Classification of Cardiovascular Disease Based on Lifestyle Using Random Forest and Logistic Regression Methods

Ajyan Brava Bietrosula[1], Indah Werdiningsih[2], Eto Wuriyanto[3]
[1,2,3]Program study of Information System, Faculty of Science and Technology, Universititas Airlangga, Indonesia

Article Info

Article history:

Received Jan 17, 2024
Revised Mar 30, 2024
Accepted May 5, 2024

Keywords:
Cardiovascular Disease
Classification
Heart Disease
Lifestyle
Logistic Regression
Random Forest

ABSTRACT

Cardiovascular disease is a non-communicable disease caused by a disturbance in the function of the heart or blood vessels. According to WHO country profile data released in 2018 regarding non-communicable diseases, cardiovascular disease is the highest cause of death in Indonesia. This study aims to classify cardiovascular disease based on lifestyle using the Random Forest and Logistic Regression methods. In the classification process with the Random Forest and Logistic Regression machine learning methods, a combination of parameters from each machine learning method will be tested to see which parameter combination is the best for processing and classifying cardiovascular disease datasets. The dataset used in this research is obtained from Kaggle called Cardiovascular Disease. The dataset was processed through several pre-processing stages, namely missing value imputation, outlier detection, and extreme data checking. After going through the pre-processing process, the amount of data that entered the classification process was 62478 rows of data with 13 attributes or columns, namely age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity, and cardiovascular disease. Dividing the dataset into different percentage distributions of training data and testing data was also tested to see the difference in classification performance of the two methods. The division of training data was 90% and testing data is 10%. The results obtained from this study were the Logistic Regression method had better accuracy results of 73.07% compared to Random Forest with an accuracy result of 71.87%.

Corresponding Author:
Indah Werdiningsih
Program study of Information Systems,
Faculty of Science and Technology,
Airlangga University,
Dr. Ir. H. Soekarno Road, Mulyorejo, Surabaya 60115, Indonesia.
Email: indah-w@fst.unair.ac.id

1. INTRODUCTION

According to World Health Organization (WHO) country profile data released in 2018 regarding non-communicable diseases, cardiovascular disease is the highest cause of death in Indonesia [1]. Riset Kesehatan Dasar (Riskesdas) states that the number of heart disease cases is increasing from year to year with a prevalence of 1.5% or 15 out of 1,000 Indonesians suffering from heart disease in 2018. Congestive heart failure (CHF) is one of the a health problem in the cardiovascular system with the number of cases continuing to increase every year [2]. In research conducted at Sumber Waras Hospital in 2019, the results showed that age, gender and lifestyle factors influence a person's susceptibility to coronary heart disease. Coronary heart disease often occurs in people over 35 years of age with a risk of 0.143 times greater than those under 35 years of age. Coronary heart disease also occurs more often in men with a risk that is 8 times greater than in women. In addition, lifestyle factors such as smoking and lack of exercise by a person have a higher susceptibility to coronary heart disease [3].

In preventing an increased risk of cardiovascular disease, physical activity and a positive lifestyle are important factors in preventing cardiovascular disease and improving quality of life. Decreased physical activity may lead to an increased burden of cardiovascular disease. Regular physical activity has been proven to help prevent and treat non-communicable diseases such as heart disease, stroke, diabetes, and breast and colon cancer [4]. Apart from physical activity, diet modification is also one of the most important strategies for preventing cardiovascular disease. Implementing a healthy lifestyle such as choosing a diet can also reduce a person's susceptibility to cardiovascular disease [5].

Cardiovascular diseases, encompassing conditions like heart attacks and strokes, are often influenced by an individual's lifestyle. Unhealthy habits such as consuming high-fat and high-cholesterol diets, lack of physical activity, smoking, and stress can elevate the risk of these diseases. Research indicates that embracing a healthy lifestyle, including balanced dietary intake, regular exercise, abstaining from smoking, and stress management, can mitigate the likelihood of cardiovascular diseases[6]. Furthermore, monitoring cardiovascular risk factors and taking appropriate preventive measures, such as regular health check-ups and managing blood pressure and cholesterol levels, can also aid in preventing these diseases. Therefore, it's crucial for individuals to adopt a healthy lifestyle as the primary preventive measure in maintaining the health of their heart and blood vessels.

Over the years, the significant development of technology and information has provided benefits in various fields, one of which is in the health sector. The development of technology and information in the health sector has shown a great contribution to improving medical support services and facilities for the people of Indonesia. Machine learning has created various facilities in the medical field, such as medical imaging, disease identification, disease diagnosis, smart health records, disease prediction, and other facilities. Medical teams can diagnose or predict diseases earlier and more accurately, if methods from machine learning can be applied optimally.

Machine learning methods are widely applied in the fields of health and natural sciences by inputting biological, medical, and life science data into machine learning models for academic purposes, such as making decisions about patient care, developing medicines, or developing new medical methods. Classification-based machine learning methods are commonly used in various studies to predict events, one of which is to classify diseases. In classification-based machine learning methods, a type of supervised learning method is used where an input data set is divided into training data and target data (output) to create a prediction model based on the results of the level of conformity between the target or test data and the training data [7].

There are many types of classification techniques in machine learning methods, such as Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Decision Tree, Support Vector Machine, and Random Forest. In this research, the classification technique used as the basis of the learning model is Random Forest. Random Forest is one of the most frequently used and powerful ensemble machine learning classification techniques in pattern recognition and high-dimensional classification. Random Forest has several capabilities to provide explanations of results based on the ranking of the input features with a superior level of accuracy. In Random Forest training, two parameters in the form of the number of trees and the number of features or variables that have been randomly selected will be used to evaluate each node in the tree [8]. Random Forest is used in this study to create a predictive model in cardiovascular disease.

The Logistic Regression method is also used in this study to build a predictive model in cardiovascular disease. Logistic Regression is a predictive machine learning method that evaluates the relationship between target variables and predictor variables. The target variable or dependent variable is a variable that is categorical data with a nominal or ordinal scale. Predictor variables or independent variables are variables that are categorical data with interval or ratio scales. In Logistic Regression, the target variable is presented as a binary variable with a value of 1 (positive) or 0 (negative)[9]. The target variable in Logistic Regression determines whether a person has cardiovascular disease based on the predictor variables.

In the research conducted by Sharma, cardiovascular disease prediction is carried out using Support Vector Machine, Decision Tree, Naïve Bayes, and Random Forest methods. The dataset used in the study uses a combination of 4 cardiovascular disease datasets from the UCI Machine Learning Repository with a total of about 1025 rows of data and has 14 attributes. These attributes include age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num. In Sharma's research, the Random Forest method has the greatest accuracy rate with a value of 99%. The Support Vector Machine, Decision Tree, and Naïve Bayes methods received accuracy values of 98%, 85%, and 90%, respectively [10].

