JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 8 (ISSUE 1) JUNE 2024: 99 - 113
https://doi.org/10.21009/JSA.08109

COMPARATIVE ANALYSIS OF CLASSICAL METHODS WITH MACHINE LEARNING ALGORITHM ON SURVIVAL CLASSIFICATION OF HEART FAILURE PATIENTS

Sa'idah Zahrotul Jannah [1*], Grace Lucyana Koesnadi[2], Elly Pusporani[3]

[1,2,3] Statistics Study Program, Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga
Dr. Ir. H. Soekarno St., Surabaya, 60115, Indonesia

*Corresponding author's e-mail: * **s.zahrotul.jannah@fst.unair.ac.id***

ABSTRACT

Cardiovascular disease is a global threat and is the main cause of death worldwide. More than 17.9 million people died from heart and blood vessel problems. Most of these deaths, around 80%, occurred in countries with low or middle economies, including Indonesia. This research aims to find the most accurate and efficient model for classifying cardiovascular disease data so that cardiovascular disease can be detected early.

This research uses heart failure patient data with predictor and response variables. The response variable has two categories such as passed away and alive. Moreover, predictor variables are obtained from the patient's behavioral risk factors. Data preprocessing was done before the modeling and divided into 0% training and 20% testing data. Modeling in training data was done with multiple algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Each model was evaluated with metrics such as Accuracy, Precision, and Recall obtaining the best model.

This study found that the use of all research variables in the classification analysis leads to a decrease in classification performance, so this study used SelectKBest with a total of 8 significant variables. Furthermore, the Random Forest algorithm with optimal parameters using entropy criterion and a maximum depth of 8 is the method with the most optimal performance, achieving a precision of 90.51% for the 'alive' category, recall of 88.27% for 'alive', the precision of 88.55% for 'deceased', recall of 90.74% for 'deceased', training accuracy of 89.51%, AUC of 0.895, and testing accuracy of 87.80%, placing it in the category of good classification.

Although this research is limited to medical records and behavioral risk factors of heart failure patients to classify patient survival resilience, it addressed data imbalance, employed feature selection, and compared multiple algorithms to provide insights into their effectiveness for this specific classification task and improve model efficiency.

Article History:

Received: 20 April 2024
Revised: 3 June 2024
Accepted: 27 June 2024
Published: 30 June 2024
Available online.

Keywords:

Featured Selection, Heart Failure, Imbalanced Data, Classification, Machine Learning

How to cite this article:
S.Z. Jannah, G.L. Koesnadi, E. Pusporani, "COMPARATIVE ANALYSIS OF CLASSICAL METHODS WITH MACHINE LEARNING ALGORITHM ON SURVIVAL CLASSIFICATION OF HEART FAILURE PATIENTS", Jurnal Statistika dan Aplikasinya, vol. 8, iss. 1, pp. 99 – 113, June 2024

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Cardiovascular disease is the narrowing or blockage of blood vessels that can lead to heart attacks due to lack of blood, chest pain (angina), or strokes [1]. Typical symptoms of coronary heart disease include chest pain lasting more than 20 minutes, occurring during both activity and rest, accompanied by cold sweats, weakness, nausea, and dizziness. There are non-modifiable risk factors such as family history, age, gender, and obesity, as well as modifiable risk factors including hypertension, diabetes, dyslipidemia, lack of physical activity, unhealthy diet, and stress levels [2]. Cardiovascular disease remains a global threat and is a leading cause of death worldwide. More than 17.9 million people die annually due to heart and vascular issues [3]. Most of these deaths, around 80%, occur in low- or middle-income countries, including Indonesia [4].

In Indonesia, deaths due to cardiovascular diseases reach 651,481 people each year, with the majority caused by stroke (331,349 deaths), coronary heart disease (245,343 deaths), hypertensive heart disease (50,620 deaths), and other types of cardiovascular diseases [5]. This high prevalence in Indonesia is caused by unhealthy lifestyles, such as smoking habits and unbalanced diets. These behaviors contribute significantly to coronary heart disease and can even lead to sudden cardiac arrest.

Research on quality of life focusing on cardiovascular diseases is still very limited. There is no specific research on cardiovascular diseases, even though cardiovascular conditions are critical and related to the vital organ of the heart. Therefore, it is important to undertake swift and appropriate treatment actions [4]. Therefore, it is necessary to integrate artificial intelligence-based technology into the health sciences field to rapidly develop the healthcare industry in the Society 5.0 Era. This will have a positive impact on improving accurate and timely cardiovascular diagnosis, enabling more effective treatment, and enhancing overall public health.

Therefore, by applying several machine learning algorithms, this research aims to classify cardiovascular disease data and analyze the comparison of machine learning algorithms to obtain the most accurate and efficient algorithm for early detection of cardiovascular diseases. In this study, several methods used by researchers to classify the heart failure clinical records dataset include logistic regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM).

To optimize classification performance and reduce potential bias, several actions will be taken. First, the data balancing process will be conducted using resampling. Second, the data will be normalized, and feature selection will be performed to identify the most contributing variables in the heart failure patient dataset. Subsequently, the best model will be selected based on the highest evaluation metrics values.

2. METHODS

Material and Data

The data used in this study comes from a previous study that measured 299 heart failure patients collected in 2015 [6]. It can be accessed in the following link: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records. Empirical studies have proven that the best model is obtained when the dataset is divided into 70-80% training data and 20-30% testing data [7]. The division of the training and testing dataset in this study uses a split ratio of 80% and 20%. In the analysis process, the training data is used for modeling, while the testing data is used to compare the prediction results with the actual data to determine the goodness of the obtained classification model.

