kelas dan weighted information gain untuk feature selection pada algoritma support vector machine untuk prediksi loyalitas pelanggan [Application of bootstrapping for class imbalance and weighted information gain algorithm for feature selection in support vector machine to predict customer loyalty]. *Journal of Intelligent Systems*, 1(2), 98-108.

Nurmalasari, E., Soesanto, O., & Indriani, F. (2017). Algoritma particle swarm optimization (pso) untuk optimasi nilai center radial basis probabilistic neural network (rbpnn) pada klasifikasi data breast cancer (Particle swarm optimization (pso) algorithm for optimizing the center radial value of probabilistic neural network (rbpnn) base on the classification of breast cancer data). *Jurnal Elektronik Nasional Teknologi dan Ilmu Komputer*, 1(1), 137-150.

Sabransyah, M., Nasution, Y. N., & Amijaya, F. D. (2017). Aplikasi metode Naive Bayes dalam prediksi risiko penyakit jantung [Application of the Naive Bayes method in predicting the risk of heart disease]. *Jurnal Eksponensial*, 8(2), 111-118. Retrieved from http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/31

Sudibyo, U., Astuti, Y. P., & Kurniawan, A. W. (2017). High School Major Classification towards University Students Variable of Score Using Naïve Bayes Algorithm. *Scientific Journal of Informatics,* 4(2), 193-194.

Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive evaluation of performance of computer science students of UNNES using data mining based on Naïve Bayes Classifier (NBC). *Journal of Theoretical and Applied Information Technology*, 95(4), 902–911. Retrieved from https://lib.unnes.ac.id/33089/

Suyanto. 2017. *Data mining untuk klasifikasi dan klasterisasi data* [Data mining for data classification and classification]. Bandung: Informatika.

Widiastuti, N. A., Santosa, S., & Supriyanto, C. (2014). algoritma klasifikasi data mining Naive Bayes berbasis Particle Swarm Optimization untuk deteksi penyakit jantung [Naive Bayes data mining classification algorithm based on particle swarm optimization for heart disease detection]. *Pseudocode*, 1(1), 11-14. doi:10.33369/pseudocode.1.1.11-14