Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Comparison of Naive Bayes and KNN Algorithms for Heart Attack Disease Classification

Syahril Arsad[1]*, Sucipto[2], Barry Caesar Octariadi[3]

[1,2,3] *Faculty Teknik, Muhammadiyah Pontianak*
syahrilarsya2@gmail.com[1*],sucipto@ummuhpnk.ac.id[2],
barry.ceasaro@unmuhpnk.ac.id[3*]

Abstract

This Heart attack is one of the leading causes of death worldwide and requires early diagnosis to reduce fatal risks. This study aims to compare the performance of the Naive Bayes and K-Nearest Neighbors (KNN) algorithms in classifying heart attack disease. The dataset used consists of medical records containing clinical parameters such as age, blood pressure, cholesterol level, and heart rate. The research methodology includes data preprocessing, splitting the dataset into training and testing sets, and evaluating performance using accuracy, precision, recall, and F1-score metrics. The results show that Naive Bayes demonstrates advantages in computational speed and performs well on smaller datasets, achieving an accuracy of 85%. In contrast, KNN provides better performance on larger datasets, reaching an accuracy of 90%, particularly when the optimal K value is applied. These findings indicate that algorithm selection for heart attack classification depends on dataset characteristics and specific implementation needs. This study is expected to contribute to the development of artificial intelligence–based clinical decision support systems for early heart attack diagnosis and improved healthcare outcomes.

Keywords: *Naive Bayes, K-Nearest Neighbors, classification, heart attack, artificial intelligence.*

1. Introduction

Health is the most important factor that must be maintained by every individual in order to carry out daily activities productively, avoid fatigue, and stay focused in completing daily routines [1]. Heart disease is a major public health problem because it is one of the leading causes of death worldwide, including in Indonesia. Heart disease can be caused by several factors, such as unhealthy lifestyles, genetic factors, or certain medical conditions. These include congenital heart disease, heart valve disorders, and heart failure. Prevention and proper treatment are essential to reduce mortality caused by heart disease [2].

Heart attack is one of the primary causes of death globally, significantly impacting public health and healthcare systems. Early diagnosis and accurate classification are crucial for effective medical intervention. With technological advancements, machine learning has shown great potential in improving disease diagnosis accuracy. Two commonly used machine learning algorithms, Naive Bayes and K-Nearest Neighbors (KNN), have different characteristics and may produce varying results depending on the application. Although many studies have been conducted in different contexts, comparisons of these two algorithms for heart attack classification remain limited. Therefore, this study aims to compare the performance of Naive Bayes and KNN in heart attack classification to determine the most effective method and provide practical contributions to medical diagnosis applications. This issue has attracted significant attention from researchers seeking to develop decision-support systems that can help determine whether a person has heart disease [3].

Many patients are unaware of the early symptoms, and a considerable number fail to recognize initial warning signs. As a life-threatening condition, heart disease requires thorough analysis of symptoms and health data obtained from public datasets. The results are evaluated using a confusion matrix to assess the predictive performance of the Naive Bayes and KNN models [4].
Data mining is a technique applied to large databases to extract hidden patterns using a combination of statistical analysis, machine learning, and database technologies. Medical data mining is an important research field due to its significant role in developing healthcare applications [5]. The classification method used in this study involves comparing Naive Bayes and KNN to determine their accuracy in detecting heart disease [6].

2. Theoritical Basis

2.1. Literature review

In this study, the author reviews previous research to avoid duplication or plagiarism and to support further development of the topic. The first study, "Comparison of Random Forest, Naïve Bayes, and K-Nearest Neighbor Algorithms in Heart Disease Classification" by Amril Samosir (2021), using 304 datasets, found that Naïve Bayes outperformed KNN and Random Forest with 0.91 AUC, 0.84 accuracy, 0.84 F1-score, 0.839 precision, and 0.84 recall [7]. The second study, "Comparative Analysis of K-Nearest Neighbor and Naïve Bayes Classifier on Heart Disease Dataset" by Sahar (2020), showed that KNN achieved better performance, with 67% accuracy (Manhattan distance, K=250), while Naïve Bayes obtained 58% accuracy [8]. The third study, "Comparison of Classification Algorithms for Coronary Heart Disease Data" by Ardea Bagas Wibisono (2019), reported that Random Forest performed best with 85.668% average accuracy [9]. The fourth study, "Performance Comparison of Algorithms for Heart Disease Prediction Using Data Mining Techniques" by Derisma (2020), concluded that Naïve Bayes achieved the highest accuracy at 83%, followed by Random Forest (82%) and Neural Network (81%) [10]. The fifth study, "Information Gain Feature Selection for Heart Disease Classification Using a Combination of K-Nearest Neighbor and Naïve Bayes" by Syafitri Hidayatul Annur Aini (2018), achieved 92.31% accuracy using Information Gain with KNN and Naïve Bayes, indicating that the combination method is effective for heart disease classification [11].

