**KLIK: Kajian Ilmiah Informatika dan Komputer**
ISSN 2723-3898 (Media Online)
Vol 4, No 3, December 2023, Hal 1227-1234
DOI 10.30865/klik.v4i3.1382
https://djournals.com/klik

# Preparing Dual Data Normalization for KNN Classfication in Prediction of Heart Failure

**Alya Masitha, Muhammad Kunta Biddinika*, Herman**

Faculty of Industrial Technology, Magister Informatika, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
Email: [1]alyamasitha@gmail.com, [2,*]muhammad.kunta@mti.uad.ac.id, [3]hermankaha@mti.uad.ac.id
Email Penulis Korespondensi: [2,*]muhammad.kunta@mti.uad.ac.id

**Abstract**−Heart failure disease is a serious condition that is significant in affecting both a person's quality of life and health. Therefore, it is important to develop classification methods that can help detect this disease. In this research, a data preprocessing stage is performed before being used to classify heart failure diseases using machine learning models, such as K-NN. Data preprocessing is an effort to simplify data analysis and ensure accurate results, and it is a very essential step in analyzing data to improve the quality of the data used. The dataset used in this research is raw data that has not gone through the preprocessing stage. The dataset consists of 918 data with target attributes of 0 and 1, where a value of 0 indicates a normal condition and a value of 1 indicates a potential heart failure condition. Data preprocessing includes data cleaning, data transformation, and data normalization. The main objective of this research is to carry out the preprocessing stage on data derived from heart failure disease datasets. Based on the comparison between two normalization methods, namely Min-Max and Simple Feature Scale, it is found that the Simple Feature Scale normalization method has the best performance, with an accuracy rate of 85%, while the Min-Max normalization method only reaches 84%.

**Keywords**: Heart Failure; Min-Max; Simple Feature Scale; K-NN; Classification; Normalization; Preprocessing

## 1. INTRODUCTION

The heart is a vital organ in the body responsible for pumping blood to supply oxygen and nutrients throughout the body. Even during sleep, the heart maintains a continuous rhythm. Humans lack the ability to regulate their heart rate while the heart is actively pumping blood [1]. Heart failure is a prevalent and lethal global health issue. Early identification and prediction of heart failure are crucial to minimize risks and increase survival rates of patients [2]. Patient medical data is commonly utilized and analyzed by machine learning algorithms to predict the probability of heart failure [3]. Heart failure is a condition where the heart is unable to pump blood efficiently enough to meet the body's demands. Its diagnosis and management require a thorough understanding of cardiac physiology and a multidisciplinary approach. Certain factors, including high blood pressure, diabetes, obesity, smoking, and hereditary conditions, may contribute to the development of this disease [4].

The diagnosis of heart failure necessitates a comprehensive assessment of symptoms, medical history, physical examination findings, and medical test results. Nonetheless, technological advances and machines' data processing capabilities have enabled the application of machine learning approaches to bolster diagnosis and forecasting of this ailment. [5]. One machine learning method that can be utilized is the K-Nearest Neighbor (K-NN) technique [6]. This method works by classifying new data based on training data that has characteristics or features similar to the new data. In this case, the K-NN method can be used to predict a person's likelihood of experiencing heart failure based on collected medical data, such as age, gender, blood pressure, blood sugar levels, etc [7]. Conducting analysis and predictions using the K-NN method can be used for early prevention and treatment so as to minimize the risk of heart failure in a person.

Preprocessing is essential as the initial stage in this research. Its aim is to ready raw data for the subsequent phase. Data preprocessing methods exist in several forms, one of which is the conversion of information to a compatible format for processing by the system. Data preprocessing methods exist in several forms, one of which is the conversion of information to a compatible format for processing by the system. Data preprocessing methods exist in several forms, one of which is the conversion of information to a compatible format for processing by the system. Data is preprocessed to obtain more precise outcomes [8]. Data preprocessing involves removing irrelevant information and converting data into a format that is optimized for system processing. In line with conventional academic structure, the text will adhere to established guidelines for citation, footnote style, and institution formatting. Its purpose is to enhance the accuracy of results, decrease computational time, and reduce the space occupied by the data while preserving its content. Technical terminology will be explained upon first use, and sentences will be written in simple, objective language using standard syntax. Data preprocessing can comprise several activities, such as data cleaning, data integration, data reduction, and data transformation [9]. Some datasets have varying value ranges for individual attributes, causing certain attributes to underperform due to their significantly smaller value range. To ensure accuracy, normalization is required to standardize the value range for each attribute on a specific scale. To ensure accuracy, normalization is required to standardize the value range for each attribute on a specific scale. This process improves accuracy in data analysis [10].

