p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
11(2) 2023: 44-50

Jurnal Statistika
Universitas Muhammadiyah Semarang

OPTIMIZATION OF NAÏVE BAYES USING BACKWARD ELIMINATION FOR HEART DISEASE DETECTION

Saeful Amri[1*], Ariska Fitriyana Ningrum[2] , Prizka Rismawati Arum[3]
[1,2] Departement of Data Science, Faculty of Science and Agriculture Technology, University of Muhammadiyah Semarang
[3] Departement of Statistics, Faculty of Science and Agriculture Technology, University of Muhammadiyah Semarang

***e-mail**: saefulamri@unimus.ac.id

Article Info:
Received: October 9, 2023
Accepted: November 27, 2023
Available Online: November 30, 2023

Keywords:
Heart Disease Detection; Naïve Bayes; Backward Elimination
.

Abstract: Heart disease is the main cause of death in humans. Even though preventive measures have been taken such as regulating food (diet), lowering cholesterol, and treating weight, diabetes, and hypertension, heart disease remains a major health problem. There are several factors that cause heart disease, including age, type of chest pain, high blood pressure, sugar levels, ECG test values, maximum heart rate, and induced angina. To reduce the percentage of deaths due to heart disease, we need a system that can predict heart disease. The algorithm used in this research is a combination of the Backward Elimination and Naive Bayes algorithms to increase accuracy in diagnosing heart disease. According to the results of this research, the Naive Bayes algorithm has an accuracy value of 78.90% and an Area Under Curve (AUC) value of 0.86, which is included in the good classification category. Combining the Backward Elimination and Naïve Bayes algorithms has an accuracy value of 82.31% and an Area Under Curve (AUC) value of 0.88.

1. INTRODUCTION

The healthcare industry possesses a vast amount of health data, but most of this data is not processed to uncover hidden insights for effective decision-making by healthcare practitioners. Making decisions based on accurate data and information can lead to precise disease diagnoses and predictions. Heart disease is the number one leading cause of high mortality rates, and it continues to be feared by people.

Coronary Heart Disease (CHD) is a cardiovascular disease caused by the blockage of coronary arteries due to plaque buildup, pollutants, or environmental chemicals typically ingested through food, drink, or inhaled as gases that accumulate on the coronary artery walls. This possibility of blood clot formation in the narrowed arteries can block blood flow entirely because the hardened blood clot obstructs the artery's passage [1]. CHD increases with aging in both men and women aged 71-75 years due to the progressive accumulation of atherosclerosis in the coronary arteries as people age. In the cardiovascular system, the aging process results in a decrease in heart rate, and narrowing the coronary arteries' lumen can disrupt blood flow to the heart muscle, leading to damage and impaired heart muscle function [2].

According to the World Health Organization (2022), cardiovascular diseases are the number one cause of death worldwide. Currently, there are an estimated 17.9 million deaths

[DOI: 10.14710/JSUNIMUS.11.2.44-50]

attributed to cardiovascular diseases every year. Heart failure accounts for 85% of deaths among cardiovascular disease patients [3].

Indonesia ranks third in the world in terms of the highest death rate due to cardiovascular diseases, following Laos and the Philippines. The Basic Health Research (Riskesdas) data from the Indonesian Ministry of Health in 2018 reported that heart failure cases in Indonesia are increasing every year, with an estimated 2,784,064 individuals affected. The 2018 Riskesdas data also reported that the prevalence of heart disease based on doctor diagnoses in Indonesia reached 1.5%, with the highest prevalence in North Kalimantan at 2.2%, Yogyakarta at 2%, and Gorontalo at 2%. Besides these three provinces, there were eight other provinces with higher prevalence rates compared to the national average. These eight provinces are Aceh (1.6%), West Sumatra (1.6%), DKI Jakarta (1.9%), West Java (1.6%), Central Java (1.6%), East Kalimantan (1.9%), North Sulawesi (1.8%), and Central Sulawesi (1.9%) [4].

