



Implementation of the Naive Bayes Algorithm to Predict the Safety of Heart Failure Patients

^{1*}Okky Putra Barus , ²Kevil Lauwren, ³Jefri Junifer Pangaribuan , ⁴Romindo
^{1,2,3,4}Information Systems, Pelita Harapan University, Indonesia
E-mail: ^{1*}okky.barus@uph.edu, ²jefri.pangaribuan@uph.edu, ³romindo@uph.edu
*Corresponding author

(Received November 8, 2023 Revised November 18, 2023 Accepted November 26, 2023, Available online December 19, 2023)

Abstract

Heart disease stands as a prominent contributor to global mortality, as indicated by data released by the World Health Organization (WHO). In 2019 alone, an estimated 17.9 million individuals succumbed to cardiovascular disease, accounting for 32% of all worldwide deaths. Of these fatalities, 85% were attributed to heart disease and stroke. Individuals harboring the potential for heart failure often persist in unhealthy lifestyles, regardless of their awareness of underlying heart conditions. To address this issue, the research explores the application of machine learning to identify an optimal method for classifying heart failure patients, employing the Naive Bayes technique. This algorithm has found extensive use in the health sector, demonstrating success in classifying various conditions such as hepatitis, stroke, respiratory infections, and more. The Naive Bayes algorithm, applied in this study, exhibited notable accuracy, precision, sensitivity, and overall classification efficacy. Specifically, the classification accuracy for heart failure patients reached 74.58%, the precision level was 97.67%, sensitivity achieved 75%, and the AUC (Area Under ROC Curve) stood at 0.857, indicating excellent classification within the 0.80 to 0.90 range. These findings can serve as an early warning system for individuals at risk of heart failure.

Keywords: Data Mining, Heart Failure, Naive Bayes

*This is an open access article under the [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).
Copyright © 2023 IAIC - All rights reserved.*



1. Introduction

Heart failure stands as a critical health concern, posing the highest mortality rate in both developed and developing nations. The global impact is staggering, with an estimated 17.9 million individuals succumbing to cardiovascular disease in 2019 alone, accounting for a staggering 32% of all deaths worldwide [1] [2]. The predominant contributors to this alarming statistic are heart attacks and strokes, collectively constituting 85% of cardiovascular-related fatalities. It is imperative to note that the terminal phase of any heart attack is characterized by heart failure, a condition wherein the heart loses its capacity to efficiently pump blood to meet the body's physiological demands [3].

In Indonesia itself, based on the results of the Basic Health Research Research Agency for Health Development of the Indonesian Ministry of Health [4], it was reported that the estimate of heart failure sufferers based on a doctor's diagnosis was estimated at 0.13%, or 229,696 people, while based on the diagnosis and symptoms, it was estimated by 0.3% or 530,068 people [5]. The total number of heart failure patients in 2013 was 759,764. When compared with the results of Riskesdas in 2018, it was reported that estimates of heart disease sufferers based on doctor's diagnoses among residents of all ages in each province reached 1.5%, or as many as 1,017,290 people. Compared to the results of Riskesdas in 2013 and 2018, the number of people with heart disease has increased by 33.89% [6].

Seeing how high the number of people with heart disease is and the increasing mortality rate globally and in Indonesia raises an important question: "How can we turn past patient clinical data into useful information to support health practitioners' decisions in treating heart failure patients?". In numerous healthcare settings, information systems predominantly serve operational functions like inpatient billing and inventory management, lacking a focus on decision-support capabilities for patient care [7]. The conventional approach relies heavily on the clinical expertise of experienced physicians, often sidelining valuable data within databases that could serve as a rich source of information when effectively harnessed through data mining techniques. Notably, researcher Robert Wu advocates for a paradigm shift by suggesting the integration of decision-support information systems with historical patient records. This integration [8] holds the potential to mitigate medical treatment errors, enhance patient safety, reduce practice-related mistakes, and facilitate overall patient outcomes.