The research variables used consist of 13 attributes, including 1 response variable (Y) or labeling attribute, namely Death event divided into 2 categories, and 12 predictor variables (X) or attributes detailed in Table 1.

Table 1. Research Variable

No.	Variable	Operational Definition	Data Types
1.	Age	Patient's age (year)	Numeric
2.	Anemic	Whether there is a decrease in hemoglobin levels or not.	Categorical 0: Not anemic 1: Anemic
3.	High blood pressure	Whether suffering from hypertension or not	Categorical 0: Not hypertension 1: hypertension
4.	Creatinine phosphokinase	CPK enzyme level in blood (mcg/L)	Numeric
5.	Diabetes	Whether suffering from diabetes or not	Categorical 0: No diabetes 1: Diabetes
6.	Ejection fraction	Percentage of blood leaving the heart during contraction	Numeric
7.	Sex	Gender	Categorical 0: Female 1: Male
8.	Platelets	Platelets in the blood (kiloplatelets/mL)	Numeric
9.	Serum creatinine	Creatinine level in blood (mg/dL)	Numeric
10.	Serum sodium	Sodium level in blood (mEq/L)	Numeric
11.	Smoking	Whether the patient smokes or not	Categorical 0: Non-smoker 1: Smoker
12.	Time	Follow-up period (days)	Numeric
13.	Death event (target)	Whether the patient died during the follow-up period or not	Categorical 0: Alive 1: Deceased

Research Method

Logistic Regression

Binary logistic regression is a statistical analysis technique with one or more independent variables and one response variable. The independent variables can be either categorical or continuous data, while the response variable must be binary categorical. The response variable Y follows a Bernoulli distribution; thus, it only has two possible outcomes: failure (0) and success (1).

If the response variable Y consists of n instances, and the probability of each event is the same, with each event being independent of the others, then the response variable Y will follow a binomial distribution. Logistic regression model is developed by $E(Y = 1|X = x)$ with $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ with logistic regression model in Equation (1)

$$\pi(x) = \frac{e^{g(x)}}{1-e^{g(x)}} \quad (1)$$

In logistic regression modeling, a link function that is appropriate for the logistic regression model is required, which is the logit function. The logit transformation is a function of π(x). The logistic regression model equation is given in equation (2).

$$logit[\pi(x)] = log\ log\ \left(\frac{\pi(x)}{1-\pi(x)}\right) \quad (2)$$

with

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3)$$

K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) is a simple yet effective algorithm for classification. KNN has a high level of accuracy and efficiency in classification tasks [8]. The working principle of K-Nearest Neighbor (KNN) involves classifying based on the proximity (distance) of one data point to other data points. The closeness or distance can be calculated using one of the predetermined distance measures, such as Euclidean distance, Minkowski distance, and Mahalanobis distance. The concept of Minkowski distance treats all variables as independent (uncorrelated). The standard transformation applied eliminates the influence of data variability, meaning that all variables will contribute equally to the distance. The formula for Minkowski distance is as follows.

$$d(x_i, x_j) = \left(\sum_{k=1}^{K} \left|x_{ik} - x_{jk}\right|^r\right)^{1/r} \quad (4)$$

with:

x_{ik} : the i-th test data on the k-th variable
x_{jk} : the j-th test data on the k-th variable
$d(x_i, x_j)$: distance
k : dimension of predictor variables

If the value of r in the Minkowski distance is 1, then this distance is equivalent to the Manhattan distance. If the value of r is 2, then the Minkowski distance is equivalent to the Euclidean distance [9].

Decision Tree

A Decision Tree is a tree-like structure resembling a flowchart where internal nodes represent predictor variables used as decision points connected by branches, and each leaf node represents a classification outcome class [10]. This algorithm is developed by J Ross Quinlan in early 1980, which is developed Decision Tree ID3 (Iterative Dichot-omiser).

The variable selected as a splitter is the one that has the highest goodness of split value, as this variable can reduce heterogeneity the most effectively. If the predictor variable used is categorical data, then the splitting of nodes can use the categorical values of that variable, with one branch for each category. However, if the predictor variable is ratio-scaled or numerical data, various possible midpoint values among the sorted data are used as node splitters. The midpoint value that results in the highest goodness of fit is selected [9].

Support Vector Machine

Support Vector Machine (SVM) is a learning system that utilizes a hypothesis of linear functions in high-dimensional space. It is trained using algorithms based on optimization theory, applying learning biases derived from statistical theory. The main objective of this method is to construct the OSH (Optimal Separating Hyperplane), which creates an optimal separating function that can be used for classification purposes. The equation of the hyperplane for the case where data can be linearly separated by a straight line is illustrated in Equation (5).

$$WX + b = 0 \quad (5)$$

with $W = (w_1, w_2, \dots, w_p)$ is the weight vector, p is the number of variables X, and b is a constant or commonly referred to as bias. When data can be separated with a linear hyperplane, the function in Equation (5) can change to Equation (6) as follows

$$f(x) = w^T x + b \quad (6)$$

if $f(x) \geq 0$ for $y_i = +1$ and if $f(x) < 0$ for $y_i = -1$ [11]. For cases where data cannot be linearly separated (non-linearly separable data), the search for the optimal hyperplane will consider data points that do not lie within the class, developed with ξ.

In real-world cases, linearly separable data is quite rare. Therefore, kernel functions are used to map data into high-dimensional vector spaces. Some commonly used kernel functions are showed in Table 2 [9].