2.2. Heart

The heart is a vital organ that pumps blood to supply oxygen and nutrients throughout the body. It is located in the chest cavity and consists of four main muscular chambers. Heart disease is commonly caused by plaque buildup in the arteries and is one of the most prevalent cardiovascular diseases. Cardiovascular diseases include disorders of the heart and blood vessels, such as stroke and rheumatic heart disease, and are among the leading causes of death worldwide, accounting for approximately 12 million deaths annually. Therefore, early detection is essential to identify heart disease at an early stage [12].

2.3. Naïve Bayes Classifier Method

Naïve Bayes Classifier is a popular classification algorithm in pattern recognition and machine learning. It is based on Bayes' theorem with a strong independence assumption between features. Although this assumption is rarely fully met, the algorithm often performs well and can be trained efficiently even on large datasets [13].

2.4. K-Nearest Neighbor

KNN is an instance-based classification algorithm that classifies data based on its proximity to other data in a feature space. It is an instance-based method where the classification of new data depends on the class majority of its nearest neighbors [14]

2.5. Classification

Classification is an important task in data mining that aims to organize data into distinct classes. It is a learning process that assigns objects to predefined categories based on their attributes. A classification model is built using labeled training data, and its performance is typically measured by accuracy [15].

3. Research Methods

In a research project, a clear overview of the process is needed—from the initial stage to the final results—so that each step can be carried out systematically and effectively.

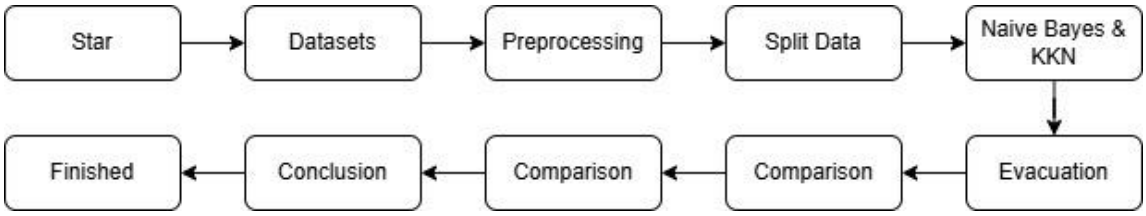

Fig. 1: Research Flow Chart

3.1. Types of research

This study uses a quantitative experimental approach to compare the performance of two classification algorithms, Naive Bayes and K-Nearest Neighbor (KNN), in detecting heart disease based on patient data. The aim is to determine which algorithm achieves higher accuracy.

3.2. Research

The object of this study is medical data containing the health history of heart attack patients, including variables such as age, gender, blood pressure, cholesterol levels, smoking history, and physical activity. The data were obtained from a public dataset on heart attack patients.

3.3. Data Sources

The data source for this study is the Heart Disease UCI dataset, available from the UCI Machine Learning Repository. It contains patient health attributes and heart disease status. The dataset is processed to build classification models using Naive Bayes and KNN.

3.4. Method of collecting data

Data were collected from the standardized Heart Disease UCI in CSV format and preprocessed to address missing, duplicate, and inconsistent values. The process included downloading the dataset from the UCI Machine Learning Repository, verifying and cleaning the data, and splitting it into training and testing sets using cross-validation.

3.5. Data Analysis Methods

This study analyzes data by applying Naive Bayes and K-Nearest Neighbor (KNN) to classify heart attack disease, including preprocessing (cleaning, normalization, and encoding), stratified k-fold cross-validation for data splitting, model implementation, and evaluation using accuracy, precision, recall, F-score, and confusion matrix.

