ILKOM Jurnal Ilmiah Vol. 14, No. 2, August 2022, pp. 120-125
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification

Pungkas Subarkah [a,1,*]; Wenti Risma Damayanti [a,2]; Reza Aditya Permana [a,3]
[a] *Universitas Amikom Purwokerto, Jl. Letjend Pol. Soemarto, Purwokerto Utara and 53127, Indonesia*
[1] *subarkah@amikompurwokerto.ac.id;* [2] *wrismadamayanti@gmail.com;* [3] *rezapermana127@gmail.com*
** Corresponding author*

Article history: Received June 03, 2022; Revised June 20, 2022; Accepted August 01, 2022; Available online August 30, 2022

Abstract

Heart failure (ARF) is a health problem that has relatively high mortality and morbidity rates in developed or developing countries, including Indonesia. In 2016, WHO stated that 17.5 million people died from cardiovascular disease, while in 2008, HF disease represented 31% of patient deaths worldwide. One of the new breakthroughs for early diagnosis was by utilizing data mining techniques. In this study, the Correlated Naive Bayes Classifier (C-NBC) and Naive Bayes Classifier (NBC) algorithms were used to obtaine the best accuracy results so that they can be used for the Heart Failure dataset. Based on the results of the tests that had been carried out, it showed that the Correlated Naive Bayes Classifier (C-NBC) algorithm accuracy of 80.6% was higher accuracy than the Naive Bayes Classifier (NBC) algorithm which was only 67.5%. With the results of this study, the use of the Correlated Naive Bayes Classifier (C-NBC) algorithm can be used to diagnose patients with potential heart failure (heart failure) because it has a high level of accuracy and is categorized as Good Classification.

Keywords: Classification; Correlated Naive Bayes Classifier; Naive Bayes Classifier; Heart failure.

Introduction

Heart failure (HF) is a major health issue with relatively high mortality and morbidity rates in developed or developing countries, including Indonesia [1]. In addition, HF is the leading cause of hospital admission, especially among the elderly[2], [3]. Based on the diagnosis results, coronary heart disease and heart failure are relatively common in the population aged 15-24 years. Death due to cardiovascular disease, especially heart failure, is 27%. About 3-20 out of 1000 people experience this disease. The incidence of heart failure increases with age (100 for every 1000 people over the age of 60 years)[4].

Based on the results of the Health Research of the Ministry of Health of the Republic of Indonesia in 2013, it was found that the estimation of heart failure patients based on a doctor's diagnosis was estimated at 0.13% or 229,696 people, with the most estimated sufferers were coming from East Java Province at 0.19% or around 54,826 people. Meanwhile, based on the diagnosis of symptoms, it is estimated at 0.3% or 530,068 people, with the highest estimated number of sufferers were coming from West Java Province at 0.3% or 96,487 people[5].

Given the number of patients with heart failure and the importance of a vital organ such as the heart, predicting heart failure has become a priority for doctors. Predicting heart failure-related issues in practice have usually failed to achieve high accuracy[6]. Rapid and accurate prediction of death for people with heart failure is very important to improve patient health care and prevent them from dying early[7]. Therefore we need a new breakthrough that is right processing for the such data in order to produce a health plan that can prioritize disease management approaches to reduce the mortality rate of heart failure patients.

Several studies have been conducted related to this research. First, the study of tuberculosis disease using the Naive Bayes algorithm based on particle swarm optimization. The result obtained in this study indicates that the accuracy value of the Naive Bayes algorithm is 92.69% [8]. Second, the research uses several datasets, including the Balance-Scale Dataset, the Iris Dataset, the Haberman Dataset, and the Servo Dataset, using the Correlated Naive Bayes Classifier (C-NBC) algorithm. Using the Correlated Naive Bayes Classifier (C-NBC) algorithm, the result shows the increase in the accuracy value that is 13.3%[9]. In addition, the study used a dataset of Covid-19 patients. This study uses various Naive Bayes techniques. The result shows that an average accuracy value was 87% [10].