Table 2. Kernel Functions in SVM

Kernel	Kernel Function
Kernel Linier	$K(x_i, x) = x_i^T x$
Kernel Radial Basis Function	$K(x_i, x) = exp\ exp\ (-\gamma \|x - x_i\|^2)$
Sigmoid Kernel	$K(x_i, x) = tanh\ tanh\ (\gamma x_i^T x + r)$

Random Forest

Random Forest is one of the methods within decision trees. This method is used to construct decision trees consisting of root nodes, internal nodes, and leaf nodes by randomly selecting attributes and data according to specified rules. The root node is used to gather data, an inner node at the root node contains questions about the data, and a leaf node is used to solve problems and make decisions. The decision tree begins by calculating the entropy value as a measure of attribute impurity and the information gain value as in Equation (7) and (8) [12].

$$Entropy(Y) = -\sum_i \quad p(c|Y)p(c|Y) \quad (7)$$

with Y is set of cases and $p(c|Y)$ is the proportion of Y to c class.

$$Information\ Gain(Y, a) = Entropy(Y) - \sum_{v \epsilon Values(a)} \quad \frac{|Y_v|}{|Y_a|} Entropy(Y_v) \quad (8)$$

with $Values(a)$ is all the possible value in set a, Y_v is the subclass of Y with class v related with class a, and Y_a is all suitable value with a [13].

Evaluation Metrics

Confusion matrix is a method to obtain accuracy metrics in calculations used in data mining techniques [14]. The confusion matrix is used to calculate the number of observations in each class that are correctly and incorrectly classified by a classification model. The results are displayed in a table format [15]. Confusion matrix can be interpreted as a tool that assesses whether a classifier can effectively recognize tuples from different classes [16]. The goodness metrics of a classification model are also based on the results of the confusion matrix in Table 3.

Table 3. Confusion Matrix

Predicted Value	Actual Value	
	True	False
True	TP (True Positive)	FP (False Positive)
False	FN (False Negative)	TN (True Negative)

with TP (True Positive) representing the total observations where the actual and predicted values are both positive, FP (False Positive) indicating the total observations where the actual value is negative but predicted as positive, FN (False Negative) denoting the total observations where the actual value is positive but predicted as negative, and TN (True Negative) indicating the total observations where both the actual and predicted values are negative. Evaluation metrics used in this research include accuracy, precision, recall, and AUC (Area Under the Curve). Accuracy is an intuitive measure of correctness, calculating the ratio of observations correctly predicted across the entire dataset, with a mathematical formulation as presented in Equation (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Precision is the ratio of true positive predictions to the total predicted positive observations, formulated mathematically as in Equation (10).

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Recall is the measure of success or ability of a system to find information from the dataset, as defined in Equation (11) [17].

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

Area Under Curve (AUC), often used as a measure of the goodness of a classification, is calculated by assessing the performance difference of the method or algorithm used, formulated in Equation (12).

$$\theta^r = \frac{1}{mn}\sum_j^n \quad = 1\sum_i^m \quad = 1\psi(xi^r, xj^r) \quad (12)$$

with

$$\psi(X,Y) = \{1 \qquad Y < X\ \frac{1}{2} \qquad Y = X\ 1 \qquad Y > X \quad (13)$$

with X is the positive output and Y is the negative output.

Feature Selection

Feature selection involves selecting a subset of informative and relevant features or variables from a larger set, thereby improving the characterization of multiclass patterns [18]. Filter methods are one of the feature selection techniques that operate without using a classifier. This enhances computational efficiency in filter methods [19]. Univariate Feature Selection (SelectKBest) is a category of filter methods used in feature selection. SelectKBest is a feature selection algorithm aimed at improving prediction accuracy and performance on high-dimensional datasets. It is part of univariate feature selection, which selects the best features based on univariate statistical tests or ANOVA tests. These statistical tests help identify features with the strongest relationship to the response variable. SelectKBest retains only the top-scoring features while discarding the rest [20]. SelectKBest selects the top K features with the maximum relevance to the target variable [21].

Data Normalization

Normalization of data aims to maintain the range of data to remain balanced during the calculation process. [22]. Normalization using Min-Max Scaler will transform the scale of all original data into values that range between 0 and 1. The mathematical formulation used in Min-Max Scaler normalization is given by Equation (14).

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{14}$$

with x' represents the normalized value of the data, where x_i is the ith original data value, x_{min} is the minimum data value, and x_{max} is the maximum data value [23].

Research Flowchart

In detail, the stages of analysis in the research are outlined and presented in a model diagram illustrating the flow of the study in Figure 1.