Therefore, there is a need for a precise method to process this data to generate healthcare plans that prioritize disease management efforts to reduce hospitalizations and deaths among heart failure patients. One approach to addressing this issue is by using data mining techniques. Several data mining techniques have been applied in the healthcare field, such as classification and prediction [5]. Classification algorithms like Naive Bayes can be used in classification tasks by using mathematical probability calculations. The advantage of using Naive Bayes is that it requires a small amount of training data to estimate the necessary parameters for the classification process [6]. Previous research has also attempted to predict the life expectancy of heart failure patients. Applying the Particle Swarm Optimization (PSO) method to the Naive Bayes algorithm for feature selection. The classification results showed that the PSO method combined with Naive Bayes achieved better accuracy and fell into the category of Excellent Classification [7]. Additionally, the Naive Bayes algorithm was optimized using Forward Selection for the classification of chronic kidney disease [8]. According to Larose, feature selection using Backward Elimination removes attributes that are not independent of the processed data, while forward elimination retains all attributes, whether independent or not, making non-independent attributes insignificant and reducing their residual square count [9]. Backward Elimination yields better performance compared to statistical significance methods in the selection phase, as demonstrated by high sensitivity, specificity, and accuracy [10]. The purpose of this research is to compare the Naive Bayes algorithm with the Naive Bayes algorithm-based Backward Elimination in detecting heart disease.

2. LITERATURE REVIEW

2.1. Heart Disease

One of the abnormal heart functions is the resulting atherosclerotic heart disease from a builtup of plaque that hardened and makes the arteries narrow. This plaque buildup causes blood flow is not smooth and ultimate obstructed. A heart attack can occur when blood flow inhibitate due to blood clots [11].

Heart disease occurs when accumulated plaque blocks blood flow to the muscles hertz in the coronary arteries [12]. When the fat deposits grew, it reaches about 70% of the diameter of coronary arteries and may begin to cause symptoms [13]. The main symptoms of heart disease is chest pain or angina. This pain could be fell in the center of the chest, but can spread or only felt in the neck, shoulders, arms, or lower jaw, especially on the left side of the body. For some for most people, these symptoms almost always occur along with or after physical activity emotional feelings and may occur after eating or in cold weather [14].

Heart disease is a type of clinical syndrome characterized by dyspnea and limited activity tolerance, which is attributed to impairment of ejection dysfunction or ventricular filling, or a

[DOI: 10.14710/JSUNIMUS.11.2.44-50]

combination of both [15]. Various types of heart diseases are Coronary heart disease, Cardiomyopathy, Cardiovascular disease, Ischemic heart disease, Heart failure, Hypertensive heart disease, Inflammatory heart disease, Valvular heart disease. As per the survey conducted by WHO, out of 10 deaths in India, eight are caused by cardiovascular diseases and diabetes. Preventive strategies to reduce risk factors are essential to reduce the alarmingly increasing burden of heart disease in our population.

2.2. Data Mining

Data mining is the process of extracting useful information and patterns from very large datasets. Data mining encompasses data collection, data extraction, data analysis, and statistical analysis of data [16]. Data mining is also known as knowledge discovery, knowledge, extraction, data/pattern analysis, and information harvesting. The goal of data mining is to uncover previously unknown patterns. Once these patterns are obtained, they can be used to address various types of problems. Data mining is a process that utilizes statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and relevant knowledge from various large databases.

2.3. Naïve Bayes

The Naive Bayes algorithm is a simple probability classifier that calculates a set of probabilities by counting the frequency and combinations of values in each data set. The algorithm uses Bayes' theorem and assumes that all variables are independent considering the value of the class variable. This conditional independence assumption is rarely valid in real-world applications, so it is characterized as Naive, but the algorithm tends to learn quickly in a variety of controlled classification problems [17]. The Naive Bayes algorithm is commonly used in predicting the probability of membership of a class. In doing a calculation of the probability value P(A|B), the Bayesian Theorem uses probabilities P(A), P(B), and P(B|A) as follows.

