Jurnal Galaksi (Global Knowledge, Artificial Inteligent and Information System) E-ISSN: 3048-2399
Vol 1, No 3, December 2024 , pp.190-197
https://doi.org/10.70103/galaksi.v1i3.45

Towards Improved Heart Disease Detection: Evaluating Naïve Bayes and K-Nearest Neighbors in Medical Data Classification

Mariane Cetty Angelyn[1*], Ida Bagus Ary Indra Iswara[2], Desak Made Dwi Utami Putra[3], Ni Nyoman Ayu J. Sastaparamitha[4]

[1*,2,3,4]*Informatika, Fakultas Teknologi dan Informatika, Institut Bisnis dan Teknologi Indonesia*
[*1]*khetyangelyn@gmail.com,* [2]*indraiswara@instiki.ac.id,* [3]*desak.utami@instiki.ac.id,* [4]*ayusasta@instiki.ac.id*
**corresponding author*

ARTICLE INFO

Article history:
Received 14 December 2024
Revised 26 December 2024
Accepted 29 December 2024
Available Online 31 December 2024

Keywords *:*
Classification;
Comparison Method;
Naïve Bayes;
K-Nearst Neighbors

ABSTRACT

The application of machine learning in healthcare is increasingly critical for improving diagnostic accuracy and timely treatment. This study explores the classification of heart disease using Naïve Bayes and K-Nearest Neighbors (KNN), focusing on evaluating their effectiveness through a comparative analysis. The research addresses the challenge of identifying an optimal method for heart disease classification, emphasizing the need for reliable algorithms. Using a dataset from Kaggle with detailed preprocessing, we implement Naïve Bayes and KNN to assess classification performance. The study introduces a comparative perspective on classification accuracy, precision, recall, and F1-score, revealing the strengths and limitations of each method. The results highlight the superior performance of Naïve Bayes with an accuracy of 88%, offering novel insights for data-driven healthcare decisions.

1. Introduction

Data is a foundational component in the advancement of machine learning. With the rapid increase in data generated daily, its proper analysis offers immense potential, especially in the healthcare sector, where precise data interpretation can save lives. One of the key challenges in healthcare analytics is the classification of data for disease diagnosis (Guo & Chen, 2023; Rehman et al., 2022). Accurate classification is critical for making informed clinical decisions. To address this, advanced analytical techniques are essential for organizing, grouping, and classifying healthcare data with efficiency and precision.

Machine learning algorithms have become indispensable tools in healthcare data analysis. Two widely used methods for data classification are Naïve Bayes and K-Nearest Neighbors (KNN), which adopt distinct approaches. K-Nearest Neighbors is an instance-based learning algorithm that classifies data by comparing it to previously stored instances rather than creating an explicit model. It works by identifying the K most similar objects in the training data to classify a new object or test data (Rahman et al., 2021; Zhang et al., 2019).

In contrast, Naïve Bayes is a classification algorithm based on Bayes' theorem, which assumes feature independence (Armaeni et al., 2024). Despite its simplicity, Naïve Bayes often delivers competitive results, particularly in applications like text classification and spam filtering (Gupta et al., 2020). This algorithm estimates the probability of a data point belonging to a specific class based on its features. Its advantages include computational efficiency and strong performance with high-dimensional data.

This study aims to compare the performance of Naïve Bayes and K-Nearest Neighbors in analyzing a dataset. The comparison is particularly relevant across diverse domains such as marketing, healthcare, and finance, where accurate and efficient data analysis provides significant competitive advantages.

The rapid growth of healthcare data necessitates robust analytical techniques to support accurate disease diagnosis. Despite the wide application of machine learning algorithms in healthcare, gaps remain in selecting the most suitable method for specific conditions like heart disease. Recent studies demonstrate the effectiveness of Naïve Bayes and KNN in various domains, but a direct comparison focusing on heart disease remains limited. This study addresses this gap by analyzing the performance of these algorithms on a well-curated dataset, aiming to provide a clear recommendation for practitioners.

This research is driven by the urgent need to enhance heart disease classification methods, given its significant impact on patient outcomes. Existing studies offer partial insights into the application of Naïve Bayes and KNN, but lack a focused comparison on heart disease datasets with diverse clinical attributes. By addressing this gap, the study contributes to the optimization of machine learning techniques, ensuring reliable and interpretable outcomes for medical practitioners.

