Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

Evaluation and Comparison of K-Nearest Neighbors Algorithm Models for Heart Failure Prediction

Alya Masitha, Nurul Huda*, Deden Istiawan, Lucky Nur Rohman Firdaus

Department of Software Engineering, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang, Indonesia
Email: [1]alya.masitha@itesa.ac.id, [2,*]nurul.huda@itesa.ac.id, [3]deden.istiawan@itesa.ac.id, [4]rohmanfirdaus11@gmail.com
Correspondence Author Email: nurul.huda@itesa.ac.id
Submitted: **17/09/2024**; Accepted: **01/12/2024**; Published**: 03/12/2024**

Abstract−Heart failure is a disease that is one of the most crucial in the world. Researchers have used several machine learning techniques to assist health professionals in the diagnosis of heart failure. K-NN is a technique of supervised learning algorithm that has been successfully used in terms of classification. However, using the K-NN algorithm has stages in terms of data analysis. The data used must also be processed in such a way that it becomes data that is easier to analyse and that the results obtained are also more accurate. Data pre-processing involves transforming raw data into a format that is appropriate for the model. The normalization technique is one of the techniques contained in pre-processing. This research uses two normalization techniques, namely the simple feature scale and min-max. The purpose of this study is to compare the performance of the KNN model to obtain an optimal prediction model. This study contributes to producing a heart failure prediction model based on the K-Nearest Neighbors (KNN) algorithm that can be optimized to improve the accuracy of early detection, so that it can help medical personnel in making more appropriate clinical decisions. The results obtained from this research show that the dataset that uses the min-max normalization method is better than data that is not normalized and data that uses simple feature scale normalization. The highest level of accuracy was achieved by employing the min-max normalisation technique, with a value of K=9, resulting in an accuracy rate of 85.05%.

Keywords: K-NN; Normalization; Min-Max; Simple Feature Scale; Heart failure

1. INTRODUCTION

Heart failure is a severe condition often referred to as cardiovascular disease and is potentially life-threatening. This disease is characterized by the inability of the heart to pump blood effectively and is feared by millions of people around the world [1], [2]. Statistics show that heart failure is one of the leading causes of death and claims more lives than lung cancer [3]. The primary factors contributing to heart failure are coronary heart disease, hypertension, and anomalies in heart valves. Common symptoms include shortness of breath, fatigue and peripheral edema. Precise diagnosis and suitable therapy are necessary to avert severe consequences and enhance the patient's quality of life in this illness [4]. Humans certainly do not want their heart organs to have problems in order to experience long life and reduce mortality, therefore early anticipation is needed regarding heart failure. To be able to predict heart failure, several tests are needed. Lack of expertise from medical staff can result in wrong predictions. Various models are applied to predict the risk of heart attack, the probability of heart attack risk is displayed through a website. Another model used is Support vector machine (SVM) which achieved an accuracy value of 85.7%. The accuracy value was obtained from several attributes analyzed in the form of cholesterol, blood pressure and blood sugar [5].

Technological advances also play a significant role in these efforts. In the 21st century, technology is developing rapidly and has had many positive impacts in various fields, including health. This development is based on human innovation and creativity in creating new things. Diagnosing heart failure is like uncovering a hidden cloak and revealing a disease that may not show obvious symptoms. Doctors rely on various methods to diagnose heart failure, such as: physical tests that check for physical signs such as leg swelling, fatigue and shortness of breath [6]. Blood tests measure levels of troponin, a protein released when the heart is damaged. Imaging tests visualize the structure and function of the heart, such as chest X-ray, echocardiogram and cardiac MRI. Data from these various sources are used to assess heart function and identify the underlying cause of a patient's symptoms. Early detection of heart failure enables timely intervention and implementation of suitable medical approaches, which can effectively impede disease advancement, alleviate symptoms, and enhance patients' quality of life [7]. Technological developments in healthcare, particularly in data analytics and machine learning, have opened up new opportunities in heart failure prediction [6]. Using available clinical data, machine learning algorithms can be trained to recognize patterns that may not be visible to conventional clinicians, thus providing more accurate predictions of heart failure risk in patients [8], [9].

The k-Nearest Neighbours (k-NN) algorithm is widely recognised in the field of machine learning as a straightforward yet remarkably powerful classification approach [10]. Originating from the field of pattern recognition, k-NN has gained widespread popularity due to its intuitive approach and robust performance in a wide range of applications [11]. The k-NN algorithm is a non-parametric and slow learning method. Non-parametric refers to a statistical method that does not rely on any assumptions about the distribution of the data. Lazy learning refers to a method where an explicit model is not learned during the training phase. Instead, the k-NN algorithm saves the complete training dataset and generates predictions only when it is executed [12].