Recognizing the significance of predictive analysis based on historical data, the Naive Bayes algorithm emerges as a favorable choice for its ability to anticipate future opportunities with minimal training data requirements for estimating essential parameters in the classification process [9], [10]. This unique advantage prompts researchers to adopt the Naive Bayes data mining method as the focal point of this study. By leveraging this method, the study aims to bridge the existing gap in healthcare information systems, emphasizing the potential for enhanced decision-making, reduced errors, and, ultimately, improved patient care [11], [12].

To provide a comprehensive perspective on the research methodology, an exploration of data mining techniques beyond the Naive Bayes algorithm has been incorporated for comparative analysis. In a noteworthy study focused on detecting heart disease, the Nearest Neighbor Classification (K-NN) algorithm was employed. Utilizing a dataset distinct from the one used in this study, the research applied the K-NN algorithm with K set to 9. The outcomes revealed an accuracy rate of 70%, showcasing the algorithm's effectiveness in identifying heart disease patients. Additionally, the Area Under the Curve (AUC) value stood at 0.875, indicating excellent classification performance in heart disease detection [13], [14].

This research employed the K-NN algorithm in the context of heart disease detection, and the ensuing results will be juxtaposed with the findings derived from the Naive Bayes algorithm in the present study [15][16][17]. By incorporating diverse data mining techniques for comparison, the study aims to discern the strengths and limitations of each algorithm, offering a nuanced understanding of their applicability in healthcare decision support [18]. This approach enhances the robustness of the research findings and contributes to a more thorough evaluation of the chosen Naive Bayes algorithm in the specific context of heart failure prediction. There is also research on heart disease prediction using several data mining techniques and datasets different from this study [19]. The data mining techniques used are the Naive Bayes, Decision Tree, and Neural Network methods. The three data mining models can answer complex questions and provide detailed information [20]. However, Naive Bayes manages to answer four of the five research objectives, whereas the Decision Tree only manages to answer three, and the Neural Network only manages to answer two [21].

2. Research Method

The framework that was created to become a reference and guideline for this research activity [22] is:

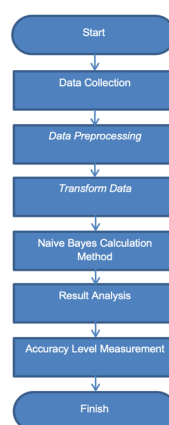


Figure. 1. Data Mining Implementation Flowchart with Naive Bayes

1. Data Collection

The dataset used in this study is data on heart failure patients from the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad from April 2015 to December 2015 [23]. The data collected consisted of 299 patient data with 13 attribute columns where 203 heart failure patients lived and 96 patients who did not survive [24]. Of the 13 attribute columns in this dataset, the researcher adjusted the attributes used in this study so that there were 12 attributes left, which can be seen in Table 1.

Table 1. Heart Failure Patient Dataset Attributes

No.	Attribute	Description
1	Age	Patient's age in years (number).
2	Anemia	Does the patient suffer from reduced red blood cells/hemoglobin? (Yes/No)
3	Creatine Phosphokinase	Blood level of the enzyme Creatine Phosphokinase (mcg/L)
4	Diabetes	Does the patient have diabetes? (Yes No)
5	Ejection Fraction	The percentage of blood leaving the heart per contraction (%)
6	Hypertension	Does the patient have high blood pressure? (Yes No)
7	Platelets	Platelet levels in the blood (kiloplatelets/mL)
8	Creatinine	Blood creatinine level (mg/dL)
9	Serum Sodium	Level of sodium in the blood (mEq/L)
10	Sex	Male or Female
11	Smoking	If the patient smokes? (Yes/No)
12	Patient Safety	If the patient died during the follow-up period? (Safe/Unsafe)

2. Data Preprocessing, which checks the dataset columns to be used, such as ensuring the completeness of data attributes and removing outliers (data that are significantly different from other data) [25].
3. Data Transformation, where the data has been cleaned through a classification process, is the first stage of calculation using the Naive Bayes method [26][27].
4. Application of Naive Bayes: After the data is classified, it is divided into training and testing data. As for the percentage distribution used by researchers, it is as much as 80% used as training data (240 out of 299 data), and as much as 20% used as data testing (59 out of 299 data) [28]. The workflow for the Naive Bayes method is as follows:
 - a. Calculate the initial probability of each class of events
 - b. Calculating the probability of a detailed attribute in a class.
 - c. Multiplies all class attributes by the occurrence class.
 - d. Comparing results between classes.

