Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 22, No. 3, July 2023, pp. 593∼604
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v22i3.2984

K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset

Alya Masitha , Muhammad Kunta Biddinika , Herman
Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Article Info

Article history:

Received May 22, 2023
Revised June 15, 2023
Accepted July 18, 2023

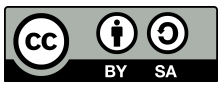
Keywords:

Heart failure
K-Nearest Neighbors
Manhattan Distance
Preprocessing

ABSTRACT

Heart failure is included in the category of cardiovascular disease. Heart disease is not easy to detect, and its detection needs to be done by experienced and skilled medical professionals. Most patients with heart failure require hospitalization. Common symptoms of heart disease, such as chest pain and high or low blood pressure, vary from person to person. This study aims to find the most optimal k value based on the accuracy obtained based on calculations by testing different k values, namely 1, 3, 5, 7, and 9. After getting the results of the accuracy of the five k values, compare which accuracy has the highest value, best for K-Nearest Neighbor (K-NN) models. The classification process uses the K-NN algorithm. This algorithm is quite easy to use because some parameters work using distance metrics and k values. Therefore, the value of k in the K-NN algorithm greatly affects the accuracy that will be produced. In the results of this study, the accuracy obtained was k = 7 and k = 9, which are the most optimal results because they have the highest accuracy compared to other k values, with an accuracy of 88%. The expected benefit of this research is that it can make a scientific contribution to research in the field of machine learning classification, especially in predicting heart failure.

Copyright ©2022 The Authors.
This is an open access article under the CC BY-SA license.



Corresponding Author:

Alya Masitha, +62813 9218 1614,
Faculty of Technology Industry, Master Program of Informatics,
Universitas Ahmad Dahlan, Yogyakarta, Indonesia,
Email: alyamasitha@gmail.com

How to Cite:
A. Masitha, M. Biddinika, and H. Herman, ”K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset”, MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 3, pp. 593-604, Jul. 2023.
This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

1. INTRODUCTION

WHO data states that over 17 million people worldwide die from cardiovascular disease. PTM is the number-one cause of death in the world every year, from 12.1 miles in 2013 to 10.9 miles in 2018. Coronary artery disease remained at 1.5 percent between 2013 and 2018. The premature death rate from heart disease is 4 percent in high-income countries and 42 percent in low-income countries. The heart is the main part of the human body. The heart is the body's operating system. Other functions of the human body will be greatly affected by the irregular functioning of the heart [1].

The heart is one of the most important organs in the body and is one of the vital organs. The heart has an important role in the body by pumping blood to supply oxygen and nutrients throughout the body. The heart will never stop beating, even when sleeping. Humans cannot regulate their heart rate when the heart pumps blood [2]. Diagnosing heart failure requires a thorough evaluation of symptoms, medical history, and the results of a physical examination and medical tests. However, with the development of technology and the ability of machines to process data, machine-learning methods can be used to support the diagnosis and prediction of this disease.

Heart failure is included in the category of cardiovascular disease. Heart failure is when the heart does not receive enough blood to meet the body's needs. Most heart failure patients are hospitalized. The prevalence of heart failure in America and Europe is around 12%. There is no epidemiological data on heart failure in Indonesia, but the National Health Survey found cardiovascular disease to be Indonesia's main cause of death (26.4%), ranked eighth. Due to the value of vital organs such as the heart, the prediction of heart failure has become a priority for doctors and medical personnel. However, until now, the prediction of heart failure in clinical practice has generally not achieved high accuracy [3]. According to the WHO, the most dangerous chronic diseases are causes of death, including breast cancer, heart disease, and diabetes. Most of the sufferers of this disease come from developed countries. Heart disease is also known as coronary artery disease. Common symptoms of heart disease, such as chest pain and high or low blood pressure, vary from person to person [4].

Heart failure is a condition where the heart weakens, so it cannot pump enough blood throughout the body. This condition occurs in many people but is more common in people over 65. All major body functions are disrupted without adequate blood flow. A person suffering from a heart attack suffers when the flow of oxygen-rich blood is restricted to a certain part of the heart, causing insufficient oxygen. During heart failure, plaque breaks off and spills cholesterol and other substances into the blooda clot forms at the rupture position. If the clot is large, it will block blood flow through the blood vessels to the heart arteries. Causes of cardiac muscle death are chemical elements and nutrition (ischemia) [4]. Heart disease is the main cause of death every year in the world. Several types of diseases attack the heart [5]. Heart disease is a certain abnormal condition that includes disorders that occur in the heart and affect blood flow. Heart failure can be determined with the help of an AI approach, in this case, machine learning or classification, where the determination is based on previously obtained data.