The k-Nearest Neighbors (k-NN) algorithm has become a common choice in various prediction applications, including in the context of disease prediction. Research has shown that the implementation of the k-NN algorithm can

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

aid in the classification of heart disease with a significant degree of accuracy [13]. In addition, another study also highlighted the role of Artificial Intelligence (AI) in detecting heart failure, with a comparison of prediction accuracy between Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest (RF) methods [14]. Although there are various other algorithms used in disease prediction, such as Naive Bayes, Neural Network, and Decision Tree, k-NN remains one of the effective choices in handling heart disease classification cases [15], [16]. A separate study demonstrated that the k-NN method outperforms the Naïve Bayes algorithm in accurately classifying liver illness [17].

Processing medical data for heart failure prediction is a very important step. Medical data often contains missing values, outliers, and variables with different scales [2]. Preprocessing includes steps such as imputation of missing data, normalization, and outlier detection and removal [18]. Patient data needs to be processed and prepared before k-NN can predict heart failure [19]. This process involves data collection, which is gathering patient data from various sources, such as medical records, diagnostic tests and medical history. Data cleaning is identifying and dealing with missing, inaccurate or inconsistent data. Data preprocessing is converting data to a format suitable for analysis, such as normalization and data transformation. Good preprocessing improves data quality and ensures *machine learning* models can learn effectively from the data [20], [21].

Pre-processing is an important early stage in research. The purpose of pre-processing is to prepare raw data for processing and analysis. Data pre-processing methods are available in various ways, one of which involves normalising data through the use of normalisation procedures [22]. Data normalization is used to remove biased values and ensure that the values have an impact on the learning process. This research compares normalized and non-normalized heart failure datasets [23], [24]. The normalization techniques used in this study are *simple feature scale* and min-max. Data that has been normalized will be used to model the K-NN algorithm.

The objective of this study is to investigate preprocessing approaches applied to data using the normalisation method. Pre-processing techniques are employed to construct a K-NN model utilising heart failure datasets. The dataset used in this research is numeric and categorical. This research will test and compare the results of the K-NN algorithm that uses the normalization method and those that do not use normalization. The normalization methods to be compared are simple feature scale and min-max.

2. RESEARCH METHODOLOGY

This study aims to develop a reliable predictive model to analyze and predict the risk of heart failure using the K-Nearest Neighbors (K-NN) algorithm [25]. This algorithm was chosen because of its ability to perform nearest neighbor pattern-based classification, which allows for effective identification of complex relationships between variables in the dataset. K-NN is known as a non-parametric algorithm that works based on data similarity, and in the context of heart failure prediction, it provides an advantage in understanding risk patterns that may not be visible with conventional methods. In addition, this algorithm offers flexibility in dealing with datasets with non-uniform distributions, something that is often encountered in medical data such as heart failure cases. The research process will be carried out in several main stages, starting with the data acquisition stage. At this stage, relevant data on patients with a history of heart failure will be collected from various reliable sources, such as clinical databases, medical journals, and previous studies. This data includes various important features such as age, blood pressure, cholesterol levels, medical history, and other risk factors associated with heart failure. This data collection is carried out carefully to ensure that the dataset used is not only representative but also large and varied enough to support comprehensive analysis. After the data is collected, the preprocessing process will be carried out to clean the data from missing values, outliers, and convert the data into a format that can be processed by the machine learning algorithm. This stage includes techniques such as imputation of missing data, feature normalization, and converting categorical data into numeric form through encoding. The importance of this preprocessing stage cannot be ignored, because good data quality will greatly affect the accuracy and reliability of the predictive model to be built. Next, the processed dataset will be divided into training data and test data to ensure the model can be evaluated objectively. This process will be followed by the application of the K-NN algorithm, where parameters such as the number of nearest neighbors (k) will be optimized through cross-validation techniques to obtain the best performance. After the model is trained, its performance will be evaluated using evaluation metrics such as accuracy, precision, recall, and AUC-ROC to understand how well the model is in predicting the risk of heart failure. At the end of the study, the resulting predictive model will be compared with other commonly used models in heart failure prediction, such as Logistic Regression and Random Forest, to assess the relative superiority of the K-NN approach in this context. Visualization and model comparison results will be presented in detail to provide a comprehensive overview of the reliability of the predictions and insights gained from this study. A flowchart of the study showing each stage of the process, from data collection to model evaluation, can be seen in Figure 1, which provides a visual illustration of the methodology used and the interrelationships between stages. Through this approach, it is hoped that this study can contribute significantly to providing more accurate and reliable predictive models to assist in the diagnosis and management of heart failure risk in the future.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