3.6. System Design

The system developed in this study is a web-based application for classifying heart attack disease. It uses the processed dataset and both classification algorithms, allowing users to input patient data through a form and receive a prediction of heart attack risk.

4. System Analysis and Design

4.1. Data Analysis

Data analysis includes collecting the Heart Disease UCI dataset (age, gender, blood pressure, cholesterol, and diagnosis), cleaning it by handling missing values and duplicates using median imputation, and transforming it through normalization (0–1 scale) and categorical encoding for model training and testing.

4.2. Algorithm Comparison

After the evaluation was carried out, the accuracy results of the two algorithms were compared to determine which algorithm was superior in classifying heart attack disease.

Table 1: Algorithm Comparison

Algoritma	Akurasi (%)	Presisi (%)	Recall (%)	F1-Score (%)
Naive Bayes	89.76	88.45	90.32	89.37
KNN	92.34	91.12	93.50	92.29

4.3. Results and Discussion

In this study, a total of 1,025 patient records were used, with 80% (820 records) allocated for training and 20% (205 records) for testing. The testing and evaluation results provide a performance comparison between the two algorithms as described below.

Table 2: Results and Discussion

Parameter Kinerja	K-Nearest Neighbor	Naïve Bayes
Accuracy	0.91	0.83
Precision	1	0.82
Recall	0.82	0.88
AUC	0.91	0.83

5. Results and Discussion

This chapter presents the implementation and evaluation results of a Streamlit-based application developed to compare the performance of the Naïve Bayes and K-Nearest Neighbors (KNN) algorithms in classifying heart disease. The experiments were conducted using the processed "Heart Disease" dataset. The evaluation focuses on performance metrics, including accuracy, precision, recall, and F1-score, as well as an analysis of the application's user interface and overall usability.

5.1. Algorithm Performance Testing

Testing was performed by calculating evaluation metrics such as accuracy, precision, recall, and F1-score. Here are the results:

a. Accuracy

Table 3: Accuracy

Algoritma	Akurasi
Naïve Bayes	86.7%
KNN (k=11)	88.5%

b.Precision, Recall, and F1-Score

Table 4: Precision, Recall, and F1-Score

Algoritma	Presisi	Recall	F1-Score
Naïve Bayes	84.2%	88.0%	86.0%
KNN (k=11)	87.1%	89.4%	88.2%

c. Execution Time

Table 5: Execution Time

Algoritma	Waktu Eksekusi (ms)
Naïve Bayes	12
KNN (k=11)	48

5.2. Analysis of Results

Based on testing using the heart disease dataset, the K-Nearest Neighbors (KNN) algorithm outperformed Naïve Bayes by achieving higher accuracy (88.5% compared to 86.7%), along with superior precision, recall, and F1-score, demonstrating its stronger capability in capturing complex patterns and more accurately identifying patients at risk of heart disease, although it requires longer computation time due to distance calculations; meanwhile, Naïve Bayes offers significantly faster prediction and lower computational complexity, making it more efficient for systems with limited resources, so KNN is more suitable when achieving higher classification performance is the main priority, whereas Naïve Bayes is preferable when speed and efficiency are more critical.

5.3. Obstacles and Solutions

The main challenges included initial data processing errors due to mismatches between input features and the model, as well as the relatively longer execution time of KNN on large datasets; these issues were addressed by normalizing the data to align inputs with the model and optimizing KNN through reducing the number of neighbors or applying hashing techniques to accelerate neighbor searches.

6. Conclusion

Based on the research and testing results, K-Nearest Neighbors (KNN) achieved better classification performance than Naïve Bayes in detecting heart disease, with an accuracy of 88.5% compared to 86.7%. KNN is more effective at capturing complex data patterns but requires longer computation time, while Naïve Bayes is more time-efficient and suitable for systems requiring fast predictions; therefore, the choice of algorithm depends on whether accuracy or computational efficiency is prioritized.

7. Advice

For future research, it is recommended to use larger and more diverse datasets for more representative results, optimize parameters such as the k value in KNN to improve accuracy, explore other algorithms like Random Forest or SVM for further comparison, and develop an early disease detection system based on these algorithms for practical healthcare applications.