The K-NN algorithm is a machine learning algorithm widely used for disease classification [6], one of which is diabetes. This research produces an accuracy of 74.59% [7] in another study predicting Atherosclerosis disease using the K-NN algorithm. The highest accuracy produced with the Hungarian dataset is 80% [8]. The K-NN algorithm is not only used to classify diseases but can also be used in other fields, such as predicting stock trends, according to Indu Kumar et al. This study uses data sourced from Amazon, Cipla, Eicher, Bata, and Bosch. The highest accuracy resulted from the Amazon dataset of 65.56%, Cipla at 45.91%, Eicher at 43.45%, Bata at 65.60%, and Bosh at 55.08% [9]. K-Nearest Neighbor (K-NN) is one of the most widely used classification algorithms in machine learning for implementation and modification. This algorithm is quite simple because some parameters use distance metrics and k values. The main goal of the K-NN algorithm is to find the nearest neighbors (plot points) in the data and the data set [10]. The K-NN algorithm also aims to classify objects according to an attribute and training data sample. However, K-NN has a drawback, namely the selection and determination of the value of k. Determining the value of k in the K-NN calculation greatly influences the accuracy of the results that will be obtained. The value of k in K-NN is the size of the neighbor or nearest neighbor, which is measured based on distance measurements. To avoid unwanted results, the value of k cannot be a multiple of the number of classes [10, 11].

K-NN is a non-parametric method used for classification and regression. Compared to other machine learning algorithms, K-NN is the simplest. Compared to other machine learning algorithms, K-NN is the simplest. This algorithm contains examples of K-closet training with space features. In this algorithm, K is a user-defined constant. Test data were classified by assigning the constant value with the highest chronicity among the K-training samples closest to that point. Literature shows that KNN has strong, consistent results [12].

Jayapandian proposes an optimization in the KNN application. The traditional Euclidean distance and cosine similarity based on the KNN approach is a modified KNN algorithm to avoid problems arising from encoding words into numeric vectors. The performance of KNN is improved by using index optimization [13]. The next research from Amit Khisor and Wilson applies the machine learning method using the K-NN algorithm. This research creates an IoT (Internet of Things) system. The model developed

using machine learning proposes to predict heart failure. This research aims to apply various machine-learning methods to the resulting data. The data collected includes heart rate, blood pressure, blood oxygen level, and other variables to assist in diagnosing heart disease. The collected data were analyzed using the K-NN method. The results of this study indicate that IoT and machine learning can significantly contribute to the diagnosis of heart failure. Patients can be monitored, and changes in cardiac condition can be detected earlier [14].

Based on the previous discussion, the author examines the most optimal k values based on the accuracy obtained based on calculations by testing different k values, namely 1, 3, 5, 7, and 9. The k value in the K-NN algorithm greatly influences the results to be obtained. Therefore, this study focuses on finding which accuracy value is the highest of the five k values to be used for the K-NN model. Testing the value of k will be carried out using 80% training data and 20% testing data, which will be used as a model. Heart failure will be the object of this study. In the heart failure dataset, two targets will be used: normal or indicated heart failure.

2. RESEARCH METHOD

This research was carried out following the framework shown in Figure 1. The research began by acquiring research data, then preprocessing the data, splitting the data, calculating the shortest distance using the Manhattan distance, testing the value of k, and finally analyzing the accuracy results.