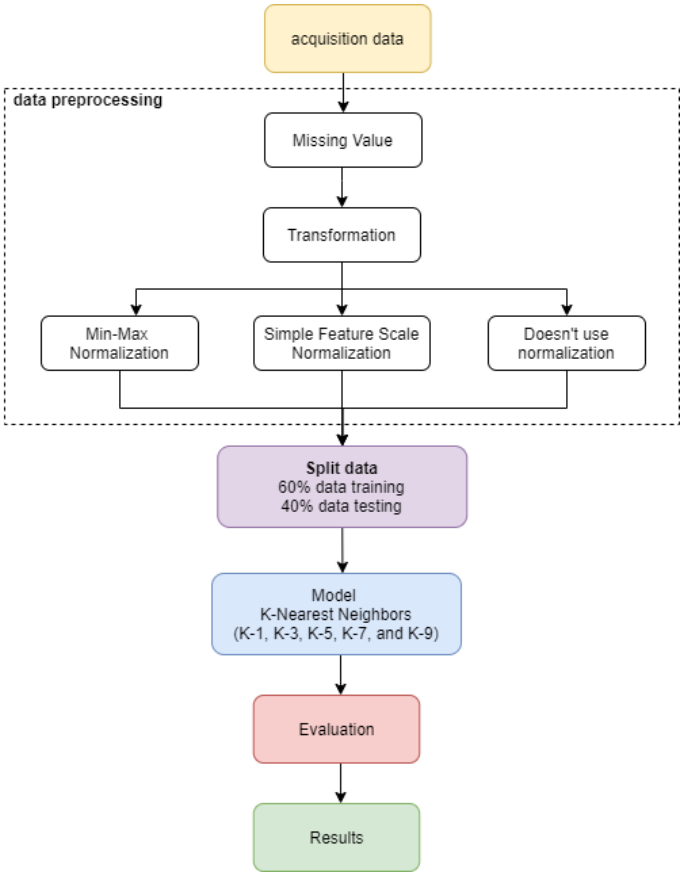

Figure 1. Research Framework

Figure 1 presents a research framework that illustrates the complete and systematic flow of the research process to be carried out, from data acquisition to final evaluation. The first stage of this research begins with data acquisition, where data related to heart failure disease is collected from various relevant sources. This data can come from clinical databases, medical journals, or previous studies. At this stage, all information obtained, both in numerical and categorical forms, is converted into a format that can be further analyzed. This process is very important because it ensures that the raw data collected is ready to be used in further analysis and minimizes the possibility of data errors due to inconsistent formats. After the data is collected, the next step is checking for missing values. This check is done to ensure that the dataset is free from problems caused by missing values or incomplete data. Missing values can affect the quality of the analysis and reduce the accuracy of the prediction model. Therefore, handling missing values is a critical step. Techniques such as imputation, where missing values are filled with the mean or mode of other similar data, will be applied if missing values are found in the dataset. With this check, the dataset used will be cleaner and more reliable, so that the research results are not disturbed by incomplete data. After the missing values check is complete, the process continues with data transformation, especially for categorical data. Categorical data such as gender, smoking status, and other medical statuses need to be converted into numeric format so that they can be processed by the K-NN algorithm. This transformation is done using techniques such as one-hot encoding or label encoding, where each category is represented as a numeric value. This process is important to prepare categorical data so that it can be analyzed mathematically by the machine learning model. The next step is the normalization experiment, where two normalization methods are applied to the data: Simple Feature Scale and Min-Max Normalization. In this stage, both normalization methods will be compared with the data without normalization. Simple Feature Scale converts each feature into standard deviation units, so that each variable has a mean of 0 and a standard deviation of 1, while Min-Max Normalization converts the feature values into the range of 0 to 1. This experiment was conducted to see the impact of normalization on the performance of the K-NN algorithm, considering that K-NN is highly dependent on the distance between data. Normalization ensures that all features have the same weight in the distance calculation, which is very important for distance-based algorithms such as K-NN. After normalization, the data is divided into two segments in the data split stage, namely training data and testing data. This division is done with a proportion of 60% of the data for model training and 40% of the data for model testing. The training data is used to build a prediction model with K-NN, while the testing data is used to evaluate the performance of the trained model. This data division is important to avoid overfitting, where the model may perform very well on the training data but fails to predict well on data that has never been seen before. The next stage is testing the model with the K-Nearest Neighbors (K-NN) algorithm using five different K values: K = 1, K = 3, K = 5, K = 7, and K = 9. Choosing the right K value is very important in the K-NN model, because K that is too small or too large can affect

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

the accuracy of the prediction. At this stage, each K value will be tested to assess the performance of the model in classifying data and predicting heart failure. After the model is tested, the next stage is the evaluation of the results. At this stage, the results of each experiment and process in the study are evaluated comprehensively. The evaluation is carried out to ensure that all steps in the research process, from data acquisition to algorithm implementation, have been carried out correctly and provide accurate results. Evaluation metrics such as accuracy, precision, recall, and F1-score will be used to assess the performance of the predictive model built. The final step in this research is the compilation and analysis of the overall results. At this stage, all the results obtained from model testing, comparison of normalization methods, and performance evaluation will be summarized to provide a comprehensive picture of the effectiveness of the approach used. These results will be compared with previous studies and discussed in the context of new findings obtained during the research process. This study is expected to provide in-depth insights into how data normalization and the selection of K values can affect the performance of the K-NN model in predicting the risk of heart failure, as well as provide recommendations for further research in the future.