Classification is a component of machine learning that involves building a model to predict target-class objects. The model typically consists of a set of training data with known labels. Classification functions are applied to predict the class of an unknown object. The process of classification involves three stages: model construction, model application, and evaluation. Model building involves using training data with pre-existing attributes and classes to create a model. This model is then applied to classify new data or objects. The next step involves evaluating the level of accuracy in building and applying the model to new data. Classification consists of two key processes: the training process and the testing process [11].

Data normalization standardizes medical feature values to remove bias and ensure they impact the learning process equally [12]. In this case, we will compare normalized data on heart failure using the K-Nearest Neighbor (K-NN) algorithm [13]. Datasets with varying ranges for each attribute possess values much smaller than others, making data normalization techniques necessary for properly normalized data. Methods for normalization include Min-Max normalization, simple feature scaling, decimal scaling, and Z-Score [14][15].

This study aims to review data preprocessing techniques to build a K-NN model using a dataset of heart failure. The dataset used in this study is numerical and categorical. This study will explain the use of the preprocessing stage using two normalization techniques, namely Min-Max and Simple Feature Scale before using the K-NN method in predicting heart failure based on medical data. So that it can be used as a comparison and reference which normalization method is suitable for use in this dataset and get precise accuracy results. It is hoped that this research can help further research to develop classification methods using preprocessing stages that can produce the best results. Previous research that has been carried out by many researchers can be seen in Table 1

**Table 1.** Comparison of previous research

| Author | Tittle and year | Dataset | Method |
|---|---|---|---|
| [9] | Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN (2019) | Dataset Wine | Min-max<br>Z-Score<br>K-NN |
| [11] | Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine (2021) | Dataset Wine | Min-max<br>Z-Score<br>Desimal Scale<br>Naïve Bayes<br>Decission Tree<br>Support Vector Machine |
| [14] | Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer (2021) | Dataset Breast Cancer | Min-max<br>Z-Score<br>K-NN |
| [16] | Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes (2022) | Dataset Diabetes | Min-max<br>Z-score<br>K-NN |
| This study | Preparing Dual Data Normalization for KNN Classfication in Prediction of Heart Failure (2023) | Dataset Heart Failure | Min-max<br>Simple feature scale<br>K-NN |

Table 1 Table 1 is a comparison of data normalization methods for various classification problems exemplified in these studies covering some related research. The first study, conducted by [9] in 2019, focused on wine classification using the K-NN algorithm. This study used two normalization methods, namely Min-max and Z-Score to process data from the Wine dataset. The next study, conducted by [11] in 2021, also focused on wine classification, but they compared more methods, including Min-max, Z-Score, Decimal Scale, Naïve Bayes, Decision Tree, and Support Vector Machine. They expanded the scope of the research by using more classification methods and normalization methods. In addition, [14] in 2021 compared the Min-max and Z-Score normalization methods in the context of the K-NN algorithm to test the accuracy in classifying breast cancer disease types from the Breast Cancer dataset. Meanwhile, [16] also used the K-NN algorithm in 2022, but they focused on the Diabetes dataset. They used Min-max and Z-Score as data normalization methods for diabetes prediction. Finally, in a recent study in 2023, a study that has not been clearly identified has been conducted to predict heart failure. They used the Heart Failure dataset and normalization methods such as Min-max and Simple feature scale, with the K-NN algorithm. This study introduced variations in the normalization methods used. Overall, this comparison table illustrates how different data normalization methods, such as Min-max, Z-Score, Decimal Scale, and Simple feature scale, are used in different classification contexts, using different datasets, and different algorithms. These studies aim to understand the effectiveness of normalization methods in improving prediction accuracy in different cases of classification problems.