2. Literature Review

The detection and diagnosis of heart disease have become increasingly reliant on machine learning techniques, particularly Naïve Bayes and K-Nearest Neighbors (KNN). These classifiers offer distinct advantages and challenges in the context of medical data classification, which is critical for timely and accurate patient care (Guo & Chen, 2023).

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, which assumes independence among predictors. It is particularly effective in high-dimensional datasets, making it suitable for medical applications where numerous patient attributes are considered. Studies have demonstrated that Naïve Bayes can achieve high accuracy rates in heart disease prediction, often outperforming other classifiers in specific contexts. For instance, Agarwal's comparative study highlights the effectiveness of Naïve Bayes in predicting heart disease based on patient attributes, showcasing its utility in clinical settings (Agarwal, 2024; Rahman et al., 2021). Furthermore, Yousef and Batiha emphasize the importance of addressing dimensionality issues in prediction systems, proposing a mixed model that incorporates Naïve Bayes to enhance prediction accuracy (Gupta et al., 2020; Yousef & Batiha, 2021).

On the other hand, K-Nearest Neighbors is a non-parametric method that classifies data points based on the majority class among their nearest neighbors. KNN is praised for its simplicity and effectiveness, particularly in scenarios where the decision boundary is irregular. Mohan et al. note that KNN, alongside other machine learning techniques, is widely employed to assess the severity of heart disease (Mohan et al., 2019). Additionally, Khan et al. report that KNN, when used in conjunction with other classifiers, can significantly improve the classification of cardiovascular diseases (Khan et al., 2022). However, KNN's performance can be adversely affected by the curse of dimensionality, which may lead to decreased accuracy in high-dimensional spaces.

The comparative analysis of these two classifiers reveals that while Naïve Bayes often excels in scenarios with high-dimensional data and independence assumptions, KNN offers robustness in cases where the data structure is more complex. For instance, a study by (Damayunita et al., 2022) indicates that KNN achieved an accuracy of 91% compared to Naïve Bayes 88%% in a classification task. Conversely, other studies suggest that KNN may outperform Naïve Bayes in different contexts, such as in predicting COVID-19 patient outcomes, where KNN achieved a higher accuracy (Romadhon & Kurniawan, 2021).

In summary, both Naïve Bayes and K-Nearest Neighbors present valuable methodologies for heart disease detection, each with its strengths and weaknesses. The choice between these classifiers should be informed by the specific characteristics of the dataset and the clinical context, as well as the need for interpretability and computational efficiency in medical applications. Future research should continue to explore hybrid approaches that leverage the strengths of both classifiers to enhance predictive accuracy and clinical utility. Previous studies have compared machine learning methods for disease classification but often focused on different datasets or did not comprehensively explore heart disease. This research aims to address this gap by specifically comparing Naïve Bayes and KNN using a heart disease dataset, evaluating their performance across critical metrics. Unlike prior work, our approach emphasizes the role of data preprocessing and hyperparameter optimization, offering a novel perspective on algorithmic performance.

3. Research Methods

3.1. Research Stages

In this research, there are stages of research to facilitate the explanation of each process at the beginning of data collection to produce forecasts of Indonesian stocks.

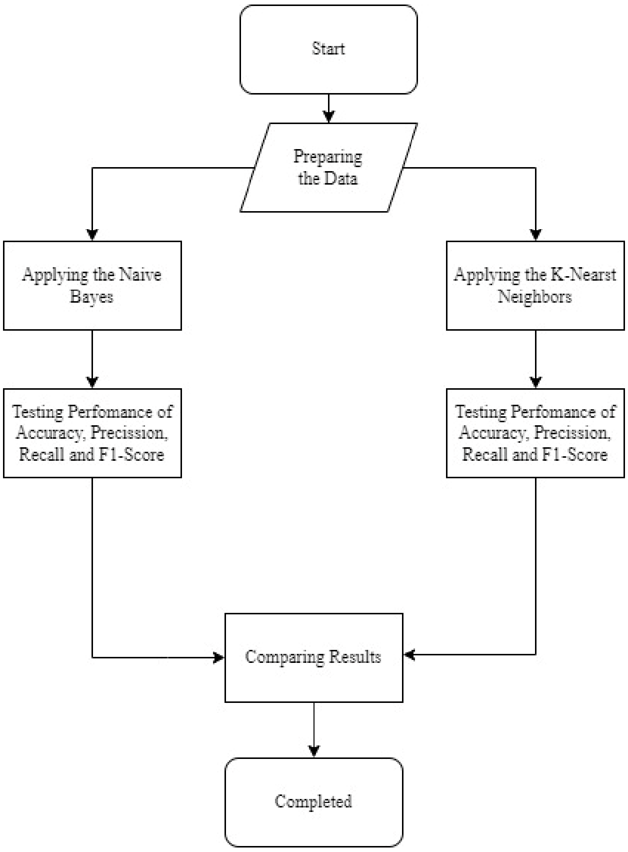

Start
Preparing the Data
Applying the Naive Bayes
Applying the K-Nearst Neighbors
Testing Perfomance of Accuracy, Precission, Recall and F1-Score
Testing Perfomance of Accuracy, Precission, Recall and F1-Score
Comparing Results
Completed