2.1 Pre-processing

Preprocessing is the initial stage in the development of a machine learning model [26], [27]. Data transformation is employed to convert data into a streamlined and optimised format, enabling machine learning algorithms to generate more precise outcomes. This research uses Missing value, data transformation and data normalization. Missing value is used to check whether there is noise in the dataset that can cause poor accuracy or cannot be implemented into a machine learning model. Noise data refers to data that includes inaccurate or aberrant numbers, which is commonly referred to as data anomaly. Data transformation is a process that aims to standardise all data in order to facilitate data analysis. Transformation in this research is employed to convert categorical data into numerical data. The normalisation process is employed to transform data into a standardised scale. An analysis-friendly method using multiple variables with a range of values that is neither too large nor excessively tiny. Normalization techniques have a variety of methods, but in this research only uses two normalization methods. The normalization methods used are simple feature scale and min-max.

2.2 Missing Value

Missing values are situations where one or more observations in a dataset do not have a value recorded, either due to technical error, omission, or unavailability of data. The presence of missing values in a dataset is a common problem in data analysis processes, especially in the context of scientific research and machine learning applications, where data integrity and completeness are critical to producing valid and accurate results. The causes of missing values can vary widely, from human error in data collection or input, technical glitches during the data collection process, or situations where data is simply not available or relevant for a particular observation. For example, in a medical survey, patients may not answer certain questions for personal reasons or because the questions do not apply to their condition. If missing values are not handled properly, they can have serious consequences for the quality of the analysis. In some cases, missing values can cause distortion or bias in the results of a study. For example, if a group of patients with certain characteristics are more likely to have missing values, a prediction model built without considering this issue may produce invalid or less representative conclusions. Furthermore, statistical methods used to analyze data, such as regression or machine learning algorithms, often assume that the data used are complete. If missing values exist, these algorithms may make errors in the modeling process, either by ignoring observations containing missing values or by producing less accurate results. Therefore, an important first step is to identify missing values in the dataset. Checking for missing values can be done using a variety of methods, including using built-in functions in programming languages or statistical software that supports missing value detection. For example, in Python, libraries such as pandas provide functions such as **isnull()** or **isna()** that can help identify which cells have missing values. Once missing values are identified, the next step is to decide how to handle the problem. There are several common strategies used to handle missing values, depending on the nature of the data and how significant the missing values are.

2.3 Simple Feature Scale Normalization

Simple feature scale is one of the methods of normalization that is commonly used to analyze data. This method aims to scale the values of each variable in the dataset so that the values have a uniform range [0 to 1]. This method is one of the simple normalization methods. Equation (1) is used in the simple feature scale method.

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

$\boldsymbol{x_{old}}$ represents the value of each attribute in the dataset, $\boldsymbol{x_{max}}$ enotes the maximum value present in the dataset, and $\boldsymbol{x_{new}}$ signifies the outcome of the normalisation process.

2.4 Min-Max Normalization

The min-max normalisation approach is employed to rescale the value of each feature in the dataset, ensuring that the values are confined to a specific range, often between 0 and 1. Equation (2) is used in the min-max normalization method.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \quad (2)$$

$\boldsymbol{x_{old}}$ represents the value of each attribute in the dataset, $\boldsymbol{x_{min}}$ represents the minimum value present in the dataset, $\boldsymbol{x_{max}}$ represents the maximum value present in the dataset, and $\boldsymbol{x_{new}}$ represents the outcome of the normalisation process.

2.5 k-Nearest Neighbors (k-NN)

K-Nearest Neighbor is a machine learning algorithm. It is a data classification algorithm and is a lazy learning algorithm. The algorithm is a data categorization algorithm and falls under the category of lazy learning algorithms. The K-NN technique is employed to identify clusters of objects in the training data that are most similar to objects in the testing data. The calculation of distances between nearest neighbours is performed using the Euclidean distance technique. The equation for calculating the distance between points x and y using the Euclidean distance measurement [28], [29]. The computation of the distance method is illustrated in the equation (3).

$$Euclidean\ (x, y) = \sqrt{\Sigma_{i=1}^{n}\ (xi\ - yi)^2} \quad (3)$$

The use of euclidean distance is a calculation between two points. This method is a derivative of the pythagorean theorem formula, which is a square root calculation. Quantification of the spatial distance between points X and Y involves the utilization of various distance measurement functions. However, in research applications, the Euclidean distance metric is often preferred and used. Euclidean is related to the Phytagorean Theorem which is usually applied to higher dimensions.

2.6 *Accuracy*

Accuracy is a matrix that measures the extent to which the model can correctly predict the class of all data that has been tested using model testing. Accuracy, in the context of mathematical formulation, is determined by dividing the total number of true predictions (True Positives and True Negatives) by the total number of observations in the test dataset. While accuracy provides an overview of model proficiency especially when the distribution of positive and negative classes is balanced, there are other metrics that are also relevant in evaluating model performance. The accuracy is calculated using equation (4).