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A)\ P(B|A)}}{\mathrm{P(B)}} \qquad (1)$$

Where;
$P(A|B)$ is the probability of the occurrence of event A when event B occurs,
$P(A)$ is the probability of the occurrence of A
$P(B|A)$ is the probability of the occurrence of event B when event A occurs
$P(B)$ is the probability of the occurrence of B
Naïve Bayes is applied to the data set and the confusion matrix is generated for class gender having two possible values i.e., Healthy controls or Patients.

2.4. Backward Elimination

Backward Elimination is used to select the best among all possible features to remove and analyze statistically inappropriate ones, the Backward Elimination method provides better performance compared to the significance statistical method in the selection phase. The best performance is evidenced by high sensitivity, specificity, and accuracy [10].

3. METHODOLOGY

This paper was conducted using the experimental method using the software Rapid Miner. The steps are as follows Dataset, Data Analysis Technique, Proposed Algorithm, Testing Method, and Evaluation Method.

[DOI: 10.14710/JSUNIMUS.11.2.44-50]

3.1 Data and Source Data

In this paper, private patient data on heart disease were used. The results obtained from the examination included a total of 209 patients, with 92 patients detected as having heart disease and 117 patients detected as healthy. The attributes in the dataset include age, type of chest pain, high blood pressure, glucose level, EKG test results, maximum heart rate, and induced wind.

3.2 Steps of Analysis

The initial data processing stage is carried out to prepare data that is truly valid before being processed at the next stage, but not all data can be used because there is some data that has missing values, and not all attributes are used because they must go through several stages of initial data processing (data preprocessing). Some of the data analysis techniques used in this paper include the following:

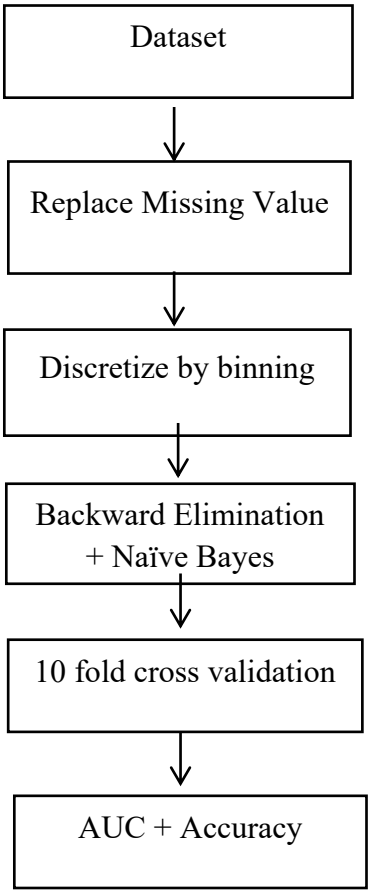

Fig 1. Flowchart Steps of Analysis

1. Dataset
   The first step is to enter heart disease data.
2. Replace Missing Value
   In this research, missing values in the nominal dataset will replaced with values that have the highest frequency in the dataset. Meanwhile, in the numerical dataset, missing values are replaced with the median value of the attribute..
3. Discretize by binning

[DOI: 10.14710/JSUNIMUS.11.2.44-50]

Discretize by binning is carried out to change numeric attribute data into nominal data (by adjusting the bin value according to the original).

4. Backward Elimination + Naïve
   Carry out testing on all attributes that are inappropriates and have no effect on modeling, then calculate the probability of the class and label.
5. 10 fold cross validation
   The testing method used in this paper employs 10-fold Cross-Validation. 10-fold cross-validation is one of the techniques used to assess or validate the accuracy of a model built based on a specific dataset. The creation of a model is usually aimed at making predictions or classifications on new data that may not have appeared in the dataset. The data used in the model-building process is referred to as training data, while the data used to validate the model is referred to as test data. In this technique, the dataset is randomly divided into K partitions. Then K experiments are performed, with each experiment using the K-th partition as the testing data and utilizing the remaining partitions as the training data.