In the research conducted by Gupta, cardiovascular disease prediction is carried out using Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbors, and Random Forest methods. The dataset used in the study uses 1 cardiovascular disease dataset from the UCI Machine Learning Repository with a total of about 303 rows of data and has 14 attributes. These attributes include age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num. In Gupta's research, the Logistic Regression method has the greatest accuracy with a value of 92.30%. The Support Vector Machine, Naïve Bayes, Decision Tree, K-Nearest Neighbors, and Random Forest methods received accuracy values of 91.20%, 87.91%, 85.71%, 90.11%, and 85.71%, respectively[11].

Based on the results of previous studies, this research chooses to classify cardiovascular diseases using methods that have the greatest accuracy results including Random Forest and Logistic Regression. The significant difference between the two previous studies and this study is that this study classifies cardiovascular disease with a dataset that has attributes related to lifestyle, such as physical activity status, smoking status, glucose level, and alcohol consumption. The cardiovascular disease dataset in this study has 70,000 rows of data with 12 attributes.

Based on the discussion above, this research focuses on how Random Forest and Logistic Regression methods can create a machine learning model to classify whether a person is prone to cardiovascular disease. The two types of machine learning methods used will be compared to see which method can classify heart disease better. Classifying cardiovascular disease is important to prevent or reduce the probability of a person developing cardiovascular disease. By looking at the factors that cause the impact of cardiovascular disease, Indonesians can organize a healthier and more regular lifestyle and diet. Research indicates that the risk of cardiovascular disease can be predicted using logistic regression and random forest approaches, enabling us to identify the relationship between lifestyle and the likelihood of developing such diseases.

2. PROPOSED RESEARCH

The research method contains research procedures that will be carried out to answer the formulation of problems and research objectives based on a scientific foundation. Research on the classification of cardiovascular disease based on a person's lifestyle with machine learning classification has several stages shown in Figure 1. The machine learning classification method used for the classification process is divided into two for comparison of the best methods, namely the Random Forest and Logistic Regression methods.

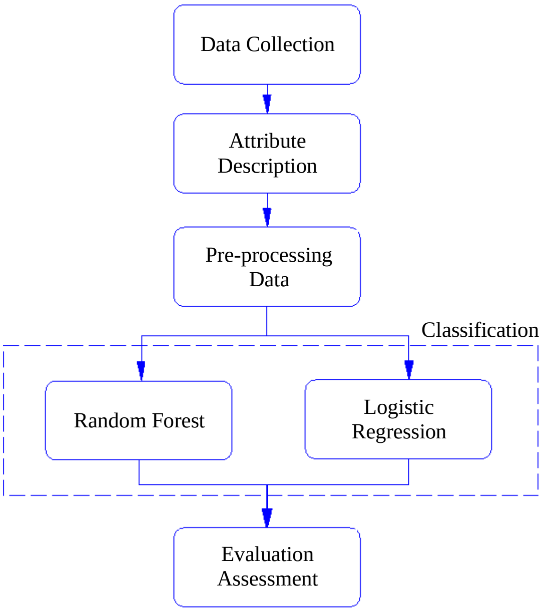

Figure 1. Research Flow Chart

2.1. Data Collection

The data used in this study is secondary data taken from Kaggle in the form of cardiovascular disease datasets. The Cardiovascular Disease dataset was uploaded to Kaggle by data scientist, Svetlana Ulianova [12]. This dataset has a total of 70,000 rows of data with 12 columns or attributes. The attributes in the dataset include information related to lifestyle, such as smoking, alcohol consumption, and physical activity to measure their relationship to a person's likelihood of developing cardiovascular disease. Data collection was done by document study taken from public data provided by Kaggle.

2.2. Attribute Description

The cardiovascular disease dataset has 12 data attributes that will be used to build a machine learning model to classify and analyze the relationship between lifestyle and a person's susceptibility to cardiovascular disease. Of the 12 attributes, 1 attribute will be selected as the target variable and the rest will be used as input variables or predictor variables. Table 1 shows the names of the attributes that will be used. The attribute selected as the target variable is the Cardiovascular Disease attribute which has a binary data type to classify cardiovascular disease.

There are some data rows in the dataset that have unrealistic values. The dataset needs to be processed and some data rows that have unrealistic data need to be eliminated to improve the accuracy of the classification results of the likelihood of a person having cardiovascular disease.

2.3. Pre-processing Data

The initial stage of this research is the data pre-processing stage. In the data processing stage, this process refers to data evaluation, data cleaning, data transformation, and selection of common features to be used in the machine learning model used for the data classification process. The preprocessing stage is carried out to improve model performance and prevent overfitting, so that it can help make it easier for the Random Forest and Logistic Regression methods to classify data and produce a high level of accuracy. Several steps are taken in the data pre-processing process, such as:

Table 1. Attributes in the Dataset

No.	Attribute Name	Description	Value
1.	*Age (age)*	Patient age (days)	10798 days to 23713 days
2.	*Height (height)*	Patient height (cm)	55 cm to 250 cm
3.	*Weight (weight)*	Patient weight (kg)	10 kg to 200 kg
4.	*Gender (gender)*	Patient gender	1: female 2: male
5.	*Systolic Blood Pressure (ap-hi)*	Systolic blood pressure (mmHg)	-150 mmHg to 16020 mmHg
6.	*Diastolic Blood Pressure (ap-lo)*	Diastolic blood pressure (mmHg)	-70 mmHg to 11000 mmHg
7.	*Cholesterol (cholesterol)*	Cholesterol level	1: normal 2: above normal 3: significantly above normal
8.	*Glucose (gluc)*	Blood glucose level	1: normal 2: above normal 3: significantly above normal
9.	*Smoking (smoke)*	Smoking status	0: not smoking 1: smoking
10.	*Alcohol Intake (alco)*	Alcohol consumption status	0: no alcohol consumption 1: consuming alcohol
11.	*Physical Activity (active)*	Physical activity status	0: not physically active 1: physically active
12.	*Cardiovascular Disease (cardio)*	Presence of cardiovascular disease	0: no cardiovascular disease 1: have cardiovascular disease

2.3.1. Missing Value Imputation

Missing value imputation techniques are divided into 2 types, namely statistical-based or machine learning-based techniques. In this research, a machine learning-based imputation technique is chosen to fill in the missing values in the dataset. According to[13] Wei-Chao Lin, it is suggested that when the dataset is divided into training and testing data to search for missing data, it is more practical to make both types of data incomplete rather than focusing on only one type of data. After the missing value imputation process is performed, the divided and processed dataset can be used for the classification process [14]. In this research, data rows that have a value of N/A or do not contain any data at all become the missing value criteria that will be imputed with the median or average method depending on the type of value distribution of each attribute.

2.3.2. Outlier Detection

Machine learning models, especially the Random Forest classification method, are sensitive to outliers. Outliers tend to emphasize statistical rarities and deviations. In machine learning, outliers refer to erroneous data points that make machine learning models more difficult to fit [15]. Outliers in the data will be removed to develop the desired machine learning model.

2.3.3. Extreme Data Value Checking

After the removal of outlier data, if the data content of the cardiovascular disease dataset still has patient data with unrealistic values or data with extreme values, these rows of data will be removed automatically by imposing upper and lower value limits on each problematic attribute. However, attributes that have binary data types will not be given value limits because they already have table values.