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (4)$$

3. RESULT AND DISCUSSION

The results of the research are a series of processes that have been carried out based on the research framework in Figure 1. The following will present the results of the research starting from the data acquisition stage to the evaluation results of the K-NN algorithm to classify data on potential heart disease patients and normal patients. This research has tested three different models using five K values, namely K1, K3, K5, K7 and K9.

3.1 Missing Value

Missing value is a stage of the process of cleaning data from noise or checking for missing data. Missing value is also a value that does not exist or is incomplete and missing value can affect the results of data analysis. Missing value handling is done by deleting and inputting data. The dataset utilised in this investigation did not identify any missing values.

```
data.isnull()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
913	False	False	False	False	False	False	False	False	False	False	False	False
914	False	False	False	False	False	False	False	False	False	False	False	False
915	False	False	False	False	False	False	False	False	False	False	False	False
916	False	False	False	False	False	False	False	False	False	False	False	False
917	False	False	False	False	False	False	False	False	False	False	False	False

Figure 2. Result of Missing Value

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

Figure 2 is a description of the research data that does not have missing values. The "False" output indicates that the attributes in the dataset are complete and there are no missing values, so this dataset does not need any additions to handle missing values. This research uses the pandas library with the "isnull()" method to detect whether or not there is a missing value in the data. If the resulting output is "True", then the dataset used has a missing value. The absence of missing values provides additional confidence that the data collected in this study has been processed properly and meets high quality standards. This is important, especially when datasets come from multiple sources that may differ in data quality and consistency. This study used data that was clean from missing values, which ensures that all relevant information has been included in the analysis, so that the prediction model developed based on this data will be more accurate and representative. In machine learning-based research like this, data quality is very important in determining the quality of the model results, so that researchers can be more confident in emitting and interpreting the results obtained from this dataset.

3.2 Evaluation Model

Model evaluation is an important stage carried out by analyzing accuracy results. The technique involves calculating the accuracy percentage by dividing the amount of correctly classified test data (true positive and true negative) by the total quantity of test data, and then multiplying the result by 100%. This method is in accordance with the accuracy formula that has been attached to Formula 4. This evaluation aims to determine how well the model predicts the data correctly, thus providing a clear picture of the overall model performance.

This stage is the result of parsing the dataset that has been tested using data preprocessing before the calculation process using the K-NN algorithm at this stage, a comparison of data normalization is carried out using three different methods. First, the data is normalized using the min-max method which scales the data values into a certain range. Second, the data is normalized using the simple feature scale method which adjusts the data feature scale simply. Third, an analysis is conducted on data that does not use normalization as a comparison. This comparison aims to evaluate the effect of each normalization method on the accuracy and performance of the K-NN algorithm, so that it can be determined which normalization method is most effective to use on the dataset.

Table 1. Dataset after Min-Max Normalization Process

Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Atr10	Atr11
0.244	0	0.333	0.700	0.479	0	0	0.788	0	0.295	1.000
0.428	1	0.666	0.800	0.298	0	0	0.676	0	0.409	0.500
0.183	0	0.333	0.650	0.469	0	0.5	0.267	0	0.295	1.00
0.408	1	0.000	0.690	0.354	0	0	0.338	1	0.465	0.500
0.530	0	0.666	0.750	0.323	0	0	0.436	0	0.295	1.00
…	…	…	…	…	…	…	…	…	…	…