http://dx.doi.org/10.33096/ilkom.v14i2.1148.120-125

Based on the description of the problems that have been described, to reduce the number of deaths caused by heart failure is to make an early diagnosis correctly. One method that can be used is to use data mining techniques. To diagnose heart failure, a form or procedure that has the best level of accuracy is required, so in this study, a comparison of several data mining classification methods is required, namely the Correlated-Naive Bayes Classifier and Naive Bayes Classifier algorithms, to obtain the best accuracy so that they can be used for the diagnosis of potential heart failure Effectively and optimally.

Method

The method proposed by researchers in research on the Comparison of Accuracy of Correlated Naive Bayes Classifier and Naive Bayes Classifier Algorithms for Classification of Heart Failure Disease from several stages can be seen in **Figure 1**.

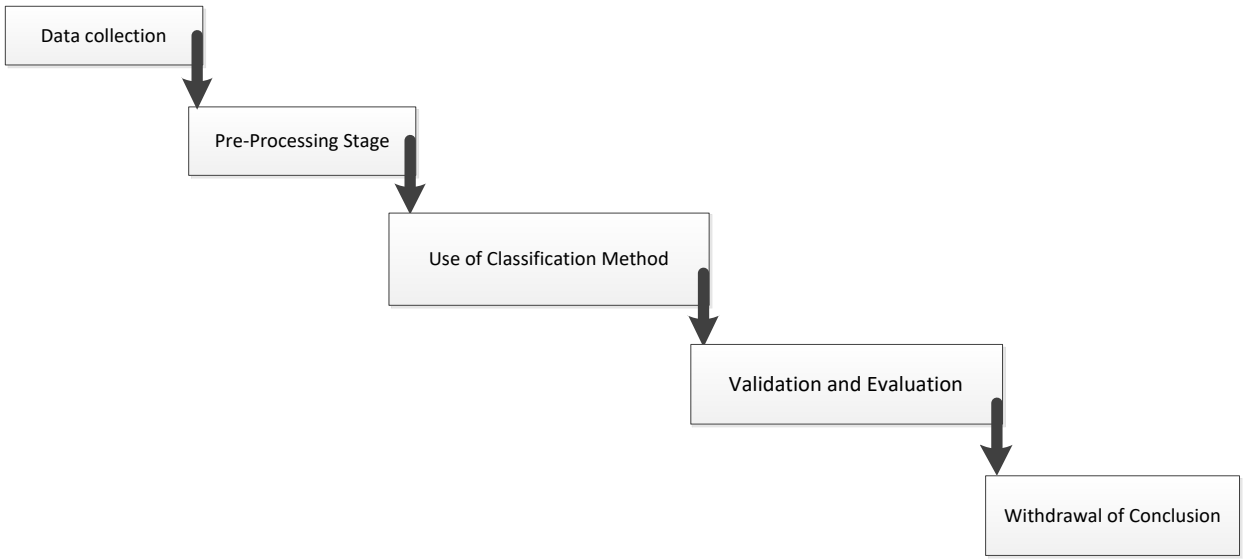

Figure 1. Research Stages

Finally, Explanation of the research stages based on the figure 1., above are as follows:

1. Data Collection

   The dataset in this study was collected from the Krembil Research Institute, Toronto, Canada, dataset taken from the UCI Repository. The dataset has 13 attributes, two classes, and 299 data see **Table 1**[11].