2.4. Classification Process

2.4.1. Logistic Regression

The data is preprocessed to remove missing values, convert categorical variables into numerical variables, and standardize the data. After preprocessing, the dataset needs to be split into training and testing data. The training data is used as input into the Logistic Regression method to train the machine learning model, while the testing data is used to evaluate the performance of the Logistic Regression method-based machine learning model in classifying cardiovascular diseases. The scikit-learn library in Python provides a function to split the data into training and testing sets. Furthermore, Logistic Regression needs to be defined and trained. In this study, the proportion of training data and testing data will be divided into several percentage combinations to see what percentage division has the least False Positive and False Negative results. This is to ensure that the classification results provided by the machine learning model are better.

In the classification process for Logistic Regression, different combinations of parameters will be tested to see which combination of parameters can produce the best Logistic Regression classification model. Training a logistic regression model involves estimating the parameters (weights) that best relate the feature variables to the probability of the target variable being in a particular class. Here's a formal model for the training process of logistic regression:

a. Model Specification

Given a dataset $D = \{(x_1, y_1), (x_2, y_2), \dots\dots\dots, (x_n, y_n)\}$, where x_1 represents the feature vector of the $i - th$ instance, and $y_i \in \{0,1\}$ represents the corresponding binary class label for each instance i, logistic regression models the probability that $y_i = 1$ given x_i as follows eqution (1):

$$P\,(y_i = 1 \mid x_i; \theta) = \sigma\,(\theta^T x_i) \tag{1}$$

where:

i. $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic (sigmoid) function.

ii. θ is the parameter vector (including the intercept term), and $\theta^T x_i$ denotes the dot product between θ and x_i.

b. Objective Function (Loss Function)

The objective of training is to find the parameter vector θ that maximizes the likelihood (or equivalently, minimizes the negative log-likelihood) of the observed data in equation (2):

$$L(\theta) = \sum_{i=1}^{n} |y_i \log\big(\sigma(\theta^T x_i)\big) + (1 - y_i) \log(1 - \sigma(\theta^T x_i))\,| \tag{2}$$

The negative log-likelihood, often used as the loss function to be minimized, is equation (3):

$$J(\theta) = -\frac{1}{n}\,L\,(\theta) \tag{3}$$

c. Optimization

The optimization of $J(\theta)$ with respect to θ can be performed using various optimization algorithms. The most common method is gradient descent (or its variants like stochastic gradient descent, mini-batch gradient descent, etc.). The gradient of $J(\theta)$ with respect to θ is equation (4):

$$\nabla_\theta\, J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} (x_i(y_i - \sigma(\theta^T x_i))) \tag{4}$$

The update rule for gradient descent is equation (5):

$$\theta := \theta - \alpha\, \nabla_\theta\, J(\theta) \tag{5}$$

where α is the learning rate, a hyperparameter that determines the step size at each iteration in the parameter space.

d. Regularization

To prevent overfitting, regularization terms might be added to the loss function. For example, L_2 regularization leads to equation (6):

$$J_{reg}\,(\theta) = J\,(\theta) + \frac{\lambda}{2n}\,||\theta||^2 \tag{6}$$

where α is the regularization strength.

e. Convergence

The training process continues iteratively updating θ until convergence, i.e., when changes in the loss function or in the parameters θ fall below a pre-defined threshold, or a maximum number of iterations is reached.

This formal model encapsulates the fundamental aspects of training logistic regression models, from the specification of the model through to the optimization and regularization strategies.

The scikit-learn library provides a LogisticRegression class that can be used for this purpose. The machine learning model is trained on the training set using the fit () function, which estimates the model parameters. Once the model is trained, it is evaluated on the test set. This is done by classifying the class label for the test set using the predict () function and comparing it with the actual class label. In addition, the GridSearchCV (Grid Search Cross-Validation) function is also provided from the scikit-learn library which is used to display the best combination of parameters. Finally, evaluation metrics will be used to assess the performance of the model, such as accuracy, precision, recall, and F1 score.

2.4.2. Random Forest

Machine learning modeling with the Random Forest method will be built with the Python programming language. In building the Random Forest method, the data will be divided into training data and testing data. Random Forest has Information Gain and Entropy measurement variables to determine the quality of features when splitting data. Information Gain in machine learning is used to select features with the best relevance. If a feature has an information gain value below the threshold, it will be removed [16]. Entropy is a tool for extracting incorrect or inappropriate features. Features in a dataset are measured based on their impurity by looking at their high or low entropy level. A dataset that has high entropy is considered impure, and vice versa. Entropy is calculated based on the proportion of each class in the data [17].

In this research, the type of feature assessment criteria in data separation used for the Random Forest classification method is Entropy. Training a random forest model involves several steps. Here's a formal model for the training process as follow:

$Let\ D = \{(x_1, y_1), (x_2, y_2), \dots\dots\dots., (x_n, y_n)\}$ be the training dataset, where x_1 represents the features and y_1 represents the corresponding labels. Each x_i is a vector of features, and each y_i is the corresponding label.

1. Bootstrap Sampling (Bagging):

   $For\ t = 1\ to\ T$ (number of trees in the forest):

   Sample a bootstrap dataset D_t of size n from D with replacement.

2. Tree Construction:

   For each t :

   a. Randomly select a subset of features of size m (typically $\sqrt{p}$ for classification and $\frac{p}{3}$ for regression, where p is the total number of features).

   b. Using the bootstrap dataset D_t and the selected subset of features, grow a decision tree T_t recursively as follows:

      At each node:

      i. Select the best feature among the randomly chosen subset based on a splitting criterion (e.g., Gini impurity for classification, mean squared error for regression).

      ii. Split the node into child nodes based on the selected feature.

      iii. Repeat the process recursively until a stopping criterion is met (e.g., maximum depth reached, minimum samples per leaf reached).

3. Model Aggregation:

   The trained trees $T_1, T_2, \dots\dots\dots, T_T$ form the random forest model.

   a. For classification:

      For each input x :

      i. $Let\ p_{ti}$be the probability of x belonging to class according to tree T_t.

      ii. Aggregate predictions by taking the majority class among all trees.

   b. For regression:

      For each input x

i. $Let\ y_{ti}$be the predicted value of x according to tree T_t.
ii. Aggregate predictions by taking the average (or median) of all y_{ti} values across trees.

This formal model outlines the process of training a random forest, which involves bootstrapping, constructing multiple decision trees with random feature subsets, and aggregating their predictions.

Several Python libraries are imported to process data and run the Random Forest classification function, such as numpy, pandas, matplotlib, and sklearn. Generally, the LabelEncoder function in sklearn is used in data processing before feeding the data into the classification method to convert categorical variables into numerical variables. However, the cardiovascular disease dataset does not have attributes containing categorical variables, so the sklearn function is only used to prepare the data for input to the Random Forest method, such as separating the data into training and testing data. In the Random Forest method, the proportion of training data and testing data will also be divided into several percentage combinations to see what percentage division has the least False Positive and False Negative results just like the Logistic Regression method.