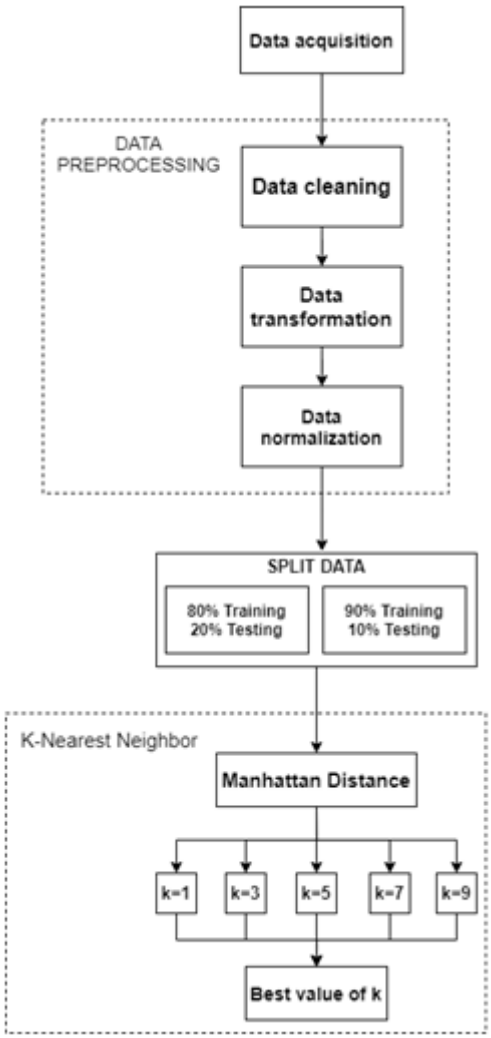


Figure 1. Research Framework

In Figure 1, The research framework is the stage for getting the best k value for this study. The stages will be passed as follows: First, data acquisition is done through Kaggle.com. Second is the preprocessing of the data in the dataset to make it easier for the data to be processed and to avoid errors in the data when the data is executed. The preprocessing is carried out in three stages: data cleaning, transformation, and normalization. Cleaning data is the initial process of the data preprocessing stage. Cleaning data functions to check the dataset. If the dataset contains noisy and missing values, then the dataset will not have maximum accuracy when processed. Data transformation is an advanced process of data cleaning. Transformation data serves to change the scale of data measurement into another form. If the data is numeric, it will be aligned with numeric data. Data normalization has three methods that can be used, but at this stage, only one normalization method is used: simple feature scaling. Normalization is used to help the training process. If there are very large range differences between the numeric variables, the variable with the highest magnitude may dominate the model. Third, after preprocessing the data, the data is split. Split data is divided into two scenarios: the first divide 80% of the training data and 20% of the testing data. The second scenario divides 90% of the training data and 10% of the testing data. Fourth, after splitting the data, the next step is calculating the distance between the testing and training data using the Manhattan distance. Manhattan distance measures the closeness distance between attributes, which will later be used as a model for testing the value of k. The last process is testing the value of k. This test is carried out individually, starting from k = 1, k = 3, k = 5, k = 7, and k = 9. After testing, it will be determined which value of k is the best for use in this study.

2.1. Data Acquisition

The dataset used in this study comes from Kaggle.com. The data obtained has been widely used in previous studies. The dataset, which totals 918 pieces, will be divided into two parts: training and testing data. The training data (training data) that will be used is 734 data points, and the testing data (test data) is 184 data points. The training data trains and builds a model, while the test data tests the model after the complete training process.

Table 1. Dataset

Attribute ID	Attribute Name	Attribute Description
1	Age	age of the patient [years]
2	Sex	sex of the patient [M: Male, F: Female]
3	ChestPainType	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4	RestingBP	resting blood pressure [mm Hg]
5	Cholesterol	serum cholesterol [mm/dl]
6	FastingBS	fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
7	RestingECG	resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8	MaxHR	maximum heart rate achieved [Numeric value between 60 and 202]
9	ExerciseAngina	exercise-induced angina [Y: Yes, N: No]
10	Oldpeak	ST [Numeric value measured in depression]
11	ST_Slope	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12	HeartDisease	output class [1: heart disease, 0: Normal]

In Table 1, the first attribute is a person's age (2977 years). The second attribute describes a person's gender ("0" for women and "1" for men). The third attribute defines the level of chest pain experienced by patients in the hospital. Four types of chest pain are converted into numerical values; each value describes the level of chest pain (TA: 0, ATA: 1, NAP: 2, ASY: 3). The fourth attribute describes the results of a person's blood pressure. The fifth attribute indicates a person's cholesterol level. The sixth attribute describes a person's fasting blood sugar level (1 if blood sugar is ¿ 120 mg/dl, 0 otherwise). The seventh attribute shows EKG results from 0 to 2, where each value indicates the severity of pain. The eighth attribute is the maximum heart rate value (minimum: 71, maximum: 202). The ninth attribute is used to understand whether exercise causes angina or not (yes: 1, no: 0). The tenth attribute defines a person's depression status. The eleventh attribute describes the slope of the peak training ST segment (Up: upsloping, Flat: flat, Down: downsloping). The final data is a collection of classes or labels describing the number of categories in a dataset. This dataset uses class binary; 0 means there is no possibility of heart failure in a person, while 1 implies a strong possibility of someone having heart failure [15].