Fig. 1. Research Stages

Naïve Bayes

Naïve Bayes is an algorithm that is part of the classification technique. Naïve Bayes is a probabilistic and statistical classification method proposed by English scientist Thomas Bayes, which predicts future probabilities based on past experiences, known as Bayes' theorem. Compared to other classifiers, Naïve Bayes works better and has a high accuracy rate based on Bayes' theorem (Armaeni et al., 2024; Wu et al., 2019). The chosen Bayes value is the highest percentage.

$$P(H|X) = \frac{P(X|H)}{P(X)} P(H) \quad (1)$$

Description:
X = Unkown class data
H = Specific class hypothesis
P(X) = Probability of X
P(H) = Probability of hypothesis H (prior probability)
P(X|H) = Probability of X given hypothesis H
P(H|X) = Probability of hypothesis H given X (posterior probability)

K-Nearst Neighbors (K-NN)

The K-Nearest Neighbors (K-NN) algorithm is one of the classification methods in data mining. K-NN classifies a set of data based on labeled training data. As a supervised learning method, K-NN is used to classify new objects based on their nearest neighbors. The result of a new query instance will be classified based on the majority category of its neighbors (Devika et al., 2019; Sudipa et al., 2024) . This means that the most frequent class will be chosen as the classification class. K-NN also calculates the distance between old cases and new cases. Classification using K-NN does not require a model, but only relies on memory. The K-NN algorithm uses the nearest neighbor classification as the predicted value for new test samples (Lubis & Lubis, 2020).

Here are the steps to perform classification using the K-Nearest Neighbors (K-NN) algorithm:

1. Determine the suitable value of K based on the data used. The minimum value of K is 1, and the maximum value is the number of training data
2. Calculate the distance between the test data and the training data. To calculate the distance, the K-NN algorithm usually uses the Euclidean distance formula, such as:

$$\mathrm{d(A,B)} = \sqrt{(\mathrm{A_1} - \mathrm{B_1})^2 + (\mathrm{A_2} - \mathrm{B_2})^2 + \cdots + (\mathrm{A_n} - \mathrm{B_n})^2} \quad (2)$$

Description :
d(A, B) = Euclidean distance between points A and B
A and B = Two points in n-dimensional space. Point A is represented by coordinates (A1, A2, …, An), and point B is represented by coordinates (B1, B2, …, Bn).
Ai and Bi = The i-th coordinate of points A and B.

3. Sort the distances from largest to smallest.
4. Determine the nearest distance up to the parameter K.
5. Pair the suitable class.
6. Find the number of classes from the nearest neighbors and assign that class as the class of the data to be classified or evaluated.

4. Results and Discussions

Dataset

The dataset used in this study was sourced from Kaggle and consists of 918 samples with both categorical and numerical attributes. Key features include patient demographics (age, gender), clinical indicators (cholesterol, blood pressure), and diagnostic results (chest pain type, ST slope, exercise-induced angina). The dataset represents a balanced mix of normal and diseased cases, ensuring comprehensive evaluation of classification algorithms.

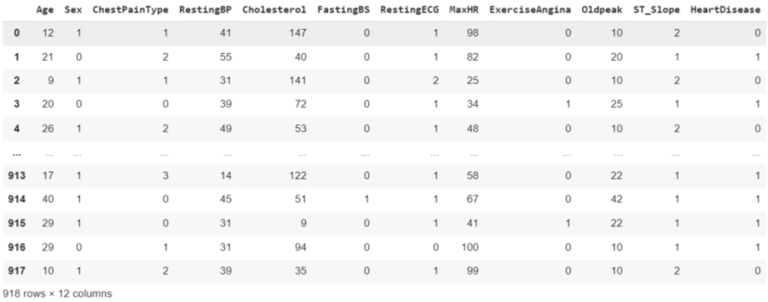
Table 1. Dataset

No	Age	Sex	ChestPain Type	RestingBP	Cholesterol	FastingBS	RestingECG	Max HR	Exercise Angina	Oldpeak	ST_Slope	Heart Disease

1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
918	57	F	ATA	130	236	0	LVH	174	N	0	Flat	1

Data Preparation

As observed in the table above, the dataset consists of both categorical and numerical data. Therefore, in this process, the researcher will convert the data into numerical data before proceeding to the testing phase. The purpose is to enable the data to be utilized by the Naïve Bayes and K-Nearest Neighbors algorithms. By employing the LabelEncoder function in Python, the data is transformed into numerical data.