In the classification process for Random Forest, several different combinations of parameters will be tested to see which combination of parameters can produce a Random Forest classification model. The GridSearchCV (Grid Search Cross-Validation) function from the scikit-learn library is used to display the best parameter combinations. The processed data will be entered into the Random Forest classification method with a different number of n_estimators to compare which accuracy results are the highest. In sklearn, the n_estimators command is the number of decision trees that will be made before taking the most votes or the average classification results from each decision tree. The classification results of the Random Forest method will be measured by evaluation metrics to assess the performance of machine learning models, such as accuracy, precision, recall, and F1 score.

2.5. Evaluation Assessment

Assessment evaluation is made to compare the accuracy results obtained between the Random Forest and Logistic Regression methods. The evaluation is based on the Confusion Matrix obtained from the classification results of each machine learning method. The assessment evaluation is built with Python programming with metrics features from the sklearn library. The assessment evaluation will be used to assess the accuracy of cardiovascular disease classification, such as accuracy, precision, recall, and F1-score. The calculation of the evaluation assessment can refer to the equation formula in the literature review chapter in sub-chapter 2.5. The machine learning method with the highest accuracy result will be selected as the best classification method in classifying a person's susceptibility to cardiovascular disease.

3. RESULTS AND DISCUSSION

3.1. Data Pre-Processing Results

The id attribute in the dataset was removed as it was not used. The dataset was checked to find and remove duplicate data rows. Then, in processing the numeric data attributes, such as age, height, weight, ap_hi, and ap_lo, they were calculated to find the interquartile range based on the 1st and 3rd quartile values of each attribute. These values will be used to find the upper and lower outliers that will be removed from the data. Then, in the Systolic Blood Pressure (ap-hi) and Diastolic Blood Pressure (ap-lo) attributes, there are data rows with extreme values that are removed from the dataset. Data rows with ap_hi attribute values exceeding 250 and less than 0 were removed from the dataset, while for the ap_lo attribute, data rows with values exceeding 200 and less than 0 were removed. In the process of handling missing values, the SimpleImputer feature of the Scikit-learn module was used to replace missing values by using the median value in each column. There are no empty values in the dataset, but there are 23 rows that have duplicate sets of values. In addition, the number of deleted data rows after the outlier and extreme data removal process amounted to 7498 data rows with the total data rows being 62478 data rows.

In this research, the BMI attribute is created by calculating the weight and height attributes and added to the data frame. The BMI formula used in this study uses BMI with a metric system. The total rows and columns in the dataset that will be used for the Logistic Regression and Random Forest classification process are 62478 rows of data with 13 columns or attributes. All data rows in the dataset do not have empty or missing values for the processing of both machine learning methods.

In Figure 2, the heatmap table is used to show the degree of correlation each attribute has. Some significant correlations can be seen, such as:
1) Height attributes have a strong correlation with gender attributes
2) Attribute weight has a very strong correlation with attribute bmi
3) The ap_lo attribute has a strong correlation with the ap_hi attribute
4) The ap_hi attribute has a strong correlation with the cardio attribute

5) The smoke attribute has a strong correlation with the gender attribute
6) The gluc attribute has a strong correlation with the cholesterol attribute
7) The cholesterol attribute has a strong correlation with the cardio attribute
8) The ap_lo attribute has a strong correlation with the cardio attribute

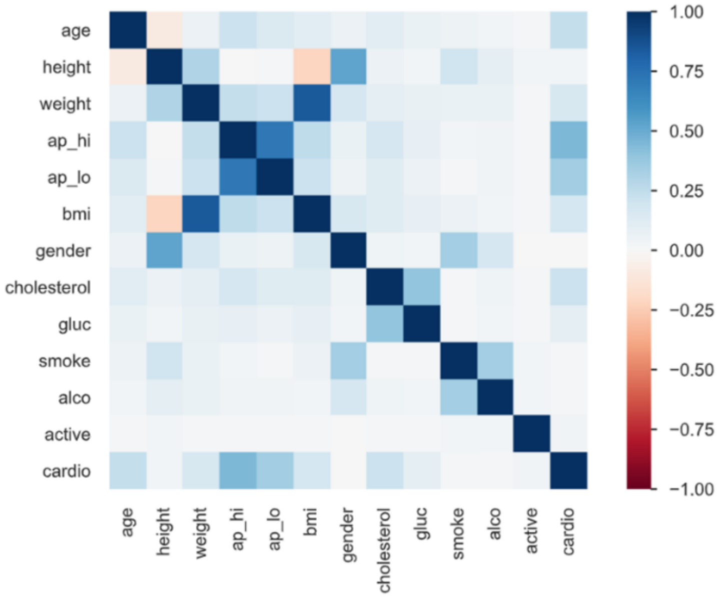

Figure 2. Heatmap of Data Attributes

In Table 2, the numeric attributes are described to see the mean, min, max, distinct, and std values. Table 3 describes the frequency of categorical attribute values in the dataset. Glucose (gluc), Smoke, and Alcohol Intake (alco) attributes have unbalanced values, i.e. one value or category has a very high frequency compared to other values or categories.

Table 2. Description of Numeric Attribute Values in the Data

	Mean	*Min*	*Max*	*Distinct*	*STD*
Age	19493.854	14282	23713	7992	2458.222
Height	164.40606	143	186	44	7.5324356
Weight	73.184049	40	107	68	12.273306
BMI	26.628765	13	50	38	4.5841866
Ap-Hi	126.42296	90	170	75	14.292042
Ap-Lo	81.699958	65	105	41	7.6746608

Table 3. Frequency of Categorical Attribute Values in the Data

	0	1	2	3
Gender		40694	21784	
Cholesterol		47187	8224	7067
Glucose		53381	4403	4694
Smoke	57081	5397		
Alcohol Intake	59217	3261		
Physical Activity	12227	50251		
Cardiovascular Disease	31615	30863		

3.2. Logistic Regression Results

Before the dataset is entered into the Logistic Regression model, the dataset is divided into training data and testing data into 4 types of percentage distribution for the iteration process. The percentage distribution of training and testing datasets of the whole dataset is divided into 0.6 and 0.4 (60% and 40%); 0.7 and 0.3 (70% and 30%); 0.8 and 0.2 (80% and 20%); 0.9 and 0.1 (90% and 10%). In the Logistic Regression process, features are used from the Scikit-learn module, namely StandardScaler, GridSearchCV, and LogisticRegression. After the training and testing data were standardized with StandardScaler, the LogisticRegression process was run using various combinations of parameters. Table 4 shows the parameters and their values used in the LogisticRegression process.