2.2. Preprocessing Data

Preprocessing data is the first step in creating machine learning and artificial intelligence models. This process transforms data into an easier and more efficient form, making it possible for machine learning models to produce more accurate results [6, 16]. In this study, three data preprocessing stages were used: first, data cleaning is an initial process carried out in data preprocessing. Cleaning data is used to select and delete data that can reduce the accuracy of the machine-learning model. In this study, there were no noisy data or missing values. Noisy data is data that contains wrong or abnormal values; this condition is called a data anomaly [16]. Second, data transformation is a function to equalize all data, such as by equating data structures, data formats, or values in data to produce the appropriate data set. Finally, data normalization is a technique for converting data into a regular scale. A process in which several variables have the same range of values, not too large or too small, to facilitate analysis [16]. Normalization of the data has three normalization methods, but in this study, we used one data normalization method, namely simple feature scaling. This method is a simple normalization method that divides each value by the maximum value on the attribute. The formula used in simple feature scaling is contained in Formula (1)

$$x_{new} = \frac{x_{old}}{x_{max}} \tag{1}$$

x_{old} is the value of each attribute in the dataset, x_{max} is the maximum value for each attribute in the dataset, and the dataset x_{new} is the normalized value.

2.3. Split Data

The process of making a model and testing it to get the best K results is done by making a model using training data and testing the model with data testing [17]. The distribution of training data and data testing in this study was done manually by dividing 80% of the data used for training by 20% of the data used for testing in the first scenario. Split the data manually using the percentage formula contained in Formulas (2) and (3).

$$\text{Data Training} = 918 * \frac{80}{100} = 734data \tag{2}$$

The above formula is the manual division of the training data into 80% of the dataset, which will be used as training or model training data.

$$\text{Data Testing} = 918 * \frac{20}{100} = 184data \tag{3}$$

The above formula is a manual division of testing data, with 20% of the dataset used as testing data.
In the second scenario, manually divide the data into 90% training data and 10% testing data using Formulas (4) and (5).

$$\text{Data Training} = 918 * \frac{90}{100} = 826data \tag{4}$$

The above formula is the manual division of the training data into 90% of the dataset, which will be used as training or model training data.

$$\text{Data Training} = 918 * \frac{10}{100} = 92data \tag{5}$$

The above formula is a manual division of testing data, with 10% of the dataset for testing data.

2.4. Manhattan Distance

Manhattan distance is one of the distance measurement methods used in the K-NN algorithm. Manhattan distance is the distance between two points in three-dimensional space, calculated by adding the absolute difference of the x and y coordinates between the two points. The Manhattan Formula distance can be seen in Formula (6).

$$d(x, y) = \Sigma_{i=1}^{n} \mid xi - yi \mid \tag{6}$$

Based on equation (6), calculate the distance between the training data and the testing data, where the xi symbol represents the testing data, the yi symbol represents the training data, and the i symbol represents the data variable.

The distance calculation used in this study calculates the distance values that exist in training and testing data attributes. The attributes used in measuring this distance are age, sex, chest pain type, resting BP, cholesterol, fasting BP, resting ECG, maxhr, exercise angina, old peak, and st_slope.

2.5. K-Nearest Neighbors (K-NN)

The K-NN algorithm is a supervised learning classification algorithm that can be implemented on labeled data. K-NN classifies the dependent variable based on how similar the independent variable is to an example similar to known data [18].

The K-NN algorithm is one of the most widely used classification and learning algorithms for implementation and modification, and this algorithm is fairly simple because some parameters work using distance metrics and k values. The main goal of the K-NN algorithm is to find the nearest neighbors (plot points) in the data and the data set [19]. K-NN classifies sample points with the most or majority values for its neighbors. K-NN uses a dataset where data points are divided into several classes to predict the classification of new sample points. This algorithm determines the distance between all data queries and selects a specific number close to the query value, then selects a repeating label in the classification case or the average label in the regression case [10].

The k-NN (k Nearest Neighbors) algorithm is a classification algorithm based on learning from previously classified data. This algorithm includes supervised learning where the results of the new instance are classified based on the majority of the nearest neighbor distance from the existing class. Meanwhile, Zhang explained that the k-NN (K-Nearest Neighbors) algorithm is a non-parametric or instance-based method and is considered one of the simplest methods in data mining and machine learning [20]. Testing the value of k, which is made using Manhattan distance measurements after the data preprocessing process and data split, the best and most optimal k value will be selected from the k values that have been determined [21]. Based on the results of the highest accuracy of all k values, the most optimal k value will be obtained for use in this study.

3. RESULTS AND ANALYSIS

Following the research framework shown in Figure 1, the following will present the research results, starting from the data preprocessing stage to applying the K-NN algorithm to classify data on patients with potential for heart disease and patients with normal hearts. To get the best classification model, this study will test several different k values.