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	12	1	1	41	147	0	1	98	0	10	2	0
1	21	0	2	55	40	0	1	82	0	20	1	1
2	9	1	1	31	141	0	2	25	0	10	2	0
3	20	0	0	39	72	0	1	34	1	25	1	1
4	26	1	2	49	53	0	1	48	0	10	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	17	1	3	14	122	0	1	58	0	22	1	1
914	40	1	0	45	51	1	1	67	0	42	1	1
915	29	1	0	31	9	0	1	41	1	22	1	1
916	29	0	1	31	94	0	0	100	0	10	1	1
917	10	1	2	39	35	0	1	99	0	10	2	0

918 rows × 12 columns

Fig. 2. Transformed Data

Comparison Results of Naïve Bayes and K-Nearest Neighbors

The following is the result and discussion of the data processing to compare the performance of the two methods in classifying heart disease. Using testing and training data, the data was tested four times, and the highest result from the overall data testing was observed.

Table 2. Split Data Naïve Bayes

Data Testing and DataTraining	Accuracy	Precission	Recall	F1-Score
0,2 - 0,8	0,88	0,88	0,88	0,88
0,3 - 0,7	0,86	0,86	0,86	0,85
0,4 - 0,6	0,86	0,86	0,86	0,86
0,5 - 0,5	0,86	0,86	0,86	0,86

The testing results show that the accuracy, precision, recall, and F1-score of the Naïve Bayes method are relatively stable and high, ranging from 0.86 to 0.88, and are not significantly affected by changes in the training and testing ratio. Therefore, it can be concluded that the Naïve Bayes method has good and stable performance in classifying data, and is not overly sensitive to changes in the training and testing ratio.

Furthermore, the testing results using K-Nearest Neighbors are presented. Using testing and training data, the data was tested four times, and the highest result from the overall data testing was observed.

Table 3. Split Data K-Nearst Neighbors

Data Testing and DataTraining	Accuracy	Precission	Recall	F1-Score

0,2 - 0,8	0,79	0,79	0,79	0,79
0,3 - 0,7	0,76	0,76	0,76	0,76
0,4 - 0,6	0,74	0,75	0,74	0,74
0,5 - 0,5	0,74	0,75	0,74	0,74

The testing results show that the accuracy, precision, recall, and F1-score of the K-Nearest Neighbors method are relatively stable, but slightly decrease as the training ratio increases. The values of accuracy, precision, recall, and F1-score range from 0.74 to 0.79. Therefore, it can be concluded that the K-Nearest Neighbors method has a fairly good performance in classifying data, but its performance slightly decreases if the training ratio increases. This is because the K-Nearest Neighbors method is more sensitive to changes in the training and testing ratio compared to the Naïve Bayes method.

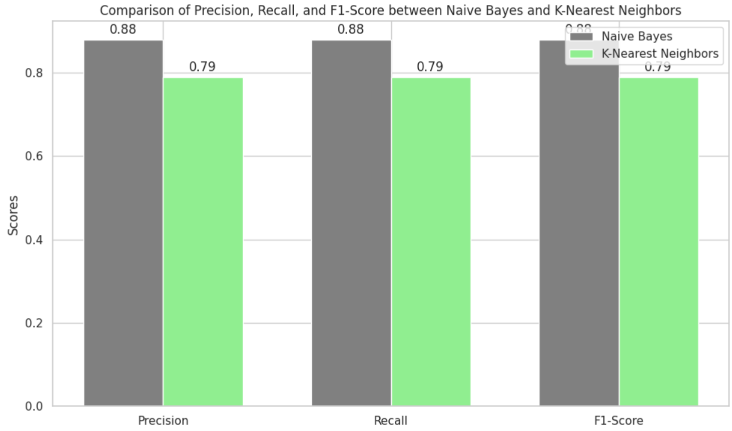


Fig. 3. Bar Chart Comparison of Precision, Recall, and F1-Score between Naive Bayes and K-Nearest Neighbors

Figure 4.2 shows the F1-Score results of the two methods. F1-Score is a measure that indicates the balance between precision and recall. It can be seen that the F1-Score of Naïve Bayes is higher than that of K-Nearest Neighbors for all data splits. The F1-Score of Naïve Bayes ranges between 0.85 and 0.88, while the F1-Score of K-Nearest Neighbors ranges between 0.74 and 0.79.