Table 4. Parameters Used for Logistic Regression Method

Parameter	Tested Parameter Values
Penalty	*l1*, *l2*, dan *elasticnet*
C	0.01, 0.1, dan 1.0
class_weight	*None* dan *balanced*
fit_intercept	*True* dan *False*
tol	1e-4, 1e-3, dan 1e-2
warm_start	*True* dan *False*
l1_ratio	0.0, 0.5, dan 1.0

In processing various combinations of parameter values for LogisticRegression, the GridSearchCV feature is used so that the results released by the machine learning model are only the best results from all combinations of parameter values (Pedegrosa et al., 2011). Table 5 shows the best parameters used for each percentage distribution of the test data.

Table 4. Best Parameter Selection for Logistic Regression

Testing Data Distribution	*C*	*class_weight*	*fit_intercept*	*l1_ratio*	*penalty*	*tol*	*warm_start*
40%	1.0	*balanced*	*True*	0.5	*elasticnet*	0.0001	*True*
30%	1.0	*balanced*	*True*	0.0	*l1*	0.0001	*True*
20%	0.1	*balanced*	*True*	0.5	*elasticnet*	0.01	*True*
10%	1.0	*balanced*	*True*	0.0	*l2*	0.001	*True*

3.2.1. Classification Results of 40% Testing Data Distribution

Table 5 shows the confusion matrix results for the 40% test data distribution with an error value of 27.64%. Table 6 shows 5 important attributes based on the largest feature importance value. The attribute that most influences the classification process is the ap_hi attribute.

Table 5. Confusion Matrix Distribution of Testing Data 40% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	10016 (*True Negative*)	2737 (*False Positive*)
Actual Positive	4172 (*False Negative*)	8067 (*True Positive*)
Error Formula	$Error = \frac{FN + FP}{TN + TP + FN + FP}$ $Error = \frac{4172 + 2737}{8067 + 10016 + 4172 + 2737}$	
Error	0.2764484635 atau 27.64%	

Table 6. Attributes with Highest Feature Importance Value

Attribute	Value
ap_hi (*Systolic Blood Pressure*)	0.884450
cholesterol	0.342455
age	0.330550
weight	0.264142
bmi	0.115820

Table 6 shows the values of the 5 attributes that have the highest feature importance values. The feature importance method used is the dependent feature importance model for logistic regression.

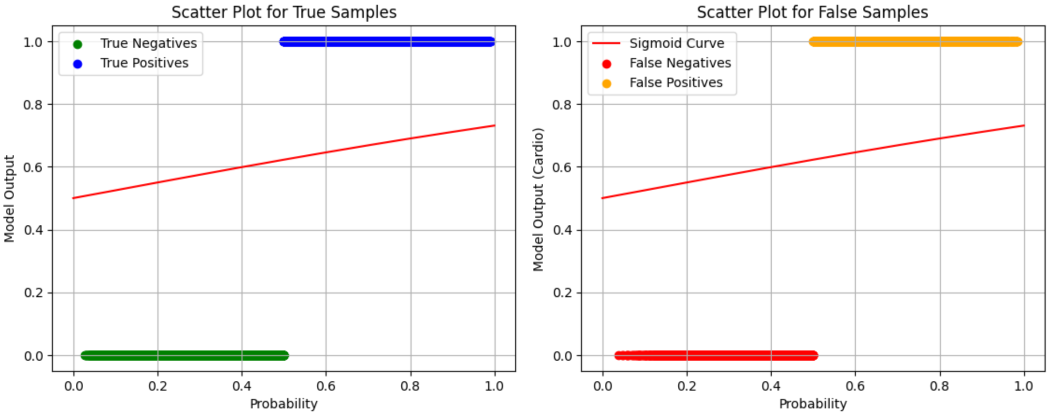

Figure 3. Sigmoid Graph of 40% Testing Data Distribution

Figures 3, 4, 5, and 6 depict the Sigmoid graph of testing data distribution for true samples and false samples. The x-axis represents the probability of the data, and the y-axis represents the model output consisting of values 0 and 1. A value of 0 indicates the class without cardiovascular disease, while a value of 1 indicates the class with cardiovascular disease. The data distribution shows points above or below the sigmoid curve. These points are interpreted as points above the sigmoid curve representing data predicted as positive class, while points below the sigmoid curve represent data predicted as negative class. Meanwhile, the red line is interpreted as the separating line between two different classes or data groups, namely positive class, and negative class. To determine positive and negative classes based on their probability values. If the probability is below 0.5, then the data has a value of 0, meaning the data belongs to the class without cardiovascular disease, and if the probability is above 0.5, then the data has a class of 1, meaning the data belongs to the class with cardiovascular disease. The red line also represents a trend or pattern in the data. The data used shows a trend or pattern of population growth, where in this case it represents the growth or increase in the probability of someone being affected by cardiovascular disease.

Table 7. Data Description of Sigmoid Graph

Sample	Color Point	Axis Position	Description
True Positive	Blue	y = 1 0.5 < x < 1	Correct classification result that the patient has cardiovascular disease.
True Negative	Green	y = 0 0 < x < 0.5	Correct classification result that the patient does not have cardiovascular disease.
False Positive	Orange	y = 1 0.5 < x < 1	Incorrect classification result that the patient has cardiovascular disease.
False Negative	Red	y = 0 0 < x < 0.5	Incorrect classification result that the patient does not have cardiovascular disease.

Table 7 explains the description of the sigmoid graph. The y-axis represents the cardiovascular disease classification result with a value of 0 indicating the patient does not have cardiovascular disease and a value of 1 indicating the patient has cardiovascular disease. The x-axis represents the range of probability values of samples belonging to the output class 0 or 1 with a probability value of 1 representing a definite classification of having cardiovascular disease and 0 representing a definite classification of not having cardiovascular disease. In the graph, True Samples are the data results that were successfully classified correctly (True Positives and True Negatives) and False Samples are the data results that were misclassified with the true results (False Positives and False Negatives).

3.2.2 Classification Results of 30% Testing Data Distribution

Table 8 shows the confusion matrix results for the 30% test data distribution with an error value of 27.47%. Table 9 shows 5 important attributes based on the largest feature importance value. The attribute that most influences the classification process is the ap_hi attribute.

Table 8. Confusion Matrix Distribution of Testing Data 30% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	7561 (*True Negative*)	2037 (*False Positive*)
Actual Positive	3113 (*False Negative*)	6033 (*True Positive*)
Error	0.2747545881 atau 27.47%	

Table 9. Five Attributes with the Highest Feature Importance Value

Attribute	Value
ap_hi (*Systolic Blood Pressure*)	0.884450
cholesterol	0.342455
age	0.330550
weight	0.264142
bmi	0.115820