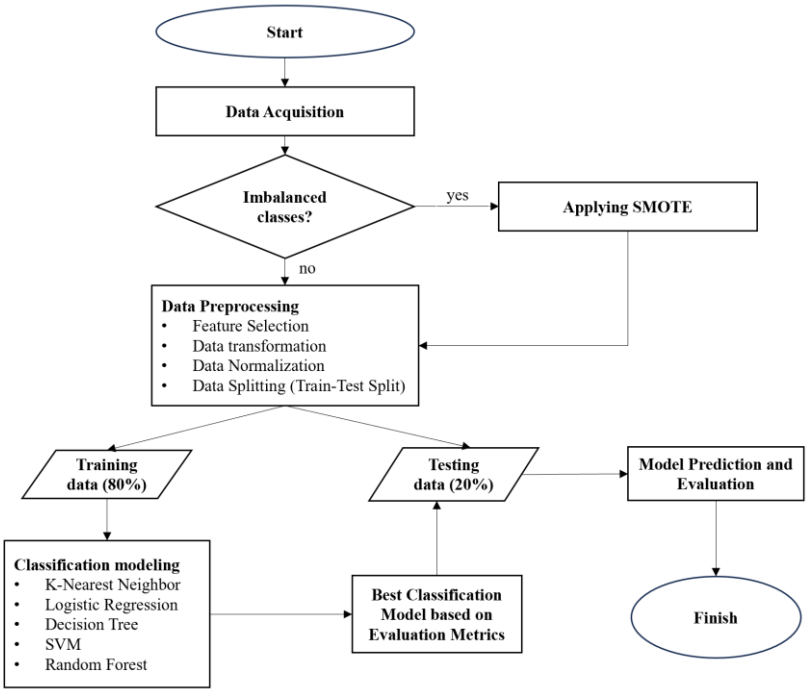


Figure 1. Research Flowchart

3. RESULTS

In this study, the analysis and discussion encompass stages such as data exploration and preprocessing, dataset splitting, classification process, determination of the best classification model based on evaluation metrics, and classification of testing data using the best classification method for heart failure patient cases.

Data Exploration and Preprocessing

The dataset related to heart failure patients will first be identified at the initial stage before further analysis. Initially, a check for missing values in the research data will be conducted, and upon verification, no missing values were found in the observations. Subsequently, exploratory data analysis for numerical predictor variables will be presented descriptively in Table 4.

Table 4. Descriptive Statistics

Variable	Mean	St. Dev.	Min.	Max.
Age	60.829	11.895	40.00	95.00
Creatinine phosphokinase	581.839	970.288	23.00	7861
Ejection fraction	38.084	11.835	14.00	80.00
Platelets	263358.029	97804.237	25100.00	850000.00
Serum creatinine	1.394	1.035	0.50	9.40
Serum sodium	136.625	4.412	113.00	148.00
Time	130.261	77.614	4.00	285.00

Based on Table 4, the platelet count of patients shows a normal average within the range of 150,000-450,000 per microliter. The average level of sodium in the blood also indicates normal results in the range of 135-145 mEq/L. However, the average levels of CPK enzyme and creatinine in the blood show high results that exceed the normal ranges, specifically 10-120 mcg/L and 0.8-1.2 mg/dL, respectively. Data exploration is also conducted on the response variable or class attribute regarding the comparison of data counts based on class categories, as presented in Figure 2.

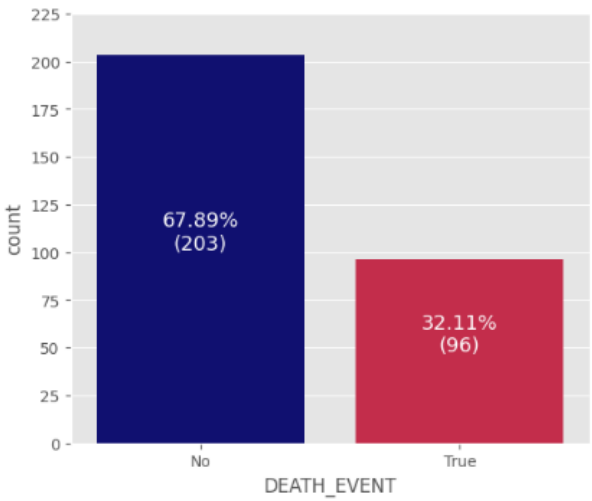


Figure 2. Comparison of Class Attribute before applying SMOTE

Figure 2 shows an imbalance in the dataset. This is evident from the number of data points in the "not deceased" category, indicated by the blue bar chart, being twice the number of data points in the "deceased" category, resulting in a ratio of 2:1. Imbalanced cases like this can lead to classification outcomes with skewed accuracy rates in favor of the dominant category, making the classification model less representative. Therefore, it is necessary to treat the dataset by performing resampling using oversampling techniques, which involves increasing the number of data points through data synthesis. In this study, the SMOTE method is used to balance the composition of data categories before conducting the analysis. A comparison of observations for categories in the response variable after applying SMOTE is presented in Figure 3.

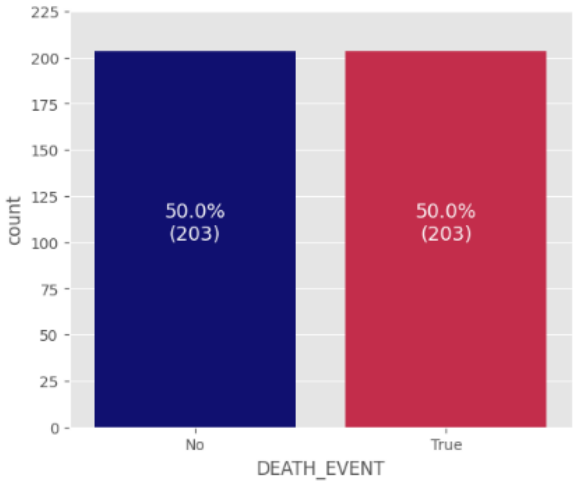


Figure 3. Comparison of Class Attribute after applying SMOTE

Next, feature selection is performed to identify variables that significantly influence and have maximum relevance to the response variable. The feature selection method used in this study is Univariate Feature Selection (SelectKBest), which falls under the category of filter methods aimed at improving prediction accuracy or enhancing performance on high-dimensional datasets. SelectKBest removes features that do not significantly impact the response variable. The process of feature selection using chi-square statistics is presented in Table 5.