Table 1. Heart Failure Dataset Attributes

No	Attribute Name	Description
1	Age	Patient Age
2	Anaemia	Decrease in red blood cells or hemoglobin
3	Creatine phosphokinase (CPK)	CPK enzyme level in blood
4	Diabetes	Have a history of diabetes
5	Ejection Fraction	Percentage of blood leaving the heart with each contraction
6	High Blood Pressure	Have hypertension
7	Platetets	Platelets in the blood
8	Serum Creatine	Creatinine level in blood
9	Serum Sodium	Sodium level in blood
10	Sex	Gender
11	Smoking	Smoke
12	Time	Patient follow-up period
13	Date Event	If the patient dies during follow-up

2. Pre-Processig Stage

   The pre-processing stage is a data mining process that is first carried out to get good quality data to be processed before the classification process, one of which is data transformation. Data transformation is an essential part of the health dataset in the pre-processing stage[12]

3. Use of Classification Method

   Stages of using the classification method. One of the tasks of data mining is classification. Classification represents the most widely used data mining technique[13]. In addition, the classification can be improved by

increasing the number of features in data mining[14]. This study uses the Correlated Naive Bayes classifier (C-NBC) and Naive Bayes Classifier (NBC) algorithms. The results of the algorithm was used to support the level of accuracy using the Confusion Matrix. The confusion matrix was obtained by calculating the value of precision, recall, and F-Measure[12], can be seen in **Table 2**.

Table 2. Confusion Matrix

Correct Classification	Clasification as	
	+	-
+	TP	FN
-	FP	TN

Details of the general calculation of the value of precision, recall, and F-Measure of an accuracy[13], :

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F\text{-}Measure = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

4. Validation and Evaluation

   The validation and evaluation stages were carried out by measuring the accuracy of the results achieved by the confusion matrix and K-fold cross-validation technique models.

5. Withdrawal of Conclusion

   This stage concluded the results obtained from the study using the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier algorithm, which provided accurate results for classifying heart failure based on the precision, recall, and F-Measure values of each algorithm, with the classification level[15], as follows :

   a. Excellent classification = 0.90 – 1.00
   b. Good classification = 0.80 – 0.90
   c. Fair classification = 0.70 – 0.80
   d. Poor classification = 0.60 – 0.70
   e. Failure = 0.50 – 0.60

Results and Discussion

A. Pre-Processing Stage

The pre-processing stage was through data transformation, which aimed to facilitate the calculation of the value between the class attributes at the classification stage. The following results from the change of non-numeric data types into numeric data can be seen in **Table 3**.

Table 3. Data transformation of heart failure dataset class

No.	Non-Numeric Data	Numerical Data
1.	*Death*	1
2.	*Survived*	0

Then at this pre-processing stage, identification and adjustment of attributes and selection of the heart failure dataset were carried out so that the data obtained was data that was really ready to be used in the next stage. The results of the heart failure dataset that have been adjusted for attributes for the weka application as shown in **Table 4**.

Table 4. Data Pre-processing

Original Data	Pre-processing Result Data	Information
45	45	Patient Age
0	0	Decrease in red blood cells or hemoglobin
2413	2413	CPK enzyme level in blood
0	0	Have a history of diabetes
38	38	Percentage of blood leaving the heart with each contraction

Original Data	Pre-processing Result Data	Information
0	0	Have hypertension
140000	140000	Platelets in the blood
1.4	1.4	Creatinine level in blood
140	140	Sodium level in blood
1	1	Gender
1	1	Smoke
280	280	Patient follow-up period
0	0	If the patient dies during follow-up

B. Use of Classification Method

After going through the pre-processing stage, then the stage was continued to the stage of using the classification method. At this stage, the aim was to produce an accuracy value from the confusion matrix. The testing technique used the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier (NBC) algorithm on the dataset with an evaluation method of 10-fold cross-validation with randomization of 20 times. The results of the algorithm for the accuracy of the heart failure dataset.

Table 5. Heart Failure Dataset Accuracy Results

No	C-NBC	NBC
1	81.94	76.58
2	81.94	76.58
3	79.6	78.26
4	79.9	76.92
5	80.27	77.26
6	80.60	76.58
7	78.93	75.92
8	80.94	76.25
9	80.94	76.26
10	80.60	76.9
11	78.60	76.59
12	81.60	76.59
13	81.27	75.58
14	79.93	75.58
15	79.60	75.58
16	79.93	75.92
17	80.27	76.92
18	80.60	76.59
19	80.66	76.59
20	79.59	76.59

In order to make it easier to see the results of the comparison of the accuracy of the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier (NBC) algorithm to the heart failure dataset based on the test results, you can see the visualization of **Figure 2**., below.