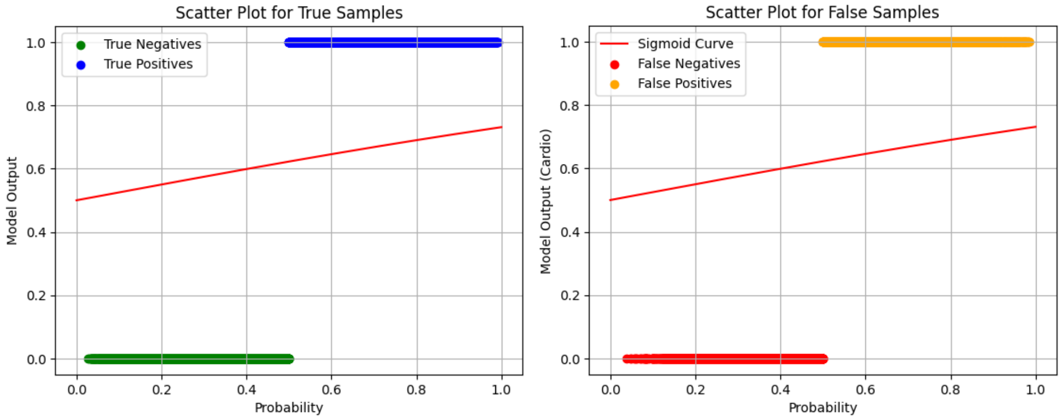

Figure 4. Sigmoid Distribution Graph of 30% Testing Data

3.2.3. Classification Results of 20% Testing Data Distribution

Table 10 shows the confusion matrix results for 20% test data distribution with an error value of 26.98%. Table 11 shows 5 important attributes based on the largest feature importance value. The attribute that most influences the classification process is the ap_hi attribute.

Table 10. Confusion Matrix Distribution of Testing Data 20% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	5082 (*True Negative*)	1350 (*False Positive*)
Actual Positive	2022 (*False Negative*)	4042 (*True Positive*)
Error	0.2698463508 atau 26.98%	

Table 11. Five Attributes with the Highest Feature Importance Value

Attribute	Value
ap_hi (*Systolic Blood Pressure*)	0.884450
cholesterol	0.342455
age	0.330550
weight	0.264142
bmi	0.115820

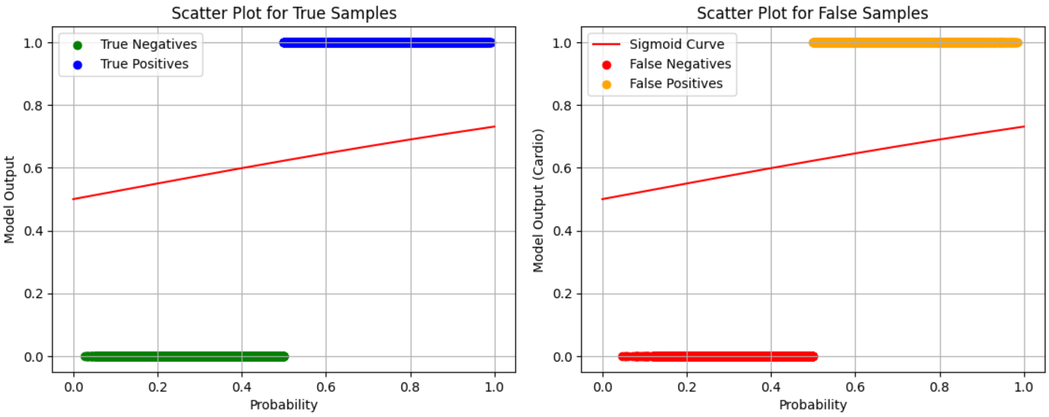

Figure 5. Sigmoid Distribution Graph of 20% Testing Data

3.2.4 Classification Results of 10% Testing Data Distribution

Table 12 shows the confusion matrix results for the 10% test data distribution with an error value of 26.92%. Table 13 shows 5 important attributes based on the largest feature importance value. The attribute that most influences the classification process is the ap_hi attribute.

Table 12. Confusion Matrix Distribution of Testing Data 10% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	2551 (*True Negative*)	677 (*False Positive*)
Actual Positive	1005 (*False Negative*)	2015 (*True Positive*)
Error	0.269206146 atau 26.92%	

Table 13. 5 Attributes with the Highest Feature Importance Value

Attribute	Value
ap_hi (*Systolic Blood Pressure*)	0.884450
Cholesterol	0.342455
Age	0.330550
Weight	0.264142
Bmi	0.115820

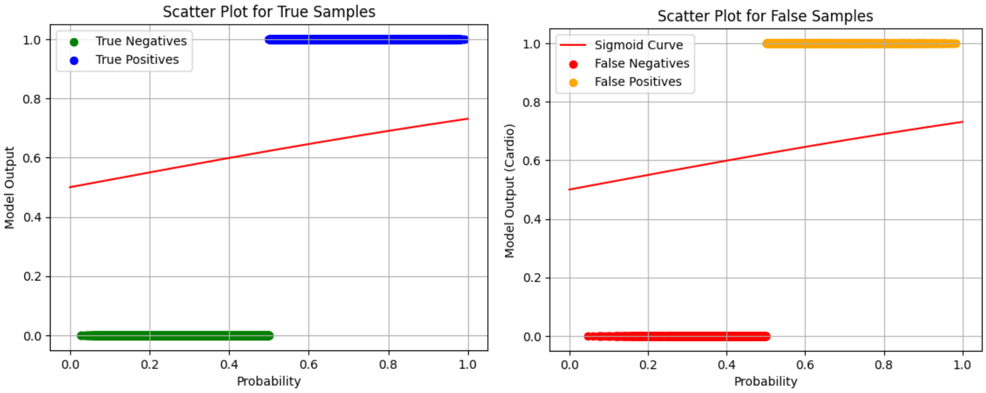

Figure 6. Sigmoid Graph of 10% Testing Data Distribution

In Table 14, the results of accuracy, precision, recall, and f1-score values for each percentage of test data distribution are shown. The best accuracy results obtained for the Logistic Regression model were obtained at a percentage distribution of 10% test data with an accuracy value of 73.07%.

Table 14. Results of the Logistic Regression Method Evaluation Matrix

Testing Data Distribution	*Accuracy*	*Precision*	*Recall*	*F1-Score*
40%	72,35%	74,66%	65,91%	70,01%
30%	72,52%	74,75%	65,96%	70,08%
20%	73,01%	74,96%	66,65%	70,56%
10%	73,07%	74,85%	66,72%	70,55%

3.3. Random Forest Results

Before the dataset is input into the Random Forest model, the dataset is divided into training data and testing data into 4 types of percentage distribution for the iteration process. The percentage distribution of training and testing datasets of the whole dataset is divided into 0.6 and 0.4 (60% and 40%); 0.7 and 0.3 (70% and 30%); 0.8 and 0.2 (80% and 20%); 0.9 and 0.1 (90% and 10%). In the Random Forest process, features are used from the Scikit-learn module, namely StandardScaler, GridSearchCV, and RandomForestClassifier. After the training and testing data were standardized with StandardScaler, the RandomForestClassifier process was run using various combinations of parameters. Table 15 shows the parameters and their values used in the RandomForestClassifier process.