The results of data calculation using the min-max method are presented in Table 1, clearly demonstrating the rescaling of each data attribute into a range of 0 to 1. The min-max normalization method is very popular in data analysis and machine learning because of its simplicity and effectiveness in handling different scales of data. The min-max normalization process works by shifting and scaling the original data so that the minimum value of each attribute becomes 0 and the maximum value becomes 1. The basic formula used is in formula (2). This normalization aims to equalize the scale of the various attributes in the dataset so that no attribute dominates or has a greater influence simply because of its wider range of values. Thus, each attribute will contribute proportionally to the analysis or machine learning model used. Attributes that have different scales will dominate the distance calculation, thus reducing the accuracy of the model. In addition, min-max normalization also helps in speeding up the training process of machine learning models, as data within a uniform range of values allows for more efficient optimization.

Table 2. Dataset after Min-Max Simple Feature Scale Normalization Process

Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Atr10	Atr11
0.519	1	0.333	0.700	0.479	0	0	0.851	0	0	1
0.636	1	0.666	0.800	0.298	0	0	0.772	0	0.161	0.500
0.480	1	0.333	0.650	0.469	0	0.500	0.485	0	0	1
0.623	1	0	0.690	0.354	0	0	0.534	0	0.241	0.500
0.701	1	0.666	0.750	0.323	0	0	0.603	0	0	1
…	…	…	…	…	…	…	…	…	…	…

Table 3 is the result of data calculation using the simple feature scale normalization method. The result of this normalization process is a scale that is in the range of 0 to 1, just like the results obtained using the min-max method. Simple feature scale normalization is a technique that aims to change the attribute values in the dataset so that all attribute values are in the interval [0, 1]. In the given table, this normalization has been applied to 11 attributes (Atr1 to Atr11) that previously may have had different value ranges. This simple feature scale normalization process is done by calculating the value of each attribute relative to the range of values of that attribute in the dataset. This method converts each original value of an attribute to a value in the range of 0 to 1 based on a comparison with the maximum

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

and minimum values of the attribute. This process uses formula (1). By applying simple feature scale normalization, the variation between attributes becomes more controllable and each attribute can contribute proportionally to the analysis or machine learning model to be used. Attributes with different scales can cause the model to give disproportionate weight to certain attributes, thus reducing the accuracy and effectiveness of the model. Table 3 shows how simple feature scale normalization has been applied to 11 attributes with various initial value ranges, so that all attributes are now in a uniform range of 0 to 1.

Table 3. Dataset without Normalization Process

Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Atr10	Atr11
40	1	1	140	289	0	1	172	0	0	2
49	0	2	160	180	0	1	156	0	1	1
37	1	1	130	283	0	2	98	0	0	2
48	0	0	138	214	0	1	108	1	1.5	1
54	1	2	150	195	0	1	122	0	0	2
…	…	…	…	…	…	…	…	…	…	…

Table 3 is a display of data that does not use the normalization method and can be seen the difference in the data attached. In table 2 and table 3, the resulting data has a scale of 0 to 1 in the data range. In Table 4, displayed raw data that does not go through the normalization process provides a clear picture of the variation in values between attributes that may be very different. Without normalization, the values of each attribute (Atr1 to Atr11) remain in their original scale which can range widely. For example, some attributes may have values in the tens or hundreds, while other attributes may have values in the small or even binary scale. Inconsistent data can cause some attributes to dominate the data analysis due to a larger range of values and ultimately reduce the accuracy of the built model. In this table, if attribute Atr1 has a value range between 20 to 100, while Atr2 has a value range between 0 and 1, Atr2's contribution to the data analysis model will be much smaller compared to Atr1 due to this different scale. This can cause bias in machine learning algorithms such as K-Nearest Neighbors (K-NN), where the calculation of distance between data points is strongly influenced by attributes with a larger range of values. Whereas in Table 1 and Table 2, the data has been ormalized using the min-max and simple feature scale methods which change the scale of attribute values to a range of 0 to 1. This normalization ensures that all attributes contribute proportionally to the analysis and no single attribute dominates due to differences in scale. The uniform scale allows machine learning algorithms to work more effectively and give equal weight to each attribute, thus improving the accuracy of the model.

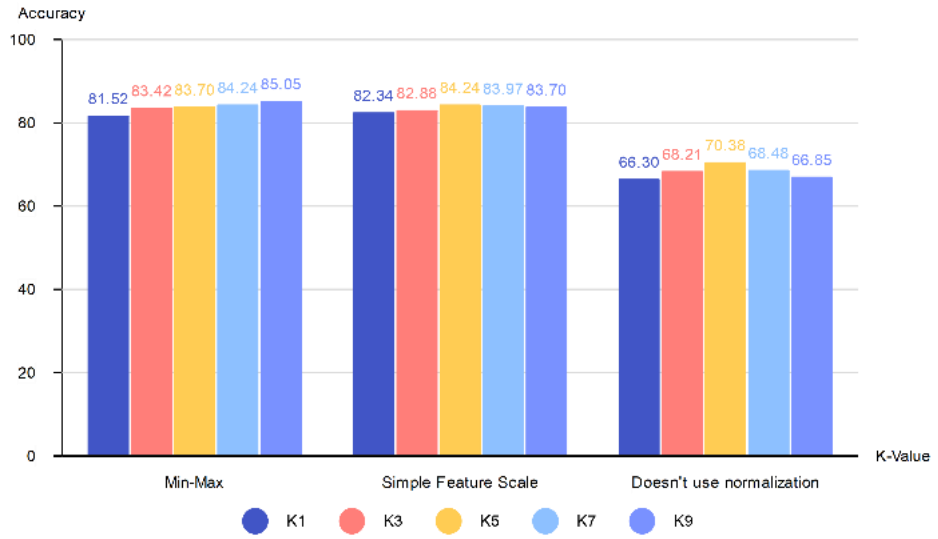

Figure 3. Accuracy graph of K-NN model comparison