The application of Naive Bayes is carried out on all data testing to determine the probability of the final result of the data, then placing them into existing classes (the classes in this study are "Safe" or "Unsafe").
5. Analysis of Results After getting the predicted results from Naive Bayes calculations with data testing in the "Safe" and "Unsafe" classes. Thus, the researchers measured the accuracy of the survival prediction in heart failure patients [29].

3. Results and Analysis

3.1. Data Collection Results

The Faisalabad Institute of Cardiology and Allied Hospital dataset in Faisalabad from April 2015 to December 2015 has 299 patient data with 13 variable columns [30]. Still, this study only used 12 variable columns.

3.2. Research Results

The dataset then goes through the Data Preprocessing process, where the data is checked against the column to be used and ensures the completeness of the data variables. After that, a data transformation

was carried out where the data was classified; then, calculations were carried out using the Naive Bayes method. The following are the steps for carrying out this research:

1. Preprocessing Data: out of 299 data from heart failure patients, no blank or incomplete data was found. So that as many as 299 data continue to the data transformation stage.
2. Data Transformation so that heart failure patient data can be used in Naive Bayes calculations.
3. Application of Naïve Bayes, namely calculating the predicted results and placing the patient into the "Recovered" class (the patient's expected results have a higher number of scores in the class recovered from heart failure) or the "Dead" type (the patient's predicted results have a higher number of values increased in the death class of heart failure). Naïve Bayes calculation divides 299 datasets into 80% training and 20% testing data. What is shown in this study is the calculation of the first testing data as an example of analysis, and the rest of the total of the other 58 testing data will be summarized.
4. Testing the level of accuracy using the confusion matrix to measure the performance of the Naïve Bayes prediction results in the form of accuracy, precision, and recall. The results of the confusion matrix calculation are shown in Table 4.2.

Table 1. Confusion Matrix

		Class Prediction	
		Safe	Unsafe
Real class	Safe	42 (TP)	14 (FN)
	Unsafe	1 (FP)	2 (TN)
TOTAL		59	

$$accuracy = \frac{TP+TN}{Total} = \frac{42+2}{59} = 74,57\%$$

$$precision = \frac{TP}{TP+FP} = \frac{42}{42+1} = 97,67\%$$

$$recall = \frac{TP}{TP+FN} = \frac{42}{42+14} = 75\%$$

From the accuracy test using the confusion matrix using 299 data consisting of 240 training data (80%) and 59 testing data (20%), it is obtained:

1. Accuracy of 74.58%, which means that as many as 44 data are correctly predicted from all classes "Safe" and "Not safe," totaling 59 data.
2. The precision is 97.67%, which means that as many as 42 data are correctly predicted into the "Safe" class of the 43 data indicated into the "Safe" class.
3. Sensitivity of 75%, which means that as many as 42 data are predicted correctly out of 56 data that are in the "Safe" class

Then, visualized with the ROC curve obtained an AUC (Area Under Curve) value of 0.857, where the results fall into the category of Good Classification (accuracy level is within 0.80 - 0.90).

From the results of the accuracy test above, the researcher compared the effects of accuracy, precision, sensitivity, and AUC values with research conducted by Mei Lestari using the nearest neighbor classification method.

Comparison results show that the naive Bayes algorithm excels in accuracy and sensitivity, while the nearest neighbor classification excels in precision and AUC value. However, both algorithms achieve good sort AUC values.

4. Conclusion

From the results of the discussion in the previous chapter, it can be concluded that using the Naive Bayes data mining method in predicting the survival of heart failure patients is as follows.

1. The steps in predicting the safety of heart failure patients using the Naïve Bayes algorithm include preprocessing the dataset used first (ensuring data is complete and removing outliers), data variables

are classified first (transformation) so they can be processed using the Naïve data mining method Bayes. After producing the prediction results, they are tested using the accuracy measurement method, the Confusion Matrix, to make the results' accuracy, precision, and sensitivity.