Table 5. Result of Feature Selection with SelectKBest

Variabel	Chi-sq	*p-value*	Notes
Platelets	83620.87	< 0.001	Significant
Time	6693.38	< 0.001	Significant
Creatinine phosphokinase	684.61	< 0.001	Significant
Ejection fraction	139.30	< 0.001	Significant
Age	95.87	< 0.001	Significant
Serum creatinine	25.86	< 0.001	Significant
Smoking	6.94	< 0.001	Significant
Diabetes	4.69	0.030	Significant
Serum sodium	3.44	0.064	Not significant
Sex	1.32	0.251	Not significant
High blood pressure	1.20	0.273	Not significant
Anemic	0.78	0.377	Not significant

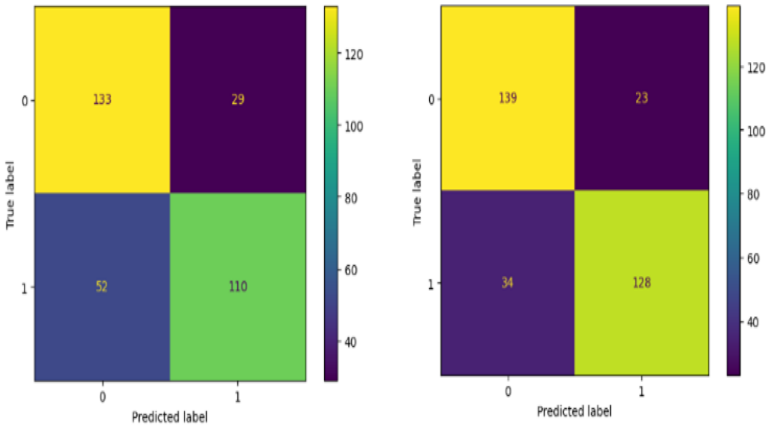
Variables with a p-value less than the significance level of 0.05 are considered significant or have a strong relationship with the response variable. Based on Table 5, 8 out of 12 top features were identified as having strong relevance to the response variable, while the other 4 features were excluded from the feature selection process. Subsequently, the dataset is transformed using the Yeo-Johnson transformation method and normalized or scaled using Min-Max Scaler to prevent variables with large unit values from dominating variables with smaller values or to address the issue of large data scales in numerical variables.

Data *Training* and *Testing Splitting*

The resampled data, which now has a balanced number of samples in both classes, will be further divided into two parts: training data and testing data. The data will be split randomly with an 80% to 20% ratio, resulting in 324 observations for the training set and 82 observations for the testing set. For subsequent research, the data will be proportioned into training and testing sets using the K-fold cross-validation method, which divides the data into K folds or sections randomly, with K typically set to 10.

K-Nearest Neighbor

The first method in the machine learning algorithm used is K-Nearest Neighbors (KNN) with the Minkowski distance metric. The parameter in this method is the number of neighbors to be used. In determining the optimal parameter, hyperparameter tuning is performed using the Grid Search method over a range of neighbors from 2 to 12. The tuning conducted on all variables resulted in an optimal number of neighbors of 3. Meanwhile, for the Feature Selection variables, the optimal number of neighbors obtained was 9. The confusion matrix results using the KNN method, both for all variables and for the Feature Selection variables, are presented in Figure 4.

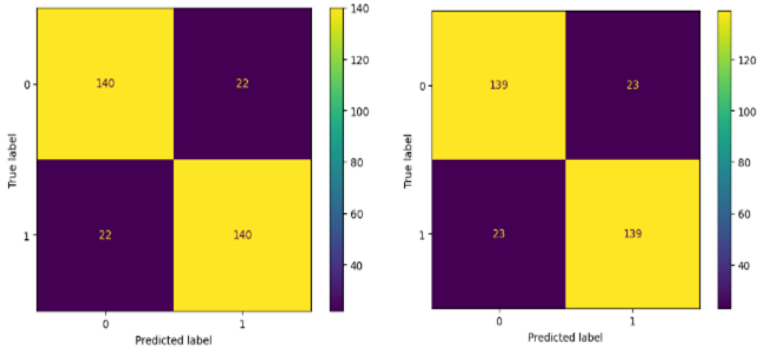


(a) (b)

Figure 4. Confusion Matrix of K-Nearest Neighbor Classification (a) Without Feature Selection and (b) With Feature Selection

Binary Logistic Regression

The next classification method used is the classic statistical method called binary logistic regression. Like other methods, this approach involves comparing the classification performance between all variables and the Feature Selection variables. Subsequently, the confusion matrix results using binary logistic regression will be presented for both all variables and the Feature Selection variables in Figure 5.



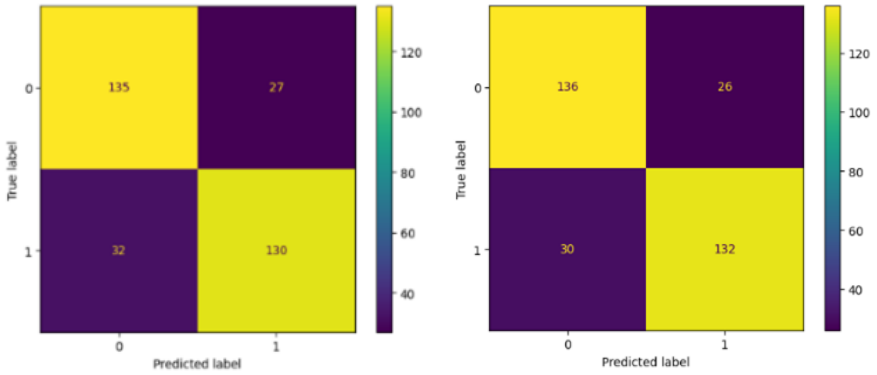
(a) (b)

Figure 5. Confusion Matrix of Binary Logistic Regression (a) Without Feature Selection and (b) With Feature Selection

Decision Tree

The decision tree method does not require specific assumptions to be met, allowing data resulting from preprocessing to be directly modeled. However, in line with the research objective of comparing classification methods, the analysis using the decision tree method will also be applied similarly. Optimal parameters obtained for all variables through hyperparameter tuning using Grid Search include combinations of criteria (Gini and entropy), maximal depth ranging from 2 to 22 with increments of 2, minimal sample split ranging from 2 to 4, and minimal sample leaf ranging from 1 to 4. The best parameter combination obtained for all variables is entropy criterion, maximal depth of 4, minimal sample leaf of 1, and minimal sample split of 2. Meanwhile, for feature selection, the best parameters are Gini criterion, maximal depth of 4, minimal sample leaf of 1, and minimal sample split of 2.