## 2. RESEARCH METHODOLOGY

In research conducted by [17] preprocessing of data is carried out to prepare data so that it can be presented using data processing algorithms that are widely used and used for each processing step in data mining. The datasets used in this study are in the form of text and numbers. In another study conducted by [18] The data underwent five stages of preprocessing: data cleaning, data reduction, data scaling, data transformation, and data partitioning. The dataset utilized for this study comprises categories and numerical figures. The results of this analysis contribute to the advancement of data-driven research in the field of buildings.

In the research that will be carried out using datasets in the form of categories and numeric This research will explain the preprocessing process on categorical and numerical data using two normalization methods as the first step before the accuracy calculation process is carried out using one of the classification methods, namely K-NN.

K-Nearest Neighbor (K-NN) is a machine learning classification method that predicts outcomes using its nearest neighbors or the majority of its k nearest neighbors. Footnotes will follow a consistent format, and citations should adhere to style guides. Filler words will be excluded, and contractions will be avoided. Bias will be avoided through indicative phrasing. Finally, correct spelling, grammar, and punctuation are essential. K-Nearest Neighbor (K-NN) is a machine learning classification method that predicts outcomes using its nearest neighbors or the majority of its k nearest neighbors. This process classifies each new data point based on its membership in the class of its closest neighbor. Precision in vocabulary will be prioritized over layman's terms. The language will be objective, with no figurative or ornamental language, and oftentimes presented in a passive tone or in the third person. The value of k, which represents the number of nearest neighbors used for classification, determines the basis for classification, and therefore, this method is called K-Nearest Neighbor. Technical terms will be explained when first used. Sentences will be kept simple and direct, and the paragraph's purpose will be apparent [19].

Cleaning data means cleaning raw data for further analysis. Raw data is often imperfect, unstructured, or contains errors, discrepancies, or missing values. Therefore, data cleaning is carried out to ensure the data is in the right condition, consistent, and reliable before being used in analysis or modeling [18].

Data transformation is a method for standardizing attribute values to a common form. The technique ensures a logical flow of information, with causal connections between statements. Complex terminology and sprawling descriptions are avoided in favor of concise, clear sentences. Technical abbreviations are explained upon first usage. Academic writing sections are consistently structured, formatted, and cited. The language maintains a formal register, free of contractions, colloquial words, and informal expressions. Clear and objective language is used, avoiding emotional, biased, figurative, or ornamental vocabulary. The text remains balanced, precise, and grammatically correct, avoiding bias and utilizing subject-specific vocabulary where appropriate. Data transformation is necessary in this study as the dataset includes categorical attributes. By converting the data into numeric form through transformation, accuracy values can be calculated more easily and accurately. Transformations can enhance the effectiveness of models or algorithms utilized for data analysis. It is imperative to maintain objectivity by excluding subjective evaluations, ensuring clear and concise language with a logical flow of information, and avoiding biased or ornamental language. Additionally, adherence to a conventional structure, high-level language with consistent technical terms, and formal register is important. Precise word choice and grammatical accuracy, including consistent citation format, are also crucial. Finally, balanced language through hedging and avoiding bias are essential [20].

Data normalization is the most frequently used technique in research. Data normalization is used to scale the data of an attribute so that it is in a smaller range [21]. Data normalization facilitates research by providing several benefits, including easier application of machine learning algorithms, increased effectiveness of algorithms, and the ability to analyze data using specific methods. Normalization is necessary to prevent bias in analysis resulting from differences in attribute value ranges. The study will utilize two methods of normalization: Min-Max normalization and simple feature scaling. Min-max normalization reduces each attribute value to its minimum and divides it by the difference between the maximum and minimum values. This method transforms data into a range between 0 and 1. Use it when an attribute has varied values and the formula for this method is:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$x_{old}$ represents the attributes for each dataset, where $x_{min}$ is the minimum value for each attribute in the dataset and $x_{max}$ is the maximum value for each attribute in the dataset.