Fig 2. 10-Fold Cross Validation

6. Auc + Accuracy
   Classification methods can be evaluated based on several criteria such as accuracy, speed, reliability, scalability, and interpretability. The testing results are conducted to measure the accuracy and AUC (Area Under Curve) of the heart disease dataset using the 10-fold cross-validation method.

Table 1. Classification AUC

AUC Value	Classification
0,90 – 1,00	Excellent
0,80 – 0,90	Good
0,70 – 0,80	Average
0,60 – 0,70	Low
0,50 – 0, 60	Fail

4. RESULTS AND DISCUSSION

The Experiment was conducted using an inter Core i3 2.53 GHz laptop with 2GB of RAM, running the Windows 7 32-bit operating system. The application used was RapidMiner 7.1 Table2 shows the AUC and Accuracy values for the proposed algorithm.

[DOI: 10.14710/JSUNIMUS.11.2.44-50]

Table 2. Result of Accuracy dan AUC

Methods	Accuracy	AUC
Naïve Bayes	78.90 %	0.86
Naïve Bayes with Backward Elimination	82.31 %	0.88

Based on the results above, the accuracy value for the Naive Bayes method is smaller than the accuracy value for Naive Bayes with backward elimination. So that using feature selection such as backward elimination has a good effect compared to just using naive Bayes without feature selection.

Apart from that, the Accuracy and AUC values also dominate when using the Naive Bayes method with backward elimination, with the results of the Accuracy and AUC values can be seen in the following figure. Apart from that, the ACC and AUC values also dominate when using the Naive Bayes method with backward elimination, with the results of the ACC and AUC values can be seen in the following figure.

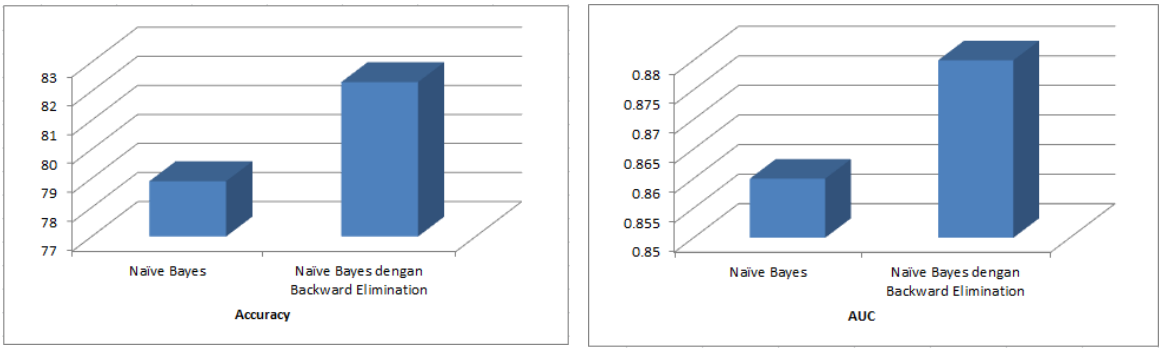

Fig 3. The results of Accuracy and AUC between Naïve Bayes and Naïve Bayes with Backward Elimination

5. CONCLUSION

The algorithm used in this research is a combination of the Backward Elimination and Naive Bayes algorithms to increase accuracy in diagnosing heart disease.The Backward Elimination method is proven to be able to reduce the dimensions of a large dataset and help improve the accuracy of Naïve Bayes classification, achieving an accuracy of 82.31% as compared to using the Naïve Bayes algorithm alone, which yielded an accuracy of 78.90%.

REFERENCES

[1] I. Iskandar, A. Hadi, and A. Alfridsyah, "Faktor Risiko Terjadinya Penyakit Jantung Koroner pada Pasien Rumah Sakit Umum Meuraxa Banda Aceh," *AcTion Aceh Nutr. J.*, vol. 2, no. 1, p. 32, 2017, doi: 10.30867/action.v2i1.34.