3.1. Implementation of Preprocessing Data

Data preprocessing is carried out in three stages: cleaning, transformation, and normalization. At the data cleaning stage, the dataset is free of noise and missing values. The dataset used is 918 data points with 12 attributes.

Table 2. Originals Dataset

Age	Sex	CP	RBP	Cho	FBS	RECG	MaxHR	ExAngina	Oldpeak	St_Slope
40	M	ATA	140	289	0	Normal	172	N	0	Up
49	F	NAP	160	180	0	Normal	156	N	1	Flat
37	M	ATA	130	283	0	ST	98	N	0	Up
48	F	ASY	138	214	0	Normal	108	Y	1,5	Flat
54	M	NAP	150	195	0	Normal	122	N	0	Up
...	...	...	...	...	...	...	...	...	...	...
38	M	NAP	138	175	0	Normal	173	N	0	Up

Table 2 is the original dataset that has not been preprocessed. The dataset needs to be preprocessed so that the results obtained produce better accuracy values.

1. Transformation Data

The data transformation carried out at this stage is to equalize or align the values in the attributes to become the same. In the dataset used in this study, attribute values are in the form of categories, so these values are difficult to process. Therefore, attribute values in the form of categories are converted into numeric form. There are five attributes whose values are converted into the numeric form: Sex, ChestPainType, RestingECG, ExerciseAngina, and St_Slope. The normalization results are presented in table form, which can be seen below.

Table 3. Transformation of Sex Data attribute

Category	Initialization
M	0
F	1

Table 3 describes the data transformation results, where the categorical M (male) and F (Female) data are converted into numeric forms 0 and 1.

Table 4. Transformation of Data ChestPainType

Category	**Initialization**
ASY	0
ATA	1
NAP	2
TA	3

Table 4 describes the results of the data transformation, where the ASY (Asymptomatic), ATA (Atypical Angina), NAP (Non-Anginal Pain), and TA (Typical Angina) data which are in the form of categories, are converted into numeric forms 0, 1, 2 and 3.

Table 5. Transformation Data RestingECG

Category	**Initialization**
Normal	0
ST	1
LVH	2

Table 5 describes the data transformation results, where the Normal, ST, and LVH data, which are in the form of categories, are converted into numeric forms 0, 1, and 2.

Table 6. Transformation Data ExerciseAngina

Category	**Initialization**
N	0
Y	1

Table 6 describes the data transformation results, where data N (No) and Y (Yes) in categories are converted into numeric forms 0 and 1.

Table 7. Transformation Data St_Slope

Category	**Initialization**
Down	0
Flat	1
Up	2

Table 7 describes the data transformation results, where Down, Flat, and Up, which are in the form of categories, are changed to numeric forms 0, 1, and 2.

Table 8. Transformed Dataset

Age	**Sex**	**CP**	**RBP**	**Cho**	**FBS**	**RECG**	**MaxHR**	**ExAngina**	**Oldpeak**	**St_Slope**
40	0	1	140	289	0	0	172	0	0	2
49	1	2	160	180	0	0	156	0	1	1
37	0	1	130	283	0	1	98	0	0	2
48	1	0	138	214	0	0	108	1	1,5	1
54	0	2	150	195	0	0	122	0	0	2
...	...	...	...	...	...	...	...	...	...	...
38	0	2	138	175	0	0	173	0	0	2

Table 8 explains that the values for each attribute result from transforming data that has changed from category to numeric form 0, 1, and 2.

2. Normalization Data

In the normalization process carried out using the simple feature scaling normalization method, normalization results are obtained on a scale of 0 to 1. Numerical normalization can help the learning process if there is a very large range difference between numeric variables because the variable with the highest magnitude can dominate the model, regardless of whether the features are informative with respect to the target or not. The normalization results can be seen in more detail in Table 7data normalization results.

Table 9. Data Normalization Result

Age	Sex	CP	RBP	Cho	FBS	RECG	MaxHR	ExAngina	Oldpeak	St_Slope
0.51	0	0.33	0.70	0.47	0	0	0.85	0	0	1
0.63	1	0.66	0.80	0.29	0	0	0.77	0	0.16	0.50
0.48	0	0.33	0.65	0.46	0	0.5	0.48	0	0	1
0.62	1	0	0.69	0.35	0	0	0.53	1	0.24	0.50
0.70	0	0.66	0.75	0.32	0	0	0.60	0	0	1
…	…	…	…	…	…	…	…	…	…	…
0.49	0	0.66	0.69	0.29	0	0	0.85	0	0	1

In Table 9, it can be explained that the values in each attribute are the result of normalizing the data that has been calculated using the simple feature scaling method.