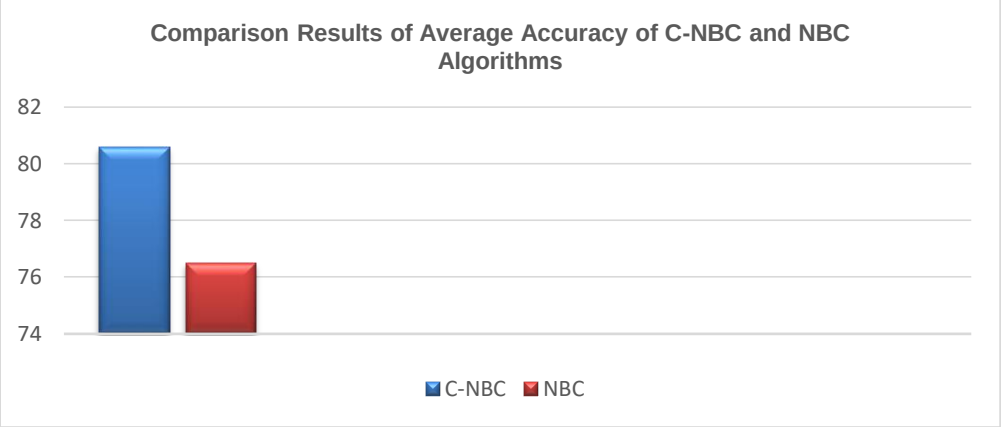

Figure 2. Comparison of Average Accuracy of C-NBC and NBC Algorithms

Based on **Figure 2** above, the highest average accuracy value in the heart failure dataset with the Correlated Naive Bayes Classifier (C-NBC) algorithm was 80.6% whereas the accuracy value of the Naive Bayes Classifier (NBC)

algorithm was 76.5%. The highest results obtained in the second test of these algorithms obtained the accuracy value of the Correlated Naive Bayes Classifier (C-NBC) algorithm compared to the Naive Bayes Classifier (NBC) algorithm because the Correlated Naive Bayes Classifier (C-NBC) algorithm took into account the R-Square (correlation value) of each dataset attribute to its class. This was in line with the fact that the addition of parameters to the correlation calculation in the Naive Bayes Classifier algorithm can increase the accuracy value optimally[16].

C. Validation and Evaluation

In this stage, validation and evaluation of the algorithms used were the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier (NBC) algorithm using a confusion matrix and 10-Fold cross-validation. After testing the classification method, the results of the average accuracy of the second algorithm were shown in the **Table 6**.

Table 6. Correlated Naive Bayes Classifier (C-NBC) Algorithm Accuracy Results and Correlated Naive Bayes Classifier (C-NBC) Algorithm

No	Algorithm	Accuracy
1	Algorithm (C-NBC)	80.6%
2	algorithm (NBC)	76.5%

Furthermore, from the accuracy described in **Table 6**., it can be seen that the accuracy results were generated by the confusion matrix from testing the heart failure dataset using the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier (NBC) algorithm, as shown in **Table 7** and **Table 8**.

Table 7. Confusion Matrix Correlated Naive Bayes Classifier (C-NBC) Algorithm

	0 (*Death*)	**1 (*Survived*)**
0 (*Death*)	189	14
1 (*Survived*)	40	56
299	229	70

While in **Table 8** below is the result of the confusion matrix algorithm Naive Bayes Classifier (NBC).