[2] D. AR and B. Indrawan, "Hubungan Usia dan Merokok pada Penderita Penyakit Jantung Koroner di Poli Penyakit Dalam RS MHPalembang Periode Tahun 2012," *Syifa' Med. J. Kedokt. dan Kesehat.*, vol. 5, no. 1, p. 16, 2014, doi: 10.32502/sm.v5i1.1420.

[3] F. Febby, A. Arjuna, and M. Maryana, "Dukungan Keluarga Berhubungan dengan Kualitas Hidup Pasien Gagal Jantung," *J. Penelit. Perawat Prof.*, vol. 5, no. 2, pp. 691–702, 2023, doi: 10.37287/jppp.v5i2.1537.

[4] Riskesdas, "Badan Penelitian dan Pengembangan Kesehatan Kementerian Kesehatan Republik Indonesia." 2018.

[5] 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, p. 16, 2020, doi: 10.1186/s12911-020-1023-5.

[6] W. Triprasojo, A., Mauliana, P., & Wiguna, "Penerapan Algoritma Naive Bayes Untuk Klasifikasi Deteksi Mesothelioma," *J. Inform.*, pp. 1–8, 2019.

[7] F. Novaldy and A. Herliana, "Penerapan Pso Pada Naïve Bayes Untuk Prediksi Harapan Hidup Pasien Gagal Jantung," *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 1, pp. 37–43, 2021, doi: 10.51977/jti.v3i1.396.

[8] M. Rizal, M. Z. Syahaf, S. R. Priyambodo, and Y. Rhamdani, "Optimasi Algoritma Naïve Bayes Menggunakan Forward Selection Untuk Klasifikasi Penyakit Ginjal Kronis," *Naratif J. Nas. Riset, Apl. dan Tek. Inform.*, vol. 5, no. 1, pp. 71–80, 2023, doi: 10.53580/naratif.v5i1.200.

[9] D. T. Larose, *Data Mining Methods and Models*. Simultaneously in Canada, 2007.

[10] A. Narin, Y. Isler, and M. Ozer, "Investigating the performance improvement of HRV Indices in CHF using feature selection methods based on backward elimination and statistical significance," *Comput. Biol. Med.*, vol. 45, pp. 72–79, 2014, doi: https://doi.org/10.1016/j.compbiomed.2013.11.016.

[11] M. CHABIB, "Persepsi Perempuan Tentang Penyakit Jantung Koroner di Puskesmas Jenangan, Kecamatan Jenangan Kabupaten Ponorogo," Universitas Muhammadiyah Ponorogo, 2017.

[12] R. Rulandi, "Hubungan Karakteristik Antara Usia, Jenis Kelamin, Tekanan Darah dan Dislipidemia dengan Kejadian Penyakit Jantung Koroner di Rumah Sakit Al-Ihsan Tahun 2014," UNISBA, 2016.

[13] RR WIDYASARI, "Studi Penggunaan Golongan Statin Pada Pada Pasien Jantung Koroner (Penelitian dilakukan di Rumah Sakit Umum Daerah Sidoarjo)," Universitas Muhammadiyah Malang, 2017.

[14] T. Alawiyah, "Gambaran Pengetahuan Penderita Pjk Tentang Bahaya dan Akibat Makanan yang Mengandung Kolesterol," Universitas Muhammadiyah Semarang, 2018.

[15] X. Chen and M. Wu, "Heart failure with recovered ejection fraction: Current understanding and future prospects," *Am. J. Med. Sci.*, vol. 365, no. 1, pp. 1–8, 2023, doi: https://doi.org/10.1016/j.amjms.2022.07.018.

[16] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," pp. 1–9, 2012, [Online]. Available: http://arxiv.org/abs/1206.1121

[17] M. Arhami and M. Nasir, *Data Mining- Algoritma dan Implementasi*. Indonesia: ANDI, 2020.