From the 324 preprocessed observations modeled, if the age is less than or equal to 0.422, the next classification process is based on variable X3; otherwise, it is based on variable X1. For the first case, if the Diabetes value is less than or equal to 0.352, it should be based on variable X5; otherwise, it is based on variable X1. This process continues until all observations are classified into the two existing class categories. The confusion matrix using the decision tree method is presented in Figure 6.

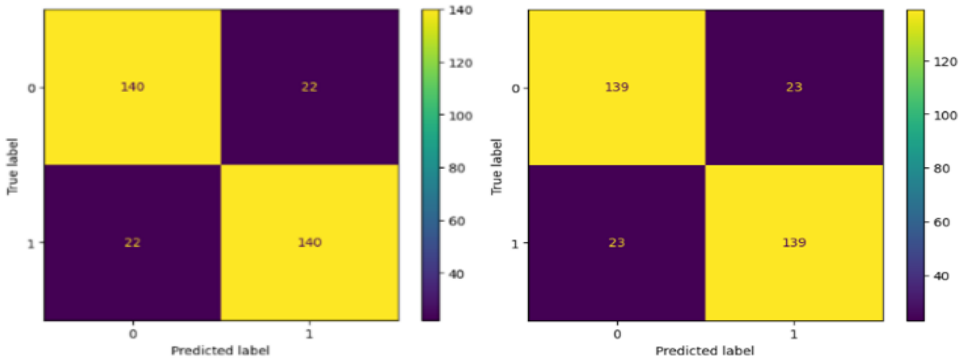


(a) (b)

Figure 6. Confusion Matrix of Decision Tree Classification (a) Without Feature Selection and (b) With Feature Selection

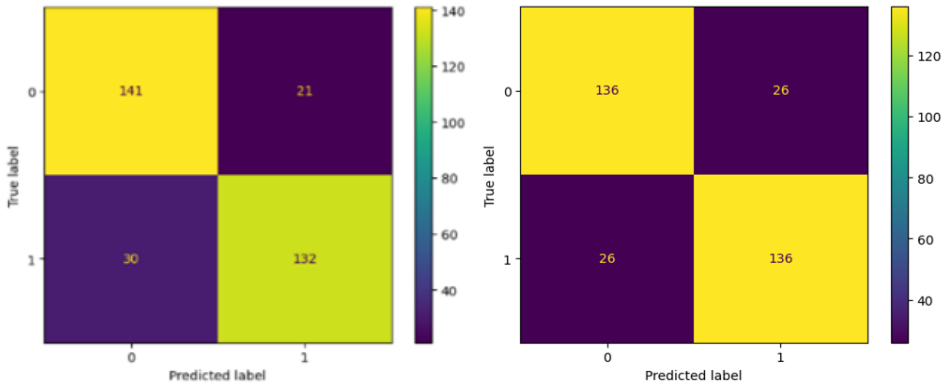
Support Vector Machine (SVM)

Next, a machine learning algorithm using the nonlinear classification method Support Vector Machine (SVM) is employed. In this study, two functions will be used: linear and radial basis function (RBF). The optimal parameters used are obtained through the hyperparameter tuning process with the Grid search method for the cost parameter with a range of 0.001 to 1000 in increments of 10. The best parameters obtained for all variables are cost with a value of 1000 for the linear function and cost with a value of 1 for the RBF function, with degree set to 3 (default) and gamma following the default of the software used for the analysis process. For the feature selection variable, a cost of 10 is obtained for the linear function and a cost of 1 for the RBF function. The confusion matrix using the SVM method with the linear kernel function is presented in Figure 7 and with the RBF function in Figure 8.



(a) (b)

Figure 7. Confusion Matrix of Linear SVM Classification (a) Without Feature Selection and (b) With Feature Selection

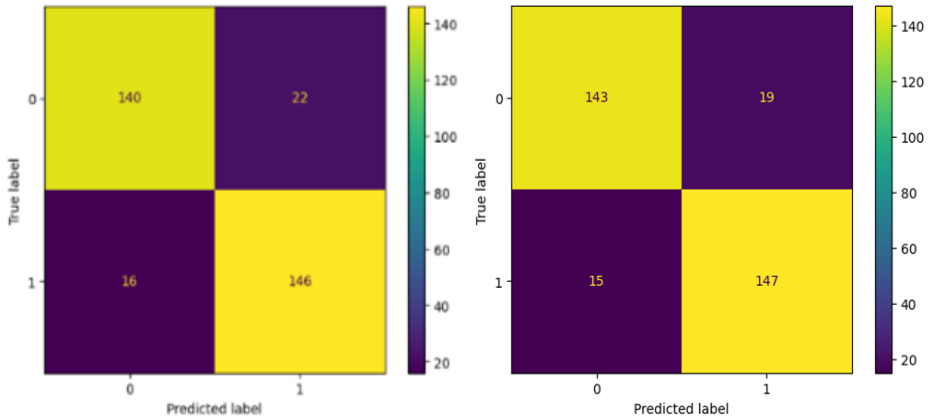


(a) (b)

Figure 8. Confusion Matrix of RBF SVM Classification (a) Without Feature Selection and (b) With Feature Selection

Random Forest

The next algorithm used is Random Forest. The optimal parameters obtained through the hyperparameter tuning process with the Grid search method for a combination of entropy and gini criteria, with a maximum depth of 2 to 22. The best parameters obtained for all variables are the entropy criterion with a maximum depth of 19. For the feature selection variable, the entropy criterion with a maximum depth of 8 is obtained. The confusion matrix using the random forest method is presented in Figure 9.