Table 8. Confusion Matrix Algorithm Naive Bayes Classifier (NBC)

	0 (*Death*)	**1 (*Survived*)**
0 (*Death*)	185	18
1 (*Survived*)	52	44
299	237	62

Conclusion

From the results of research that have been carried out related to the comparison of the accuracy of the Correlated Naive Bayes Classifier (C-NBC) algorithm and the Naive Bayes Classifier (NBC) algorithm on the heart failure dataset obtained from UCI Repository Learning, namely the Heart Failure Dataset. The Correlated Naive Bayes Classifier (C-NBC) obtained the best accuracy value, compared to the results of the Naive Bayes Classifier (NBC) algorithm. The accuracy rate of the Correlated Naive Bayes Classifier (C-NBC) algorithm was 80.6%. With the results of this study, the use of the Correlated Naive Bayes Classifier (C-NBC) algorithm can be used for the diagnosis of heart failure patients because it has a good level of accuracy and is categorized as Good Classification.

Acknowledgement

The researcher would like to thank the Research and Community Service Institute (LPPM) of Amikom University, Purwokerto, for the support given to researchers through the 2022 Amikom Young Lecturer Research (PDMA), both material and non-material. So that this research goes well and is completed on time.

References

[1] B. B. Siswanto *et al., Pedoman Tatalaksana Gagal Jantung (Edisi Pertama)*. PERKI, 2015.

[2] N. Azad and G. Lemay, "Management of chronic heart failure in the older population," *J. Geriatr. Cardiol.*, vol. 11, no. 4, 2014.

[3] E. Ammenwerth *et al.*, "HerzMobil, an integrated and collaborative telemonitoring-based disease management program for patients with heart failure: A feasibility study paving the way to routine care,"

JMIR Cardio, vol. 2, no. 1, pp. 1–12, 2018.

[4] D. P. T. Astuti and I. K. Suardamana, *Gagal jantung*. Universitas Udayana, 2017.

[5] R. Fikriana, *Sistem Kardiovaskuler (Cetakan 1)*. deepublish, 2018.

[6] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone.," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020.

[7] Z. Wang, Y. Zhu, D. Li, Y. Yin, and J. Zhang, "Feature rearrangement based deep learning system for predicting heart failure mortality," *Comput. Methods Programs Biomed.*, vol. 191, p. 105383, 2020.

[8] E. Mutiara, "Algoritma klasifikasi Naive Bayes berbasis Particle swarm optimization untuk prediksi penyakit Tuberculosis (TB).," vol. 8, no. 1, pp. 46–5, 2020.

[9] S. M. Abd-elsalam, M. M. Ezz, S. Gamalel-din, and G. Esmat, "Early diagnosis of esophageal varices using Boosted-Naïve Bayes Tree: A multicenter cross-sectional study on chronic hepatitis C patients," *Informatics Med. Unlocked*, vol. 20, p. 100421, 2020.

[10] A. H. Rabie, N. A. Mansour, A. I. Saleh, and A. E. Takieldeen, "Expecting Individuals' Body Reaction to Covid-19 Based On Statistical Naïve Bayes Technique," *Pattern Recognit.*, vol. 128, p. 108693, 2022.

[11] D. Chicco, "Dataset Heart Failure (Gagal Jantung)," *Machine Learning Repository*, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. [Accessed: 02-Nov-2021].

[12] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Third Edit., vol. 5. USA: Elsevier, 2012.

[13] M. Han, J., & Kamber, *Data Mining Concepts, Model and Techniques 2nd Edition*. San Fransisco: Elsevier, 2006.

[14] A. Urso, A. Fiannaca, M. La Rosa, V. Ravì, and R. Rizzo, *Data Mining : Classification and Prediction*. 2018.

[15] F. Gorunescu, *Data mining Concepts, Models and Techniques*. Verlen Berlin: Springer, 2011.

[16] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Pembobotan korelasi pada Naive Bayes Classifier," in *Semin. Nas. Teknol. Inf. dan Multimed.*, 2015, pp. 43–47.