2. The results of measuring the level of accuracy using the Confusion Matrix for predicting the safety of heart failure patients resulted in an accuracy rate of 74.58%, a precision level of 97.67%, and a sensitivity level of 75% with the distribution of training data by 80% and data testing by 20 % of a total of 299 heart failure patient data.
3. The prediction results obtained from applying the Naïve Bayes method can be used as supporting material for health practitioners in testing patient safety but are not allowed to be the final reference for decisions. If the patient's prediction results fall into the "Deceased" class, it is immediately recommended to be checked and receive exceptional medical treatment.

Of course, this research and calculation still require further development to obtain maximum results, so the author has some suggestions as follows:

1. Can use several variations in the distribution of training data and data testing in addition to the distribution of 80% training data and 20% data testing carried out by researchers to explore the optimal percentage of data sharing when calculated using the Naïve Bayes data mining method.
2. Researchers recommend using data mining methods other than Naïve Bayes because not all problems have to be solved with one data mining algorithm. Therefore, explore other data mining methods and compare them to determine the most accurate algorithm.

5. Acknowledgment

This work is supported by LPPM (Lembaga Penelitian dan Pengabdian kepada Masyarakat) of Universitas Pelita Harapan.

References

- [1] WHO, "Cardiovascular diseases (CVDs)." Accessed: Nov. 29, 2023. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [2] T. Hariguna, B. Bin Madon, and U. Rahardja, "User'intention to adopt blockchain certificate authentication technology towards education," in *AIP Conference Proceedings*, AIP Publishing, 2023.
- [3] O. P. Barus and N. Surantha, "The Classification Of Arrhythmia Using The Method Of Extreme Learning Machine," *ICIC Express Letters*, vol. 14, no. 12, pp. 1147–1154, Dec. 2020, doi: 10.24507/icicel.14.12.1147.
- [4] A. Dudhat, "Application of Information Technology to Education in the Age of the Fourth Industrial Revolution," *Int. Trans. Educ. Technol.*, vol. 1, no. 2, pp. 131–137, 2023.
- [5] Kementerian Kesehatan, "Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer." Accessed: Nov. 29, 2023. [Online]. Available: <https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/>
- [6] S. D. Sugiyanti, R. Widayanti, M. B. Ulum, G. Firmansyah, and A. H. Azizah, "Design Dashboard Monitoring Teacher Performance Assessment at Cinta Kasih Tzu Chi High School," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 46–56, 2022.
- [7] M. Zhang *et al.*, "A parsimonious approach for screening moderate-to-profound hearing loss in a community-dwelling geriatric population based on a decision tree analysis," *BMC Geriatr*, vol. 19, no. 1, Aug. 2019, doi: 10.1186/s12877-019-1232-x.
- [8] A. Hermawan, W. Sunaryo, and S. Hardhienata, "Optimal Solution for OCB Improvement Through Strengthening of Servant Leadership, Creativity, and Empowerment," *Aptisi Trans. Technopreneursh.*, vol. 5, no. 1Sp, pp. 11–21, 2023.
- [9] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052068.
- [10] F. Septiyana, M. S. Shihab, H. Kusumah, and D. Apriliasari, "Analysis of the effect of product quality, price perception and social value on purchase decisions for lampung tapis fabrics," *APTISI Trans. Manag.*, vol. 7, no. 1, pp. 54–59, 2023.