Table 15. Parameters Used for Random Forest Method

Parameter	Tested Parameter Values
n_estimators	50, 100, 150, 300, dan 500
max_depth	10, 15, dan 20
min_samples_split	2, 3, dan 5
min_samples_leaf	2, 3, dan 5
max_samples	0.4, 0.5, dan 0.7
max_features	*sqrt*, *log2*, dan *None*
min_weight_fraction_leaf	0.2, 0.3, dan 0.5
criterion	*gini* dan *entropy*
ccp_alpha	0.0 dan 0.1

In processing various combinations of parameter values for the RandomForestClassifier, the GridSearchCV feature is used so that the results released by the machine learning model are only the best results from all combinations of parameter values. The Random Forest classification process requires a very long processing time compared to the Logistic Regression classification process due to the large number of parameter combinations. Table 16 shows the best parameters used for each percentage of the test data distribution.

Table 16. Best Parameter Selection for Random Forest

Testing Data Distribution	*criterion*	*max_ depth*	*max_ features*	*max_ samples*	*min_ samples_ leaf*	*min_ samples_ split*	*min_ weight_ fraction_ leaf*	*n_ estimators*
40%	*entropy*	10	*sqrt*	0.5	2	2	0.3	100
30%	*entropy*	10	*sqrt*	0.4	2	2	0.2	50
20%	*entropy*	10	*sqrt*	0.4	2	2	0.2	50
10%	*entropy*	10	*sqrt*	0.5	2	2	0.3	100

3.3.1 Classification Results of 40% Testing Data Distribution

Table 17 displays the confusion matrix results for the 40% test data distribution with an error value of 26.92%. The feature importance value obtained for the 40% test data distribution is only found in one attribute, namely the ap_hi (Systolic Blood Pressure) attribute with a feature importance value of 0.2091816048. The feature importance method used is the dependent feature importance model for random forests.

Table 17. Confusion Matrix Distribution of Testing Data 40% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	10333 (*True Negative*)	2420 (*False Positive*)
Actual Positive	4805 (*False Negative*)	7434 (*True Positive*)
Error	0.2890925096 atau 28.90%	

4.3.2 Classification Results of 30% Testing Data Distribution

Table 18 displays the confusion matrix results for the 30% test data distribution with an error value of 28.77%. The feature importance value obtained for the 30% test data distribution is only found in one attribute, namely the ap_hi (Systolic Blood Pressure) attribute with a feature importance value of 0.2100512164.

Table 18. Confusion Matrix Distribution of Testing Data 30% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	7787 (*True Negative*)	1811 (*False Positive*)
Actual Positive	3582 (*False Negative*)	5564 (*True Positive*)
Error	0.2877187367 atau 28.77%	

4.3.3 Classification Results of 20% Testing Data Distribution

Table 19 displays the confusion matrix results for the 20% test data distribution with an error value of 28.21%. The feature importance value obtained for the 20% test data distribution is only found in one attribute, namely the ap_hi (Systolic Blood Pressure) attribute with a feature importance value of 0.2149861289.

Table 19. Confusion Matrix of Testing Data Distribution 20% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	5236 (*True Negative*)	1196 (*False Positive*)
Actual Positive	2330 (*False Negative*)	3734 (*True Positive*)
Error	0.2821702945 atau 28.21%	

4.3.4 Classification Results of 10% Testing Data Distribution

Table 20 displays the confusion matrix results for the 10% test data distribution with an error value of 28.12%. The feature importance value obtained for the 10% test data distribution is only found in one attribute, namely the ap_hi (Systolic Blood Pressure) attribute with a feature importance value of 0.2146286812.

Table 20. Confusion Matrix of Testing Data Distribution 10% of Total Data

	Predicted Negative	*Predicted Positive*
Actual Negative	2614 (*True Negative*)	614 (*False Positive*)
Actual Positive	1143 (*False Negative*)	1877 (*True Positive*)
Error	0.2812099872 atau 28.12%	

In Table 21, the results of accuracy, precision, recall, and f1-score values for each percentage of test data distribution are shown. The best accuracy results obtained for the Random Forest model were obtained at a percentage distribution of 10% test data with an accuracy value of 71.87%.

Table 21. Random Forest Evaluation Matrix Results

Testing Data Distribution	*Accuracy*	*Precision*	*Recall*	*F1-Score*
40%	71,09%	75,44%	60,74%	67,29%
30%	71,22%	75,44%	60,83%	67,35%
20%	71,78%	75,74%	61,57%	67,92%
10%	71,87%	75,35%	62,15%	68,11%

3.4. Comparison of Random Forest and Logistic Regression Method Results

This research involves twelve factors or attributes that affect the risk of cardiovascular disease as variables to build Logistic Regression and Random Forest classification models. Among these variables, a significant relationship was found between variables and variables or attributes of cardiovascular disease. This study uses Logistic Regression and Random Forest as methods to classify cardiovascular disease based on the available attributes. By comparing the two methods, the Logistic Regression method is a more effective and efficient algorithm in classifying cardiovascular disease compared to the Random Forest method.

In the process of finding the best parameter combination for the Random Forest and Logistic Regression methods, the process of finding parameter combinations for the Random Forest method takes quite a long time. The total combination of Random Forest parameters that need to be tested is 14,580 parameter combinations. The time required in the Random Forest method to produce a judgment evaluation for one distribution of test data can take about 4 or 8 hours to more than a day depending on the number of parameter combinations tested. Meanwhile, the process of finding parameter combinations in the Logistic Regression method takes a faster time. The total combination of Logistic Regression parameters that need to be tested is 648 parameter combinations. The time required in the Logistic Regression method to produce an assessment evaluation with one distribution of test data can take about 30 minutes to 1 hour.

Table 22. Comparison Table of Logistic Regression and Random Forest Results

Testing Data Distribution	Training Data Distribution	*Accuracy*		*Precision*		*Recall*		*F1-Score*	
		LR	*RF*	*LR*	*RF*	*LR*	*RF*	*LR*	*RF*
40%	60%	72,35%	71,09%	74,66%	75,44%	65,91%	60,74%	70,01%	67,29%
30%	70%	72,52%	71,22%	74,75%	75,44%	65,96%	60,83%	70,08%	67,35%
20%	80%	73,01%	71,78%	74,96%	75,74%	66,65%	61,57%	70,56%	67,92%
10%	90%	73,07%	71,87%	74,85%	75,35%	66,72%	62,15%	70,55%	68,11%
Previous research [17]		72%	(69 –71) %						
Previous research [18]		-	71.91%						

Based on Table 22, both methods got the highest accuracy results on the distribution of test data by 10% of the total dataset. The Logistic Regression method gets an accuracy result of 73.07% with an error rate of 26.92%. The Random Forest method gets an accuracy result of 71.87% with an error rate of 28.12%. By modeling data and classifying data using Logistic Regression and Random Forest, it was found that not all attributes have a significant influence on the classification model of cardiovascular disease. In the Logistic Regression method with a 10% test data distribution, the 5 attributes that have the highest feature importance values are ap-hi, age, cholesterol, weight, and ap-lo. While in the Random Forest method with a 10% distribution of test data, the attribute that appears to have the highest feature importance value is owned by the ap-hi attribute. This is in accordance with the statement of the Centers for Disease Control and Prevention (CDC) that blood pressure, especially high blood pressure, is a major risk factor for cardiovascular disease and causes more than 10 million deaths worldwide each year [19].