(a) (b)

Figure 9. Confusion Matrix of Random Forest Classification (a) Without Feature Selection and (b) With Feature Selection

Selection of The Best Classification Method

After performing classification analysis using several methods, the next step is to determine evaluation metrics to measure the performance of each model used. Each measure of the evaluation metrics, namely: precision, recall, accuracy, and AUC, which is obtained will be compared. The comparison of the performance of the analysis results is presented in Table 6.

Table 6. Summary of Evaluation Metrics

Methods	Precision '0'	Recall '0'	Precision '1'	Recall '1'	Accuracy	AUC
All variable						
KNN	71.89%	82.10%	79.14%	67.90%	75%	0.750
Logistic Regression	86.42%	86.42%	86.42%	86.42%	86.42%	0.864
Decision Tree	80.84%	83.33%	82.80%	80.25%	81.79%	0.818
SVM *Linear*	86.16%	84.57%	84.85%	86.42%	85.49%	0.855
SVM RBF	82.46%	87.04%	86.27%	81.48%	84.26%	0.864
Random Forest	**89.74%**	**86.42%**	**86.90%**	**90.12%**	**88.27%**	**0.883**
With feature selection						
KNN	80.35%	85.50%	84.77%	79.01%	82.41%	0.824
Regresi Logistik	84.15%	85.19%	85%	83.95%	84.57%	0.858
Decision Tree	81.93%	83.95%	83.54%	81.48%	82.72%	0.827
SVM *Linear*	86.42%	86.42%	86.42%	86.42%	86.42%	0.864
SVM RBF	83.95%	83.95%	83.95%	83.95%	83.95%	0.840
Random Forest	**90.51%**	**88.27%**	**88.55%**	**90.74%**	**89.51%**	**0.895**

Table 6 shows the classification results using all variables and feature selection variables that have accuracy, precision, and recall values that are not much different. Therefore, the four variables that are not used do not have a significant impact on determining the classification of whether a patient with heart failure will die or not during follow-up. Of all the tests on the five algorithms, the method that provides the best evaluation metrics, both using all variables and feature selection variables, is the Random Forest method. The Random Forest method has the most optimal performance on evaluation metrics with the highest precision, recall, accuracy, and AUC values compared to other methods. The best classification method in this study is Random Forest with the implementation of feature selection variables that has an AUC value of 0.895, which is included in the category of good classification.

The results of handling data imbalance and implementing feature selection can improve the performance of machine learning algorithms. This is proven by the analysis that has been done. In general, the comparison of algorithm performance in this study presented in Table 6 shows an improvement. The classification model through the process of handling imbalanced data and feature selection can be used as an alternative way to improve and optimize the classification performance of patients with heart failure who die and patients with heart failure who do not die.

Testing data classification and model evaluation

After obtaining the best classification method on the training data, the next step is to make predictions to classify the testing data based on the best classification model obtained. In the classification of testing data, the random forest method with feature selection variables is used, and then the next step is to evaluate the model. The evaluation of the classification model is based on the evaluation metrics in Table 7.

Table 7. Evaluation Metrics of Testing data

Precision '0'	Recall '0'	Precision '1'	Recall '1'	Accuracy	*AUC*
87,80%	87,80%	87,80%	87,80%	87,80%	0,878

Based on Table 7, the accuracy value of the testing data is 87.80%, which means that the method has good classification accuracy. The precision value obtained for category "0" is 87.80%, which indicates that the proportion of heart failure patients who are correctly classified as non-deceased heart failure patients is 87.80%. The recall value obtained is 87.80%, which means that the proportion of non-deceased heart failure patients who are correctly classified as non-deceased heart failure patients is 87.80%. The AUC value is 0.878, which indicates that the classification results fall into the category of good classification.

4. DISCUSSIONS

This study investigated the effectiveness of machine learning algorithms for classifying heart failure patients based on their likelihood of death during follow-up. The analysis explored the impact of data pre-processing techniques, including handling imbalanced data, feature selection, and performance comparison of five classification algorithm.

The initial dataset exhibited an imbalance between deceased and non-deceased patients. This was addressed using SMOTE, an oversampling technique that increased the number of data points in the minority class. SMOTE is a well-established technique for handling imbalanced datasets [24].

Univariate Feature Selection (SelectKBest) was employed to identify relevant features significantly impacting the response variable. The findings showed comparable or slightly improved performance with feature selection, suggesting its effectiveness in this context. It states that feature selection methods can reduce computational time, improve algorithm performance, and optimize prediction results [25].

The study compared the performance of five classification algorithms: K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Support Vector Machine (SVM) with linear and RBF kernels, and Random Forest. Both all available features and the selected features were used for training each model. Random Forest outperforms among all algorithms. It achieved the best performance based on evaluation metrics (precision, recall, accuracy, and AUC) using both all features and the selected features. Random Forest can be a key contributor to the overall performance [26].