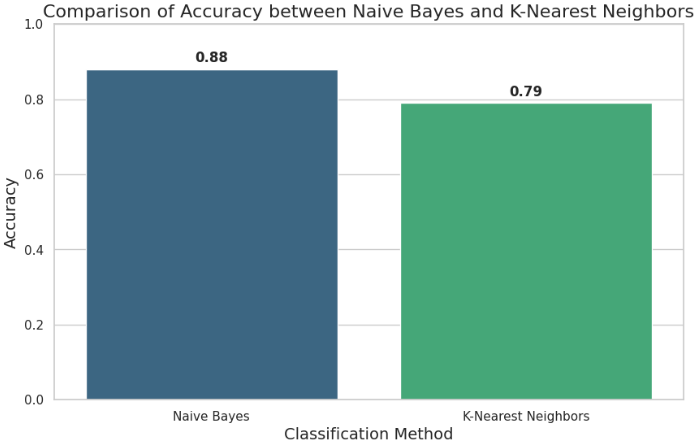


Fig. 4. Accuracy Comparison Chart of Naive Bayes and K-Nearest Neighbors

From Figure 4.3, based on the highest accuracy results between Naive Bayes and K-Nearest Neighbors, it can be seen that Bayes has a higher accuracy compared to K-Nearest Neighbors. Specifically, Naive Bayes achieved an accuracy of 0.88 or 88%, while K-Nearest Neighbors achieved an accuracy of 0.79 or 79%. From these results, it can be concluded that the accuracy of Naive Bayes in classification is better compared to K-Nearest Neighbors for the data used.

Tabel 4. Confusion Matrix Naïve Bayes and K-Nearst Neighbors

	Naïve Bayes		K-Nearst Neighbors	
Class	Score		Score	
	0= No	1= Yes	0=No	1= Yes
	86	17	85	18
	60	75	21	60

From the table above, the Naïve Bayes algorithm produces the following classification results: True Positive (TP): 86 cases of class 0 correctly predicted as class 0 (no heart disease). False Positive (FP): 17 cases actually belonging to class 0 but predicted as class 1 (having heart disease). True Negative (TN): 60 cases of class 1 correctly predicted as class 1. False Negative (FN): 75 cases actually belonging to class 0 but predicted as class 1.

In contrast, the K-Nearest Neighbors algorithm produces the following classification results: True Positive (TP): 85 cases actually belonging to class 0 correctly predicted as class 0 (no heart disease). False Positive (FP) is18 cases actually belonging to class 0 but predicted as class 1 (having heart disease). True Negative (TN): 21 cases of class 1 correctly predicted as class 1. False Negative (FN): 60 cases of class 1 correctly predicted as class 1. From the results above, it can be seen that the Naïve Bayes model has better classification performance than the KNN model, as it has more True Positives and True Negatives, and fewer False Positives and False Negatives.

While Naïve Bayes outperformed KNN in accuracy, precision, recall, and F1-score, a deeper analysis reveals the reasons behind these differences. Naïve Bayes benefits from its probabilistic framework, effectively handling the dataset's categorical features. In contrast, KNN's reliance on distance metrics struggles with higher dimensionality and data distribution. These findings underscore the importance of method selection based on data characteristics, contributing to a more informed application of machine learning in healthcare.

The novelty of this study lies in its comparative framework, offering new insights into the performance dynamics of Naïve Bayes and KNN for heart disease classification. By systematically analyzing accuracy (88% vs. 79%), precision (0.88 vs. 0.79), recall (0.88 vs. 0.79), and F1-score (0.88 vs. 0.79), the research identifies Naïve Bayes as the more reliable algorithm for this specific dataset. This analysis bridges existing gaps and provides a practical guideline for algorithm selection in clinical applications.