3.2. Best value of k

The classification process divides the data into two stages: the first stage uses split data with 80% data training and 20% data testing; the second stage uses split data with 90% data training and 10% data testing. Furthermore, in the classification process using the Manhattan distance measurement method, the Manhattan distance is used to find the closest neighbor value of each piece of data to be classified by measuring the distance between the training data and the testing data. The training data with the closest distance to the data to be classified will be selected as the nearest neighbor, and the number of selected neighbors will be adjusted to a predetermined k value. The calculation of the Manhattan distance from data that has been normalized using the simple feature scale method is carried out using the equation formula. The following is an example of calculating Manhattan distances, and the results are presented in Table 10 using the values k = 1, k = 3, k = 5, k = 7, and k = 9.

Table 10. The Result of Calculating the Distance Using The Manhattan

Age	Sex	CP	RBP	Cho	FBS	RECG	MaxHR	ExAngina	Oldpeak	St_Slope	Heart Disease	MD
0.68	1	0.33	0.7	0.35	0	0	0.7	1	0.32	0.5	1	0.01
0.75	0	0	0.75	0.44	0	1	0.54	1	0.12	1	1	0.01
0.7	0	0	0.65	1	1	0	0.61	1	0.16	0.5	1	0.02
0.76	0	1	0.8	0.45	0	1	0.61	0	0	1	1	0.02
0.81	0	0	0.7	0	1	1	0.73	0	0.32	1	1	0.04
0.7	1	0	0.63	0.55	1	0,5	0.76	0	0	0.5	1	0.04
0.77	1	0	0.75	0.42	0	1	0.77	0	0.41	0.5	1	0.04
0.51	0	1	0.7	0.33	0	0	0.88	1	0.22	1	1	0.04
0.74	0	0	0.65	0.35	0	0	0.7	1	0.32	0.5	1	0.05
…	…	…	…	…	…	…	…	…	…	…	…	…

Table 10 results from distance measurements using the Manhattan distance calculated from normalized simple feature scale data. The calculation of the Manhattan distance value is sorted based on the nearest neighbor of the k value. The value of k = 1 is obtained from the Manhattan distance, the smallest distance between the training and test data. An example of determining the value of k1 = 1, k3 = 1, k5 = 1, k7 = 1, and k9 = 1 Calculation of the Manhattan distance where k1 produces a value of 1 obtained from the results of the nearest neighbor of the class, namely 1. k3 equals 1 obtained from the results of the nearest neighbor on data 1, 2, and 3. k5 equals 1 obtained from the results of the nearest neighbors on data 1, 2, 3, 4, and 5. k7 equals 1 obtained from the results of the nearest neighbors on data 1, 2, 3, 4, 5, 6, and 7. k9 is equal to 1 and is obtained from the results of the nearest neighbors in data 1, 2,

3, 4, 5, 6, 7, 8, and 9. After getting the values k1 = 1, k3 = 1, k5 = 1, k = 7, and k = 9, we calculate the accuracy. More clear results will be presented in Table 11.

Table 11. Class and Manhattan Distance

Data	Heart Disease (Class)	Manhattan Distance (MD)
1	1	0.01
2	1	0.01
3	1	0.02
4	1	0.02
5	1	0.04
6	1	0.04
7	1	0.04
8	1	0.04
9	1	0.05

If the data has class 1 (diagnosed with heart failure) and k1 has the same class, then the value is TRUE. Class k3, k5, k7, and k9 have the same value, namely 1 (diagnosed with heart failure); therefore, it is TRUE. This testing process is carried out from the first data test to the 184th data test. If the k value is the same as the class value, the result is TRUE, and vice versa. If the k value is not the same as the class value, then the result is FALSE. Calculation of the accuracy of split data 80% training data and 20% testing data can be seen below:

Based on 184 testing data points, 155 data points with a true value or the same as the label was divided by the total amount of testing data. Then, the accuracy for $k = 1$ is:

$$Accuracy(\%) = \frac{155}{184} * 100\% = 84\%$$

Based on 184 testing data points, 158 data points with a true value or the same as the label was divided by the total amount of testing data. Then, the accuracy for $k = 3$ is:

$$Accuracy(\%) = \frac{158}{184} * 100\% = 85\%$$

Based on 184 testing data points, 164 data points with a true value or the same as the label was divided by the total amount of testing data. Then, the accuracy for $k = 5$ is:

$$Accuracy(\%) = \frac{164}{184} * 100\% = 86\%$$

Based on 184 testing data points, 164 data points with a true value or the same as the label was divided by the total amount of testing data. Then, the accuracy for $k = 7$ is:

$$Accuracy(\%) = \frac{156}{184} * 100\% = 86\%$$

Based on 184 testing data points, 156 data points with a true value or the same as the label was divided by the total amount of testing data. Then, the accuracy for $k = 9$ is:

$$Accuracy(\%) = \frac{156}{184} * 100\% = 84\%$$

This stage is the result of testing carried out to determine the most optimal k value from the split data results of 80% training data and 20% testing data. The final test results are presented in tabular form, as shown in Table 12result accuracy.

Table 12. Accuracy of Result

k value	Accuracy (%)
1	84
3	85
5	86
7	86
9	84

In Table 12, it can be explained that the value of k in the third and fourth columns is the most optimal result for each accuracy. The value of k = 1 gets 84% accuracy, the value of k = 3 gets 85% accuracy, the value of k = 5 gets 86% accuracy, the value of k = 7 gets 86% accuracy, and the value of k = 9 gets 84% accuracy.

Next are the results of the tests carried out to determine the most optimal k value from the split data results of 90% training data and 10% testing data. The final test results are presented in tabular form, as shown in Table 3result Accuracy.

Table 13. Accuracy of Result

k value	Accuracy (%)
1	0.86
3	0.87
5	0.87
7	0.88
9	0.88

In Table 13, it can be explained that the value of k in the fourth and fifth columns is the most optimal result for each accuracy. The value of k = 1 gets 86% accuracy, the value of k = 3 gets 87% accuracy, the value of k = 5 gets 87% accuracy, the value of k = 7 gets 88% accuracy, and the value of k = 9 gets 88% accuracy.

This K-NN research method produces different accuracy values. Based on the results of the accuracy obtained from testing 80% of training data and 20% of data, the highest accuracy results were found to be 86%, and testing 90% of training data and 20% of testing data obtained the highest accuracy results of 88%. Other normalization methods can be used to produce maximum accuracy values, such as Min-Max, Z-Score, or Decimal Scale. The normalization process has quite an effect on the resulting accuracy value. Increasing the accuracy value can also divide the data into 70% training data and 30% test data or 60% training data and 40% test data.

Table 14. Comparison of This Study with Existing Research

Title and Author	Publication Year	Dataset	Methodology	Accuracy (%)
Diabetes analysis and prediction using random forest, KNN, Naive Bayes, and J48: An ensemble approach (Minyechil Alehegn, Rahul Raghvendra Joshi, Preeti Mulay)	2019	Dataset From UCI Machine Learning PIDD	K-NN	74.59
Atherosclerosis disease prediction using Supervised Machine Learning Techniques (Oumaima Terrada, Bouchaib Cherradi, Abdelhadi Raihani, Omar Bouattane)	2020	Dataset from Cleveland Clinic Foundation dataset at the University of California Irvine	K-NN	80
A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction (Indu Kumar, Kiran Dogra, Chetna Utreja, Premlata Yadav)	2021	Dataset from UCI repository dataset, Heart Disease	K-NN	77.04
This Study	2023	Dataset from Kaggle, Heart Failure Prediction Dataset	K-NN	86

Table 14 is a comparison of the research with previous studies. Based on the research that has been done, this research has a very low accuracy value, namely 86% and 88%. Accuracy can be further improved by using other classification methods from machine learning, such as the SVM method, Decision Tree, random forest, and the like. Distance measurement methods such as Canberra, Euclidean, and others can also be used and tried to improve accuracy values. As can be seen from previous research, the use of the K-NN algorithm can be used not only in the health sector but also in other fields, such as stock trends.

4. CONCLUSION

Based on the results of research conducted on the heart failure dataset that has been done, it can be concluded that the application of the K-Nearest Neighbor algorithm to classify heart failure as a target attribute has a value of 0 and 1, where 0 is normal or does not have the potential for heart failure and 1 has the potential for heart failure. The k value most recommended for use and has the best accuracy is k = 7 and k = 9 in testing 90% training and 10% data, producing the highest accuracy of 88%. This accuracy is obtained by manually dividing the dataset, or split data, with 90% testing data from 92 data points and 80% training data from 826 data points. The testing process is done by normalizing the dataset using simple feature scaling and the Manhattan distance calculation method. In this study, the accuracy results obtained did not change significantly. Suggestions for further research can be added to or compared with other machine learning methods and can use other normalization methods to get even better accuracy results.