Simple feature scale normalization is a normalization method that is carried out by dividing each attribute value by the maximum value that exists in each attribute. The formula for this method is:

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

$x_{old}$ represents the attributes for each dataset and $x_{max}$ is the maximum value for each attribute in the dataset.

The study framework initially involves selecting the dataset in the first step. Next, the data will undergo preprocessing utilizing data cleaning techniques, data transformation, and data normalization in the second step. The complete data will then undergo preprocessing in the third step to calculate the accuracy of the K-NN classification method. For more information, please refer to Figure 1. Research Framework.

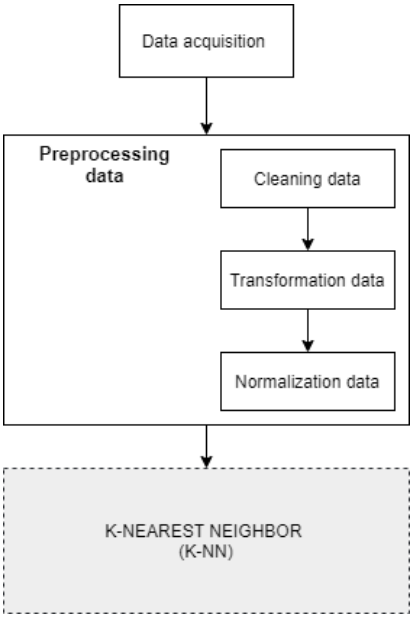


**Figure 1.** Research framework

Figure 1 depicts the steps or framework procedures that are to be conducted in this study to facilitate data analysis with the K-NN model. The initial step consists of selecting the data for analysis. The dataset used in this study was obtained from Kaggle.com, which has been previously utilized in several research studies. The second stage involves data preprocessing, encompassing various stages such as data cleaning, data transformation, and data normalization. Following the rigorous preprocessing of data, it will be utilized for the K-NN model. In this study, data cleaning will be executed by removing or correcting data that may cause inaccuracies harmful to analysis and/or machine learning applications. The data transformation process in this study converts categorical data into numerical data, thereby simplifying data processing for analysis. The normalization process aims to convert data values into a more specific or relative range, ensuring that each variable has a balanced and equal proportion. This study utilized two normalization methods: Min-Max and simple feature scaling. This study compares the values derived from the min-max normalization and simple feature scale. The aim is to determine which method produces more accurate results. The findings reveal that the simple feature scale yields more precise values. The study highlights the relevance of selecting the most appropriate normalization method based on the data characteristics and objectives of the research.

# 3. RESULT AND DISCUSSION

## 3.1 Preprocessing Data

At this early stage in the data processing chain, the main focus is on data preprocessing which involves a series of essential actions to prepare the data before it goes into further analysis. The steps involved data transformation, which included changing the format or scale of the data if needed. In addition, data normalization is also a priority, and to achieve this goal, two normalization methods are used, namely Min-Max Normalization and Simple Feature Scale. Through this normalization, the data will be transformed in such a way that the range of values corresponds to a consistent scale, which allows for more accurate comparison and analysis across datasets. This data preprocessing process is a critical step in ensuring good data quality, which in turn will support more reliable and informative analysis results at a later stage.

**Table 2.** Original 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 |

In Table 2, the dataset's original values have not undergone processing or preprocessing. Due to the data's diverse ranges, preprocessing is necessary for optimal test results. The dataset undergoes several preprocessing stages, beginning with data transformation from categorical to numeric. This data transformation utilizes encoder labels. The label encoder is a Python library derived from 'SciKit Learn', and its purpose is to convert categorical data and strings into numerical

values, making it easier for the model to comprehend the data. The outcomes of these modifications are displayed in Table 2. Results of the data transformation can be observed. Results of the data transformation can be observed.

**Table 3.** Results of data transformation

| 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 3 presented is a numerical representation of the initial categorical data. This conversion process is an essential step in the data preparation process, with the main purpose of facilitating further data processing. Categorical data naturally has different characteristics from numerical data, and therefore, conversion into numerical values allows for better compatibility with various data analysis algorithms. Once this conversion is complete, the data that originally contained categories will turn into a numerical representation, which will later be used in the normalization process. Data normalization is an important step to ensure that all variables in the dataset have a uniform scale, thus facilitating more accurate and consistent analysis. In other words, the ultimate goal of converting categorical data to numerical values is to achieve data alignment into a numerical format ready for use in further stages of analysis.