In Figure 3, a comparison of the accuracy test results of the K-NN algorithm using three different normalization approaches is shown: min-max, simple feature scale, and no normalization. Based on the test results, it can be visually concluded that the min-max normalization method provides the highest accuracy results. The highest accuracy value obtained from the min-max method reached 85.05% at a value of K = 9. This shows that the use of min-max normalization can significantly improve the performance of the K-NN algorithm compared to other methods. Meanwhile, the simple feature scale normalization method also performed quite well with the highest accuracy of 84.24% at K=5. Although slightly lower than the min-max method, this result still shows that simple feature scale can improve the accuracy of K-NN compared to non-normalized data. The non-normalized data gives a much lower highest accuracy result of 70.38% at K=5. This confirms the importance of normalization in processing data before it is applied to the K-NN algorithm. Without normalization, different scales of data attributes can cause a significant

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

decrease in accuracy as attributes with larger scales can dominate the distance calculation in K-NN. Based on the test results visualized in Figure 3, it is clear that min-max normalization is the most effective method for improving K-NN accuracy in heart failure datasets. This approach successfully optimizes the performance of the algorithm by reducing the effect of different scales on the data attributes, resulting in a more accurate model.

4. CONCLUSION

Data preparation is a crucial initial stage in this research aimed at enhancing the accuracy of the analysis outcomes. Past research has shown the importance of preprocessing steps in preparing datasets for further analysis, especially in the context of machine learning algorithms. This study compares the effectiveness of three different approaches in data preprocessing, namely: min-max normalization, simple feature scale normalization and non-normalized data. The primary aim of this research is to quantify and compare the precision of data analysis outcomes that have undergone different normalizing techniques. Min-max normalization is a method that adjusts data values so that they are within a certain range, generally between 0 and 1, with the aim of reducing the disproportionate scale effect between data features. Another normalization used is simple feature scale, which is a simpler method that sets the data feature values based on a certain scale determined by the data distribution itself. This study analyzed data that did not go through the normalization process to evaluate its impact on the accuracy of the analysis results. The results of this study show that the use of min-max normalization produces the highest level of accuracy, which is 85.05%. This indicates that min-max normalization is very effective in preparing data for further analysis with machine learning algorithms. Simple feature scale normalization also improves the accuracy of the results although it is not as good as min-max normalization with an accuracy rate of 84.24%. In comparison, the unnormalized data gave the lowest accuracy result of 70.38%, indicating that without proper preprocessing steps the performance of the data analysis model can be significantly reduced. Overall, this study confirms the importance of preprocessing steps in improving the quality of data analysis and suggests the use of min-max normalization as the most effective method to improve the accuracy of results in the context of the dataset used. The findings make a valuable contribution to the field of machine learning and data analysis, showing that the selection of an appropriate preprocessing method is a crucial factor for achieving optimal analysis results.

REFERENCES

[1] B. Rahman, H. L. Hendric Spits Warnars, B. Subirosa Sabarguna, and W. Budiharto, "Heart Disease Classification Model Using K-Nearest Neighbor Algorithm," *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, 2021, doi: 10.1109/ICIC54025.2021.9632918.

[2] H. A. U. Rehman, C.-Y. Lin, and Z. Mushtaq, "Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease," *J. Chinese Inst. Eng.*, vol. 44, no. 1, pp. 77–87, Jan. 2021, doi: 10.1080/02533839.2020.1831967.

[3] J. C. Youn *et al.*, "Cardiovascular disease burden in adult patients with cancer: An 11-year nationwide population-based cohort study," *Int. J. Cardiol.*, vol. 317, pp. 167–173, 2020, doi: 10.1016/j.ijcard.2020.04.080.

[4] 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, doi: 10.1109/icses55317.2022.9914044.

[5] A. A. Shanbhag, C. Shetty, A. Ananth, A. S. Shetty, K. Kavanashree Nayak, and B. R. Rakshitha, "Heart Attack Probability Analysis Using Machine Learning," *2021 IEEE Int. Conf. Distrib. Comput. VLSI, Electr. Circuits Robot. Discov. 2021 - Proc.*, pp. 301–306, 2021, doi: 10.1109/DISCOVER52564.2021.9663631.

[6] D. Chicco and G. Jurman, "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, 2020, doi: 10.1186/s12911-020-1023-5.