5. Conclusion

Based on the research conducted, it can be concluded that the Naive Bayes method provides better results in classifying heart disease data compared to the K-Nearest Neighbors method, based on accuracy, precision, recall, and F1-score. The Naive Bayes method shows high accuracy, with values ranging between 0.86 and 0.88. It also exhibits high precision, recall, and F1-score, with values ranging from 0.86 to 0.88. In contrast, the K-Nearest Neighbors method achieved an accuracy of 79% with precision, recall, and F1-score values of 0.79. From the confusion matrix, it is evident that Naive Bayes has a higher number of true positives, which enables it to better identify patients who actually have heart disease. Based on these findings, it is recommended to consider using the Naive Bayes method for predicting heart disease in similar datasets. Future research could incorporate additional features or improved data processing techniques and explore methods such as parameter optimization or alternative testing methods to further enhance classification performance.

References

Agarwal, N. (2024). Predictive Modelling for Heart Disease Diagnosis: A Comparative Study of Classifiers. *Eai Endorsed Transactions on Pervasive Health and Technology*, *10*(1), 1–11. https://doi.org/10.4108/eetpht.10.5518

Armaeni, P. P., Wiguna, I. K. A. G., & Parwita, W. G. S. (2024). Sentiment Analysis of YouTube Comments on the Closure of TikTok Shop Using Naïve Bayes and Decision Tree Method Comparison. *Jurnal Galaksi*, *1*(2), 70–80. https://doi.org/10.70103/galaksi.v1i2.15

Damayunita, A., Fuadi, R. S., & Juliane, C. (2022). Comparative Analysis of Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) Algorithms for Classification of Heart Disease Patients. *Jurnal Online Informatika*, *7*(2), 219–225. https://doi.org/10.15575/join.v7i2.919

Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 679–684. https://doi.org/10.1109/ICCMC.2019.8819654

Guo, C., & Chen, J. (2023). Big Data Analytics in Healthcare. In Y. Nakamori (Ed.), *Knowledge Technology and Systems: Toward Establishing Knowledge Systems Science* (pp. 27–70). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1075-5_2

Gupta, A., Kumar, L., Jain, R., & Nagrath, P. (2020). Heart Disease Prediction Using Classification (Naive Bayes). In P. K. Singh, W. Pawłowski, S. Tanwar, N. Kumar, J. J. P. C. Rodrigues, & M. S. Obaidat (Eds.), *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)* (pp. 561–573). Springer Singapore. https://doi.org/10.1007/978-981-15-3369-3_42

Khan, A., Khan, A., Khan, M. M., Farid, K., Alam, M. M., & Su'ud, M. B. M. (2022). Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers With SVM as a Meta Classifier. *Diagnostics*, *12*(11), 2595. https://doi.org/10.3390/diagnostics12112595

Lubis, A. R., & Lubis, M. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics*, *9*(1), 326–338. https://doi.org/10.11591/eei.v9i1.1464

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *Ieee Access*, *7*(1), 81542–81554. https://doi.org/10.1109/access.2019.2923707

Rahman, B., Warnars, H. L. H. S., Sabarguna, B. S., & Budiharto, W. (2021). Heart disease classification model using k-nearest neighbor algorithm. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, 1–4. https://doi.org/10.1109/ICIC54025.2021.9632918

Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, *28*(4), 1339–1371. https://doi.org/10.1007/s00530-020-00736-8

Romadhon, M. R., & Kurniawan, F. (2021). A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia. *3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 41–44. https://doi.org/10.1109/eiconcit50028.2021.9431845

Sudipa, I. G. I., Azdy, R. A., Arfiani, I., & Setiohardjo, N. M. (2024). Leveraging K-Nearest Neighbors for Enhanced Fruit Classification and Quality Assessment. *Indonesian Journal of Data and Science*, *5*(1), 30–36. https://doi.org/10.56705/ijodas.v5i1.125

Wu, M., Huang, Y., & Duan, J. (2019). Investigations on Classification Methods for Loan Application Based on Machine Learning. *Proceedings - International Conference on Machine Learning and Cybernetics*, *2019-July*, 1–6. https://doi.org/10.1109/ICMLC48188.2019.8949252

Yousef, M. M., & Batiha, K. (2021). Heart Disease Prediction Model Using NaÃ¯ve Bayes Algorithm and Machine Learning Techniques. *International Journal of Engineering & Technology*, *10*(1), 46–56. https://doi.org/10.14419/ijet.v10i1.31310

Zhang, Y., Cao, G., Wang, B., & Li, X. (2019). A novel ensemble method for k-nearest neighbor. *Pattern Recognition*, *85*, 13–25. https://doi.org/10.1016/j.patcog.2018.08.003