This statement is very relevant in the context of data processing and statistical analysis. Data normalization is a very important step in ensuring the integrity and power of data analysis. Without normalization, variables with very large or very small value ranges have the potential to dominate influence in the analysis process or in the creation of machine learning models. In an effort to avoid any distortions or biases that may arise, this study adopts the Min-Max Normalization method as the primary approach to data normalization. This method allows converting them into a uniform scale that ranges from 0 to 1, keeping all variables in a balanced value framework. With Min-Max normalization, scale differences between variables are eliminated, and this allows for fairer analysis, as well as the establishment of more accurate and reliable machine learning models. In other words, the use of the Min-Max Normalization method is a wise move to ensure that the data used in this study is treated consistently, and the results are reliable in the analysis and decision-making process.

Min-Max normalization is one of the normalization methods used in data analysis to change the value range of an attribute so that it corresponds to the same range between the minimum and maximum values present in the attribute. This technique is very useful in maintaining consistency in data analysis, especially when we work with various attributes that have various scales. Using Min-Max normalization, the attribute values will be adjusted into a range of 0 to 1, where the minimum value will be 0 and the maximum value will be 1, with the values in between being comparably distributed. This allows for a more precise comparison between attributes, and ensures that no attribute dominates in the analysis process. As an illustration, consider a set of numerical data that needs to be normalized using the Min-Max method. With this normalization, each value will be changed so that its range is standardized and makes it easier to understand and compare the data, and minimizes the risk of bias in data analysis.

$$x_{age} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{40-28}{77-28} = 0.244$$

$$x_{sex} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{0-0}{1-0} = 0$$

$$x_{\text{CP}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{1-0}{3-0} = 0.333$$

$$x_{\text{RBP}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{140-0}{200-0} = 0.700$$

$$x_{\text{CHO}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{289-0}{603-0} = 0.479$$

$$x_{\text{FBS}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{0-0}{1-0} = 0$$

$$x_{\text{RECG}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{1-0}{3-0} = 0.333$$

$$x_{\text{MaxHR}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{172-60}{202-60} = 0.788$$

$$x_{\text{ExAngina}} = \frac{x_{old}-x_{min}}{x_{max}-x_{min}} = \frac{0-0}{1-0} = 0$$

$x_{\text{oldpeak}} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} = \frac{0-(-2,6)}{6,2-(-2,6)} = 0.295$

$x_{\text{St\_Slope}} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} = \frac{2-0}{2-0} = 1$

**Tabel 4.** Results of Data Normalization Using Min-Max

| Age | Sex | CP | RBP | Cho | FBS | RECG | MaxHR | ExAngina | Oldpeak | St_Slope |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |
| … | … | … | … | … | … | … | … | … | … | … |
| 0.204 | 0 | 0.666 | 0.690 | 0.290 | 0 | 0 | 0.795 | 0 | 0.295 | 1.00 |

Table 4 is a visual representation of the Min-Max normalization results that have been applied to the dataset. Previously, this dataset has undergone a transformation, where categorical data has been converted into numerical data. This is a typical characteristic of Min-Max normalization, which aims to convert the data into a uniform scale, with the minimum value equivalent to 0 and the maximum value equivalent to 1, according to Table 3 which shows the conversion of categorical to numerical data. If the dataset initially contains negative values, the normalization method will convert the data range into an interval (min_value, max_value), where min_value is the lowest value in the dataset, and max_value is the highest value in the dataset. In Table 3, we see that Min-Max normalization has resulted in standardized data, with all variables being on the same scale. The result of this normalization is that all values are within the range of 0 to 1, so there are no values that are more than 1 or less than 0. This is important because Min-Max normalization ensures that no variable dominates the analysis or model, making the results more fair and consistent. Besides Min-Max normalization, there is also another technique used, which is simple feature scaling. This technique aims to adjust the data values into a range between 0 and 1, and it is very beneficial in ensuring that the data used in the analysis has a uniform scale. An example of calculating data using the simple feature scaling technique will provide a clearer understanding of this data normalization process.