[7] F. Meng *et al.*, "Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China," *BMJ Open*, vol. 9, no. 5, p. e023724, May 2019, doi: 10.1136/bmjopen-2018-023724.

[8] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[9] M. Mamun, A. Farjana, M. Al Mamun, M. S. Ahammed, and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?," in *2022 IEEE World AI IoT Congress (AIIoT)*, Jun. 2022, pp. 194–200, doi: 10.1109/AIIoT54504.2022.9817303.

[10] A. Syukur, D. Istiawan, W. Sulistijanti, and A. Ilham, "Hybrid genetic feature selection and support vector machine for prediction LQ45 index in Indonesia stock exchange," in *AIP Conference Proceedings*, 2023, vol. 2720, p. 020017, doi: 10.1063/5.0153673.

[11] P. Rahman, A. Rifat, I. A. Chy, M. M. Khan, M. Masud, and S. Aljahdali, "Machine Learning and Artificial Neural Network for Predicting Heart Failure Risk," *Comput. Syst. Sci. Eng.*, vol. 44, no. 1, pp. 757–775, 2022, doi: 10.32604/csse.2023.021469.

[12] I. Mahmud, M. M. Kabir, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, "Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel," *Diagnostics*, vol. 13, no. 15, p. 2540, Jul. 2023, doi: 10.3390/diagnostics13152540.

[13] R. W. Putri, A. Ristyawan, and M. N. Muzaki, "Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease," *JTECS J. Sist. Telekomun. Elektron. Sist. Kontrol Power Sist. dan Komput.*, vol. 2, no. 2, p. 143, Jul. 2022, doi: 10.32503/jtecs.v2i2.2708.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1332−1340
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.5925

[14] T. A. Assegie, S. J. Sushma, B. G. Bhavya, and S. Padmashree, "Correlation Analysis for Determining Effective Data in Machine Learning: Detection of Heart Failure," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–5, 2021, doi: 10.1007/s42979-021-00617-5.

[15] A. Masitha, M. K. Biddinika, and H. Herman, "K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 3, pp. 593–604, 2023, doi: 10.30812/matrik.v22i3.2984.

[16] C. Sowmiya and P. Sumitra, "Analytical study of heart disease diagnosis using classification techniques," *Proc. 2017 IEEE Int. Conf. Intell. Tech. Control. Optim. Signal Process. INCOS 2017*, vol. 2018-Febru, pp. 1–5, 2018, doi: 10.1109/ITCOSP.2017.8303115.

[17] H. Hartatik, M. B. Tamam, and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," *2020 2nd Int. Conf. Cybern. Intell. Syst. ICORIS 2020*, 2020, doi: 10.1109/ICORIS50180.2020.9320797.

[18] N. Huda, A. Y. Dewi, and A. Mahiruna, "Plasmodium falciparum Identification Using Otsu Thresholding Segmentation Method Based on Microscopic Blood Image," *Sci. J. Informatics*, vol. 10, no. 4, 2023, doi: https://doi.org/10.15294/sji.v10i4.47924.

[19] Y. Liang and C. Guo, "Heart failure disease prediction and stratification with temporal electronic health records data using patient representation," *Biocybern. Biomed. Eng.*, vol. 43, no. 1, pp. 124–141, Jan. 2023, doi: 10.1016/j.bbe.2022.12.008.

[20] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data," *Front. Energy Res.*, vol. 9, p. 652801, 2021, doi: 10.3389/fenrg.2021.652801.

[21] P. Ghosh *et al.*, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021, doi: 10.1109/ACCESS.2021.3053759.

[22] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," *Proc. - Int. Conf. Data Eng.*, vol. 2019-April, pp. 2086–2090, 2019, doi: 10.1109/ICDE.2019.00245.

[23] Imran, F. Qayyum, D.-H. Kim, S.-J. Bong, S.-Y. Chi, and Y.-H. Choi, "A Survey of Datasets, Preprocessing, Modeling Mechanisms, and Simulation Tools Based on AI for Material Analysis and Discovery," *Materials (Basel).*, vol. 15, no. 4, p. 1428, Feb. 2022, doi: 10.3390/ma15041428.

[24] P. Mamatha Alex and S. P. Shaji, "Prediction and diagnosis of heart disease patients using data mining technique," *Proc. 2019 IEEE Int. Conf. Commun. Signal Process. ICCSP 2019*, pp. 848–852, 2019, doi: 10.1109/ICCSP.2019.8697977.

[25] A. Masitha, M. K. Biddinika, and Herman, "Preparing Dual Data Normalization for KNN Classfication in Prediction of Heart Failure," *Klik - Kumpul. J. Ilmu Komput.*, vol. 4, no. 3, pp. 1227–1234, 2023.

[26] B. Lewandowicz and K. Kisiała, "Comparison of Support Vector Machine, Naive Bayes, and K-Nearest Neighbors Algorithms for Classifying Heart Disease," *Commun. Comput. Inf. Sci.*, vol. 1979, pp. 274–285, 2024, doi: 10.1007/978-3-031-48981-5_22.

[27] A. Kumar, E. R. Khan, and Deepika, "A Review On Heart Disease Detection Using Machine Learning Techniques," in *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Apr. 2024, pp. 317–323, doi: 10.1109/CCICT62777.2024.00059.

[28] P. Tabaghi, I. Dokmanić, and M. Vetterli, "Kinetic Euclidean Distance Matrices," *IEEE Trans. Signal Process.*, vol. 68, pp. 452–465, 2020, doi: 10.1109/TSP.2019.2959260.

[29] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bull. Electr. Eng. Informatics*, vol. 9, no. 1, pp. 326–338, 2020, doi: 10.11591/eei.v9i1.1464.