-
- [11] F.-J. Yang, "An Implementation of Naive Bayes Classifier," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.
- [12] N. Wiwin, P. A. Sunarya, N. Azizah, and D. A. Saka, "A Model for Determine Upgrades for MSMEs using Analitical Hyrarcy Process," *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 20–32, 2023.
- [13] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-Nn) Untuk Mendeteksi Penyakit Jantung," 2014.
- [14] S. A. Yakan, "Analysis of development of artificial intelligence in the game industry," *Int. J. Cyber IT Serv. Manag.*, vol. 2, no. 2, pp. 111–116, 2022.
- [15] O. Putra Barus and T. Sanjaya, "Prediksi Tingkat Keberhasilan Pengobatan Kanker Menggunakan Imunoterapi Dengan Metode Naive Bayes," vol. 5, no. 1, Jan. 2020, Accessed: Jan. 19, 2023. [Online]. Available: <https://ejournal-medan.uph.edu/index.php/isd/article/view/406>
- [16] U. Rahardja, Q. Aini, P. A. Sunarya, D. Manongga, and D. Julianingsih, "The Use of TensorFlow in Analyzing Air Quality Artificial Intelligence Predictions PM2. 5," *Aptisi Trans. Technopreneursh.*, vol. 4, no. 3, pp. 313–324, 2022.
- [17] K. Wabang, Oky Dwi Nurhayati, and Farikhin, "Application of The Naïve Bayes Classifier Algorithm to Classify Community Complaints," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 872–876, Nov. 2022, doi: 10.29207/resti.v6i5.4498.
- [18] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some Effective Techniques for Naive Bayes Text Classification," *IEEE Trans Knowl Data Eng.*, vol. 18, no. 11, pp. 1457–1466, 2006, doi: 10.1109/TKDE.2006.180.
- [19] B. Rawat and D. Maulidditya, "Entrepreneurship in Information Technology as a Method for Improving Student Creativity in the Digital Economy," *IAIC Trans. Sustain. Digit. Innov.*, vol. 4, no. 1, pp. 32–37, 2022.
- [20] M. Sohail Arshad Associate Professor *et al.*, "EVALUATION OF THE HOSPITAL CARE IN CARDIOVASCULAR DISEASE PATIENTS," 2017. Accessed: Nov. 30, 2023. [Online]. Available: <https://gjms.com.pk/index.php/journal/article/view/727>
- [21] A. Groenewegen, F. H. Rutten, A. Mosterd, and A. W. Hoes, "Epidemiology of heart failure," *Eur J Heart Fail*, vol. 22, no. 8, pp. 1342–1356, Aug. 2020, doi: <https://doi.org/10.1002/ejhf.1858>.
- [22] R. A. O. P. B. Stephanie, "Penerapan UCD dalam Aplikasi Tracking Kalori: OnTrack Solusi Kalori Seimbang," *Buletin Gemastik*, vol. 1, no. 1, pp. 6–10, 2023, Accessed: Nov. 29, 2023. [Online]. Available: <https://buletingemastik.id/index.php/bg/article/view/3/2>
- [23] J. J. Pangaribuan and O. P. Barus, *Extreme Learning Machine: Penerapan dan Aplikasi*. Eureka Media Aksara, 2022.
- [24] R. Wu, W. Peters, and M. W. Morgan, "The next generation of clinical decision support: linking evidence to best practice," *J Healthc Inf Manag*, vol. 16, no. 4, pp. 50–55, 2002, [Online]. Available: <http://europepmc.org/abstract/MED/12365300>
- [25] U. Rahardja, "Camera trap approaches using artificial intelligence and citizen science," *Int. Trans. Artif. Intell.*, vol. 1, no. 1, pp. 71–83, 2022.
- [26] D. Lowd and P. Domingos, "Naive Bayes Models for Probability Estimation," in *Proceedings of the 22nd International Conference on Machine Learning*, in ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 529–536. doi: 10.1145/1102351.1102418.
- [27] S. G. Fitri, R. Selsi, Z. Rustam, and J. Pandelaki, "Naïve bayes classifier models for cerebral infarction classification," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2020. doi: 10.1088/1742-6596/1490/1/012019.
- [28] O. Putra Barus and A. Tehja, "Prediksi Kesembuhan Pasien Covid-19 Di Indonesia Melalui Terapi Menggunakan Metode Naïve Bayes," *Journal Information System Development (ISD)*, vol. 6, no. 2, pp. 59–66, Jul. 2021, Accessed: Jan. 19, 2023. [Online]. Available: <https://ejournal-medan.uph.edu/index.php/isd/article/view/460/267>
- [29] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, 2019. doi: 10.1088/1757-899X/495/1/012033.
- [30] S. H. A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naive Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 2546–2554, Feb. 2018, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/2346>
-