5. ACKNOWLEDGEMENTS

The author would like to thank profusely all parties who have provided assistance in writing this research paper. The author also thanks the Directorate General of Higher Education, Research and Technology (DRTPM), Ministry of Education, Culture, Research and Technology (Kemendikbudristek) of the Republic of Indonesia for funding this research in the Master's Thesis Research (PTM) scheme with contract No. 0423.11/LL5-INT/AL.04/2023 and subcontract No. 054/PPS-PTM/LPPM UAD/VI/2023.

6. DECLARATION

CONTRIBUTION AUTHOR

Alya Masitha is responsible for compiling and designing studies, collecting, analyzing, and interpreting data, and preparing articles. Muhammad Kunta Biddinika supervised the project and the drafting of the manuscript, and Herman is responsible for making critical revisions to articles and giving final approval to the final version to be published.

FUNDING STATEMENT

This research was funded by the Directorate General of Higher Education, Research and Technology (DRTPM), Ministry of Education, Culture, Research and Technology (Kemendikbudristek) of the Republic of Indonesia for funding this research in the Master's Thesis Research (PTM) Scheme with contract No. 0423.11/LL5-INT/AL.04/2023 and subcontract No. 054/PPS-PTM/LPPM UAD/VI/2023.

COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] P. Mamatha Alex and S. P. Shaji, "Prediction and diagnosis of heart disease patients using data mining technique," *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019*, pp. 848–852, 2019.

[2] B. Rahman, H. L. Hendric Spits Warnars, B. Subirosa Sabarguna, and W. Budiharto, "Heart Disease Classification Model Using K-Nearest Neighbor Algorithm," *2021 6th International Conference on Informatics and Computing, ICIC 2021*, 2021.

[3] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 2020.

[4] H. Agrawal, J. Chandiwala, S. Agrawal, and Y. Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," *2021 International Conference on Intelligent Technologies, CONIT 2021*, 2021.

[5] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, vol. 2018-Febru, pp. 1–5, 2018.

[6] R. Yunus, U. Ulfa, and M. D. Safitri, "Application of the K-Nearest Neighbors (K-NN) Algorithm for Classification of Heart Failure," *Journal of Applied Intelligent System*, vol. 6, no. 1, pp. 1–9, 2021.

[7] M. Alehegn, R. R. Joshi, and P. Mulay, "Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach," *International Journal of Scientific and Technology Research*, vol. 8, no. 9, pp. 1346–1354, 2019.

[8] O. Terrada, B. Cherradi, A. Raihani, and O. Bouattane, "Atherosclerosis disease prediction using Supervised Machine Learning Techniques," *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2020*, 2020.

[9] I. Kumar, K. Dogra, C. Utreja, and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction," *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, no. Icicct, pp. 1003–1007, 2018.

[10] A. Upadhyay, S. Nadar, and R. Jadhav, "Comparative study of SVM & KNN for signature verification," *Journal of Statistics and Management Systems*, vol. 23, no. 2, pp. 191–198, 2020.

[11] A. Almomany, W. R. Ayyad, and A. Jarrah, "Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3815–3827, 2022.

[12] Z. Ma, Z., Li, H., & Wang, "Early detection of heart failure using k-nearest neighbors algorithm with feature selection," *International Journal of Advanced Robotic Systems*, vol. 17, no. 4, 2020.

[13] N. Jayapandian, C. P., & Sundararajan, "A comparative analysis of K-nearest neighbors and support vector machine algorithms for heart disease prediction," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 10, pp. 4157–4167, 2019.

[14] A. Kishor and W. Jeberson, "Diagnosis of Heart Disease Using Internet of Things and Machine Learning Algorithms," *Lecture Notes in Networks and Systems*, vol. 203 LNNS, pp. 691–702, 2021.

[15] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[16] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Computational and Mathematical Organization Theory*, vol. 25, no. 3, pp. 319–335, 2019.

[17] A. Murugan, S. A. H. Nair, and K. P. Kumar, "Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers," *Journal of Medical Systems*, vol. 43, no. 8, 2019.

[18] G. S. Reddy Thummala and R. Baskar, "Prediction of Heart Disease using Decision Tree in Comparison with KNN to Improve Accuracy," pp. 1–5, 2022.

[19] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, 2019.

[20] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, "Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets," *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers,Series A*, vol. 43, no. 1, pp. 80–92, 2020.

[21] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, 2022.