$x_{age} = \frac{x_{old}}{x_{max}} = \frac{40}{77} = 0.519$

$x_{\text{Sex}} = \frac{x_{old}}{x_{max}} = \frac{0}{1} = 1$

$x_{CP} = \frac{x_{old}}{x_{max}} = \frac{1}{3} = 0.333$

$x_{\text{RBP}} = \frac{x_{old}}{x_{max}} = \frac{140}{200} = 0.700$

$x_{CHO} = \frac{x_{old}}{x_{max}} = \frac{289}{603} = 0.479$

$x_{FBS} = \frac{x_{old}}{x_{max}} = \frac{0}{1} = 0$

$x_{\text{RECG}} = \frac{x_{old}}{x_{max}} = \frac{0}{2} = 0$

$x_{MaxHR} = \frac{x_{old}}{x_{max}} = \frac{172}{202} = 0.851$

$x_{ExAngina} = \frac{x_{old}}{x_{max}} = \frac{0}{1} = 0$

$x_{oldpeak} = \frac{x_{old}}{x_{max}} = \frac{0}{6.2} = 0$

$x_{St\_Slope} = \frac{x_{old}}{x_{max}} = \frac{2}{2} = 1$

**Tabel 5.** Results of Data Normalization Using Simple Feature Scale

| Age | Sex | CP | RBP | Cho | FBS | RECG | MaxHR | ExAngina | Oldpeak | St_Slope |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |
| … | … | … | … | … | … | … | … | … | … | … |
| 0.493 | 1 | 0.666 | 0.690 | 0.290 | 0 | 0 | 0.856 | 0 | 0 | 1 |

Table 5 is a visual representation of the results of Simple Feature Scale normalization on a dataset that has previously gone through transformation from categorical attributes to numerical data. Before applying Simple Feature Scale normalization, the data has undergone a Min-Max normalization process, as shown in Table 4. This normalization aims to transform the attribute values into a range from 0 to 1, which is the hallmark of Min-Max normalization. One of the main benefits of this normalization is that it ensures that the resulting values remain within the range of 0 to 1. This is obtained because the normalization process transforms the data into intervals (new_min, new_max), where new_min is the minimum value in the dataset, and new_max is the maximum value in the dataset. The main purpose of this step is to prevent the dominance of variables that have a wider range of values, so that each variable can contribute equally to the analysis or modeling being performed. Simple Feature Scale Normalization is also very useful in the context of distance-based machine learning algorithms, such as K-Nearest Neighbors, where the scale of variables greatly affects the calculation of distances between data points. Moreover, this technique helps in modeling that requires numerical data to be of comparable scale, making the analysis results more consistent and easier to read. As reflected in Table 4, the results of Simple Feature Scale normalization are all values that fall within the range of 0 to 1, ensuring that every variable in the dataset has a uniform scale and does not result in adverse dominance during the analysis or modeling process.

## 4. CONCLUSION

The study's findings offer essential insights into the impact of normalization method selection on the accuracy of K-Nearest Neighbors (K-NN) model to detect heart failure disease. The comparison of two normalization methods indicates that the Simple Feature Scale method outperforms the other, yielding an 85% accuracy rate. This shows that this technique can generate a more precise K-NN model in categorizing the state of hypothetical heart failure patients. Meanwhile, employing the Min-Max normalization approach yielded a marginally lower accuracy rate of 84%. These outcomes demonstrate that the selection of normalization techniques has a noticeable impact on the ultimate model outcomes, and in this circumstance, Simple Feature Scale proved to be more effective in handling the information. An applicable suggestion for future research is to explore further by considering various other preprocessing techniques and different variations of normalization methods. This will allow for a more comprehensive analysis and can strengthen the validity of the data used. In addition, the examination of the K parameter in the K-NN algorithm could also be an important focal point in future research, as proper setting of the K parameter can have a significant impact on model performance. Thus, future research can delve deeper into the series of preprocessing stages and parameters of the K-NN algorithm to achieve more accurate and useful results in detecting heart failure diseases.