4. CONCLUSION

Based on the previous description, it can be concluded in the research "Classification of Cardiovascular Diseases Based on Lifestyle Using Random Forest and Logistic Regression Methods" is cardiovascular disease is significantly associated with lifestyle factors, as evidenced by the feature importance values of the data attributes indicating their correlation with cardiovascular disease. Additionally, attributes such as weight, BMI, and gender exhibit strong associations with lifestyle factors or cardiovascular disease. Pre-processing of the cardiovascular disease dataset is essential for accurate classification. This involves handling missing values through imputation, identifying, and removing outliers and extreme data points. After pre-processing, the dataset contains 62,478 data rows, ensuring a more robust dataset for classification. These steps aim to minimize noise or abnormal data, thereby enhancing the accuracy of classification results. The processed dataset is then utilized for classification using machine learning methods, assessing attribute importance through feature importance, and evaluating classification outcomes. Logistic Regression outperforms Random Forest in accurately classifying cardiovascular disease based on lifestyle attributes from the available dataset. Both methods achieve their highest accuracy when the dataset is divided into 90% training data and 10% testing data. The Logistic Regression method attains a maximum accuracy of 73.07%, whereas the Random Forest method achieves a maximum accuracy of 71.87%.

REFERENCES

[1] World Health Organization. (2018). Noncommunicable diseases country profiles 2018. Geneva: World Health Organization, 2018. [cited 2 November 2022]. Available from: https://apps.who.int/iris/handle/10665/274512.

[2] Prakasa, R. A., Valentina, D. C. D., Abdiana, R., Handayani, R., & BP, N. I. (2020). Analisis Faktor Risiko Pasien Gagal Jantung dengan Reduced Ejection Fraction di RSUD Dr. H. Abdul Moeloek Provinsi Lampung. Essential: Essence of Scientific Medical Journal, 18(1), 22. https://doi.org/10.24843/estl.2020.v18.i01.p07

[3] Karyatin, K. (2019). Faktor-Faktor Yang Berhubungan Dengan Kejadian Penyakit Jantung Koroner. Jurnal Ilmiah Kesehatan, 11(1), 37-43. https://doi.org/10.37012/jik.v11i1.66

[4] Mattioli, A. V., & Puviani, M. B. (2020). Lifestyle at Time of COVID-19: How Could Quarantine Affect Cardiovascular Risk. American Journal of Lifestyle Medicine, 14(3), 240–242. https://doi.org/10.1177/1559827620918808

[5] Shan, Z., Li, Y., Baden, M. Y., Bhupathiraju, S. N., Tang, B. Z., Hu, F. B., Rexrode, K. M., Rimm, E. B., Qi, L., Willett, W. C., Manson, J. E., Qi, Q., & Hu, F. B. (2020). Association Between Healthy Eating Patterns and Risk of Cardiovascular Disease. JAMA Internal Medicine, 180(8), 1090. https://doi.org/10.1001/jamainternmed.2020.2176

[6] Wardah Hanifah., Wanda Septi Oktavia., dan Hoirun Nisa.,(2021). Lifestyle Factors And Coronary Heart Disease: A Systematic Review Among Indonesian Adults., The Journal of Nutrition and Food Research., 44(1), 45-58.

[7] Charbuty, B., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

[8] Petkovic, D., Altman, R., Wong, M., & Vigil, A. (2018). Improving the explainability of Random Forest classifier–user centered approach. In PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium (pp. 204-215). https://doi.org/10.1142/9789813235533_0019

[9] Ciu, T., & Oetama, R. S. (2020). Logistic Regression Prediction Model for Cardiovascular Disease. International Journal of New Media Technology, 7(1), 33–38. https://doi.org/10.31937/ijnmt.v7i1.1340

[10] Sharma, V., Yadav, S., & Gupta, M. (2020). Heart Disease Prediction using Machine Learning Techniques. In 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). https://doi.org/10.1109/icacccn51052.2020.9362842

[11] Gupta, C. K., Saha, A., Reddy, N., & Acharya, U. D. (2022). Cardiac Disease Prediction using Supervised Machine Learning Techniques. Journal of Physics: Conference Series, 2161(1), 012013. https://doi.org/10.1088/1742-6596/2161/1/012013

[12] Cardiovascular Disease dataset. (2019, January 20). Kaggle. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[13] Zhou, Q., Lan, W., Zhou, Y., & Mo, G. (2020). Effectiveness Evaluation of Anti-bird Devices based on Random Forest Algorithm. 2020 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS). https://doi.org/10.1109/iccss52145.2020.9336891

[14] Lin, W. C., & Tsai, C. F. (2019). Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53(2), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4

[15] Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier Detection. ACM Computing Surveys, 53(3), 1–37. https://doi.org/10.1145/3381028

[16] Daeli, N. O. F., & Adiwijaya, A. (2020, May 11). Sentiment Analysis on Movie Reviews using Information Gain and K-Nearest Neighbor. Journal of Data Science and Its Applications, 3(1), 1-7. https://doi.org/10.34818/jdsa.2020.3.22

[17] Wei, Y., Yang, Y., Xu, M., & Huang, W. (2021). Intelligent fault diagnosis of planetary gearbox based on refined composite hierarchical fuzzy entropy and random forest. ISA Transactions, 109, 340–351. https://doi.org/10.1016/j.isatra.2020.10.028

[18] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. https://doi.org/10.1109/iciccs48265.2020.9121169

[19] Martins, B. A., Ferreira, D., Neto, C., Abelha, A., & Machado, J. (2021). Data mining for cardiovascular disease prediction. Journal of Medical Systems, 45(1). https://doi.org/10.1007/s10916-020-01682-8

[20] Centers for Disease Control and Prevention. (2021). Cardiovascular Diseases [Fact Sheet]. https://www.cdc.gov/globalhealth/healthprotection/ncd/cardiovascular-diseases.html

BIOGRAPHY OF AUTHORS

Ajyan Brava Bietrosula, I am a 2019 Information Systems student from Universitas Airlangga, with a focus on Business Intelligence. Research areas include Machine Learning, and Data Mining

Indah Werdiningsih, graduated with a bachelor’s degree in mathematics from the Sepuluh Nopember Institute of Technology (ITS), Surabaya, and obtained a master’s degree in computer engineering from the Sepuluh Nopember Institute of Technology (ITS), Surabaya. Currently I am a lecturer in the Information Systems Study Program at Airlangga University, Surabaya. Research interests include Artificial Intelligence, Expert Systems, and Data Mining.

Eto Wuryanto, graduated with a bachelor’s degree in mathematics from Universitas Airlangga, Surabaya, and obtained a master’s degree in Bioatatistique from Universite Montpellier II, France. Currently I am a lecturer in the Information Systems Study Program at Universitas Airlangga, Surabaya. Research interests: Soft computing, Metaheuristic optimization and Artificial Intelligence.