5. CONCLUSION

The analysis of the data revealed an imbalanced data condition that tends to predict the majority class, necessitating the implementation of resampling using the SMOTE method. Utilizing all research variables in the classification analysis resulted in a decrease in classification performance, necessitating feature selection using SelectKBest with a total of 8 significant variables. Furthermore, the Random Forest algorithm with optimal parameters using the entropy criterion and maximal depth 8 is the method with the most optimal performance, achieving a precision for the living category of 90.51%, a recall for the living category of 88.27%, a precision for the deceased category of 88.55%, a recall for the deceased category of 90.74%, a training accuracy of 89.51%, an AUC of 0.895, and a testing accuracy of 87.80%, which falls into the category of good classification.

6. REFERENCES

[1] Ministry of Health Republic Indonesia, “Pathfinder : KARDIOVASKULAR.” Library of Ministry of Health Republic Indonesia, 2023.

[2] I. A. E. Widiastuti, R. Cholidah, G. W. Buanayuda, and I. B. Alit, “Deteksi Dini Faktor Risiko Penyakit Kardiovaskuler pada Pegawai Rektorat Universitas Mataram,” *Jurnal Pengabdian Magister Pendidikan IPA*, vol. 4, no. 1, 2021.

[3] World Health Organization, “Cardiovascular diseases.” WHO International, 2023.

[4] Jumayanti, A. L. Wicaksana, and E. Y. A. B. Sunaryo, “Kualitas Hidup Pasien Dengan Penyakit Kardiovaskular Di Yogyakarta,” *Jurnal Kesehatan*, vol. 13, no. 1, pp. 1–12, 2020.

[5] S. N. Tarmizi, “Cegah Penyakit Jantung dengan Menerapkan Perilaku CERDIK dan PATUH.” SehatNegeriku, 2023.

[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. 16, 2020, doi: https://doi.org/10.1186/s12911-020-1023-5.

[7] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation." The University of Texas at El Paso, 2018.
[8] M. H. Hanafi, N. Fadillah, and N. Insan, "Optimasi Algoritma K-Nearest Neighbor untuk Klasifikasi Tingkat Kematangan Buah Alpukat Berdasarkan Warna.," vol. 4, no. 1, pp. 10–18, 2019.
[9] Elly Pusporani, Siti Qomariyah, and Irhamah, "Klasifikasi Pasien Penderita Penyakit Liver," *Inferensi*, vol. 2, no. 1, pp. 25–32, 2019.
[10] J. Han, J. Pei, and H. Tong, *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc, 2022.
[11] S. Ramkishore, P. Madhumitha, and P. Palanichamy, ""Comparison of Logistic Regression and Support Vector Machine for The Classification of Microstructure and Interfacial Defects in Zircaloy-1," *Int. Conf. Soft Comput. Mach. Intell.*, 2014.
[12] K. Schouten, F. Frasincar, and R. Dekker, "An Information Gain-Driven Feature Study for Aspect-Based Sentiment Analysis," *Nat. Lang. Process. Inf. Syst.*, pp. 48–59, 2016.
[13] P. R. Sihombing and I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK J. Manaj. Tek. Inform. Dan Rekayasa Komput.*, vol. 20, no. 2, pp. 417–426, 2021, doi: 10.30812/matrik.v20i2.1174.
[14] L. Mutawalli, M. T. A. Zaen, and W. Bagye, "Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto)," *J. Inform. Dan Rekayasa Elektron.*, vol. 2, no. 2, pp. 43–51, 2019.
[15] Z. Chen et al., "The Lao text classification method based on KNN," *Procedia Comput. Sci.*, vol. 166, pp. 523–528, 2020.
[16] A. Yulianto, P. Sukarno, and N. A. Suwastika, *Improving adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset*. IOP Publishing, 2019.
[17] M. Riyyan and H. Firdaus, "PERBANDINGAN ALGORITME NAIVE BAYES DAN KNN TERHADAP DATA PENERIMAAN BEASISWA (Studi Kasus Lembaga Beasiswa Baznas Jabar)," *J. Inform. Dan Rekayasa Elektron.*, vol. 5, no. 1, pp. 1–10, 2022, doi: 10.36595/jire. v5i1.547.
[18] Z. M. Hira and D. F. Gillies, "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," *Advances in Bioinformatics*, vol. 5, pp. 1–13, 2015, doi: 10.1155/2015/198363.
[19] M. R. Al-Masud and M. R. H. Mondal, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms," *PLoS One*, vol. 15, no. 2, pp. 1–21, 2020.
[20] B. George, "A study of the effect of random projection and other dimensionality reduction techniques on different classification methods," *A Biannu. J. Interdiscip. Stud. Res.*, vol. 18, no. 1, 2017.
[21] A. C. Muller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. Sebastopol, CA: O'Reilly Media, Inc., 2016.
[22] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *CESS J. Comput Eng Syst Sci*, vol. 4, no. 1, pp. 78–82, 2019.
[23] U. M. Kusman, A. Hamid, D. C. R. Novitasari, W. D. Utami, and I. A. Wijaya, "Optimasi Model Penugasan Berdasarkan Peramalan Layanan Kapal Tunda Di Pelabuhan Tanjung Perak Menggunakan Metode Backpropagation," *J. Mnemon.*, vol. 6, no. 1, pp. 41–47, 2023, doi: 10.36040/mnemonic. v6i1.6008.
[24] R. Blagus, N. Zech, and D. Stefanik, "SMOTE-NC: Addressing the problem of noisy minorities in classification problems," *Information Sciences*, vol. 517, 2020.
[25] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.
[26] M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Expert Systems with Applications*, vol. 41, no. 16, pp. 7155–7161, 2014.