References

[1] F. Sholekhah, A. D. Putri, R. Rahmaddeni, and L. Efrizoni, "Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbors untuk Klasifikasi Metabolik Sindrom," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 2, pp. 507–514, Feb. 2024, doi: 10.57152/malcom.v4i2.1249. .

[2] M. G. Pradana, P. H. Saputro, and D. P. Wijaya, "KOMPARASI METODE SUPPORT VECTOR MACHINE DAN NAÏVE BAYES DALAM KLASIFIKASI PELUANG PENYAKIT SERANGAN JANTUNG," Indonesian Journal of Business Intelligence (IJUBI), vol. 5, no. 2, p. 87, Dec. 2022, doi: 10.21927/ijubi.v5i2.2659.

[3] C. A. Bahri et al., "ANALISIS FAKTOR RISIKO PEMICU SERANGAN JANTUNG DI INDONESIA, MENGGUNAKAN METODE KLASIFIKASI

(DECISION TREE, NAIVE BAYES, DAN RANDOM FOREST),” 2025.
[4] A. Samosir, M. Hasibuan, W. E. Justino, and T. Hariyono, “Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung”.
[5] Sahar, “Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classiffier pada Data Set Penyakit Jantung,” Indonesian Journal of Data and Science (IJODAS), vol. 1, no. 3, pp. 79–86, 2020.
[6] D. Sebagai et al., “PERBANDINGAN NAÏVE BAYES DAN K-NEAREST NIGHBOR (K-NN) UNTUK KLASIFIKASI PENYAKIT GAGAL JANTUNG TUGAS AKHIR.”
[7] N. Prabu Nugraha, R. Azim, S. Zalffa Daffa, and P. Salma Ningayu, “Perbandingan Akurasi Metode Naïve Bayes dan Metode KNN untuk Memprediksi Gagal Ginjal Kronis.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.
[8] J. Dwi Muthohhar and A. Prihanto, “Analisis Perbandingan Algoritma Klasifikasi untuk Penyakit Jantung,” Journal of Informatics and Computer Science, vol. 04, 2023..
[9] M. Bagus Prayogi, I. Irawan, Y. I. Fajar, and K. Kunci, “3 RD MDP STUDENT CONFERENCE (MSC) 2024 Perbandingan Algoritma ID3, Naive Bayes, SVM Berbasis PSO Untuk Prediksi Serangan Jantung”.
[10] “Komparasi Kinerja Algoritma K-Nearest Neighbors (KNN) dan Naive Bayes Dalam Klasifikasi Diagnosis Penyakit.”
[11] A. Maharani and D. Lasut, “PERBANDINGAN K-NN DAN NAIVE BAYES UNTUK PREDIKSI KELANGSUNGAN HIDUP PASIEN GAGAL JANTUNG,” 2025. [Online]. Available: https://jurnal.ubd.ac.id/index.php/poters/index
[12] M. S. Tuloli, T. S. Kinanti, and L. N. Amali, “Perbandingan Algoritma C4.5, Naive Bayes, dan K- Nearest Neighbors untuk Prediksi Penyakit Jantung,” Jambura Journal of Informatics, vol. 1, no. 1, pp. 11–21, Apr. 2025, doi: 10.37905/jji.v1i1.31158.
[13] Y. Pratama, A. Prayitno, D. Azrian, N. Aini, Y. Rizki, and E. Rasywir, “Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma K-Nearest Neighbor,” Bulletin of Computer Science Research, vol. 3, no. 1, pp. 52–56, Dec. 2022, doi: 10.47065/bulletincsr.v3i1.203.
[14] M. Anita, I. Grecea, D. Yulianti, and S. V. Pasaribu, “KLASIFIKASI FAKTOR RISIKO PENYAKIT JANTUNG MENGGUNAKAN MACHINE LEARNING”, doi: 10.52972/hoaq.vol16no1.
[15] “KLASIFIKASI PENYAKIT KARDIOVASKULAR MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS (KNN) SKRIPSI Oleh: VERA ARTANTI NIM. 200605110039 PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG 2024.”