## ACKNOWLEDGEMENT

The author expresses his deepest gratitude to those who provided assistance in writing this scientific work. especially for the Master of Informatics Program at University of Ahmad Dahlan. The author also thanks the Directorate General of Higher Education. Research. and Technology. Ministry of Education. Culture. Research. and Technology of the Republic of Indonesia for funding this research in the Master's Thesis Research Scheme with contract No. 181/E5/PG.02.0.0PL/2023.

## REFERENCES

[1] B. Rahman, H. L. H. S. Warnars, B. S. Sabarguna, and W. Budiharto, "Heart Disease Classification Model Using K-Nearest Neighbor Algorithm," *2021 6th Int. Conf. Informatics Comput. ICIC 2021*, pp. 1–4, 2021, doi: 10.1109/ICIC54025.2021.9632918.

[2] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, 2019, doi: 10.1186/s12911-019-1004-8.

[3] H. Agrawal, J. Chandiwala, S. Agrawal, and Y. Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," *2021 Int. Conf. Intell. Technol. CONIT 2021*, 2021, doi: 10.1109/CONIT51480.2021.9498561.

[4] 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, 2020, doi: 10.1186/s12911-020-1023-5.

[5] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics Med. Unlocked*, vol. 16, p. 100203, 2019, doi: 10.1016/j.imu.2019.100203.

[6] A. Upadhyay, S. Nadar, and R. Jadhav, "Comparative study of SVM & KNN for signature verification," *J. Stat. Manag. Syst.*, vol. 23, no. 2, pp. 191–198, 2020, doi: 10.1080/09720510.2020.1724619.

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

[8] S. Hafeez and N. Kathirisetty, "Effects and Comparison of different Data pre-processing techniques and ML and deep learning models for sentiment analysis: SVM, KNN, PCA with SVM and CNN," *2022 1st Int. Conf. Artif. Intell. Trends Pattern Recognition, ICAITPR 2022*, 2022, doi: 10.1109/ICAITPR51569.2022.9844192.

[9] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan

Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
[10] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Theory*, vol. 25, no. 3, pp. 319–335, 2019, doi: 10.1007/s10588-018-9266-8.
[11] F. Adams, R. A. D. Anggoro, M. B. Satria, and A. W. Oktavia, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, no. September, pp. 260–268, 2021.
[12] 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.
[13] C. S. Wu, M. Badshah, and V. Bhagwat, "Heart disease prediction using data mining techniques," *ACM Int. Conf. Proceeding Ser.*, pp. 7–11, 2019, doi: 10.1145/3352411.3352413.
[14] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.47738/ijiis.v4i1.73.
[15] D. Borkin, A. Némethová, G. Michaľčonok, and K. Maiorov, "Impact of Data Normalization on Classification Model Accuracy," *Res. Pap. Fac. Mater. Sci. Technol. Slovak Univ. Technol.*, vol. 27, no. 45, pp. 79–84, 2019, doi: 10.2478/rput-2019-0029.
[16] M. Sholeh, D. Andayati, and R. Y. Rachmawati, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes," *TeIKa*, vol. 12, no. 02, pp. 77–87, 2022, doi: 10.36342/teika.v12i02.2911.
[17] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, *Data preprocessing in predictive data mining*, vol. 34. 2019.
[18] 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, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.
[19] G. S. R. Thummala and R. Baskar, "Prediction of Heart Disease using Decision Tree in Comparison with KNN to Improve Accuracy," in *Proceedings of the 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2022*, 2022, pp. 1–5, doi: 10.1109/ICSES55317.2022.9914044.
[20] 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, 2021, doi: 10.1007/s42979-021-00617-5.
[21] K. Burse, V. P. S. Kirar, A. Burse, and R. Burse, "Various Preprocessing Methods for Neural Network Based Heart Disease Prediction," *Adv. Intell. Syst. Comput.*, vol. 851, pp. 55–65, 2019, doi: 10.1007/978-981-13-2414-7_6.