Social Science and
Everant Journal
Humanities Journal

ISSN: 2456-2653
Volume 09 Issue 01 January 2025

https://sshjournal.com/
Impact Factor: 2024: 6.576
2023: 5.731

DOI: https://doi.org/10.18535/sshj.v9i01.1444

# Heart Disease Prediction with Decision Tree

**Nicholas[1, a *] | Genrawan Hoendarto[2,b] | Jimmy Tjen[3,c]**

[1,2,3]Widya Dharma Pontianak University, Pontianak, West Kalimantan, Indonesia
[a]21421390_nicholas@widyadharma.ac.id, [b]genrawan@widyadharma.ac.id, [c]jimmytjen@widyadharma.ac.id

*Received 10-12-2024*
*Revised 11-12-2024*
*Accepted 02-01-2025*
*Published 04-01-2025*

**Abstract:**

Heart disease remains a major global health issue, which emphasizes the need for accurate prediction models to aid early diagnosis and effective intervention. This study explores the use of a Decision Tree algorithm to predict heart disease. The dataset used consists of 272 entries, with seven key variables including age, sex, blood pressure, cholesterol, fasting blood sugar, maximum heart rate, and ST depression. The data underwent preprocessing to handle missing values and convert categorical data into numerical format. The model was trained with 80% of the data and tested with the remaining 20%. Evaluation metrics such as accuracy, precision, recall, and f1-score, were used to evaluate the model's performance. The results demonstrated the model's efficacy, achieving an accuracy of 81.48%, a recall of 82.93%, a precision of 91.89%, and an f1-score of 87.18%. These results highlight the potential of the Decision Tree Algorithm in heart disease prediction, particularly for its simplicity and interpretability. Despite the study's limitations, such as the small dataset size that could affect generalizability, this study demonstrates significant predictive potential and a strong foundation for future work. Future research should explore alternative machine learning algorithms to improve prediction accuracy and enhance the model's robustness for real-world application.

**Keywords:** Decision Tree, Heart Disease, Prediction.

## Introduction:

Heart disease is one of the leading causes of death worldwide. According to a World Health Organization (WHO) report, approximately 17,9 million people die each year from cardiovascular disease, which accounts for 31% of all global deaths [1]. Early diagnosis and prediction of heart disease can go a long way in reducing these deaths by enabling faster and more appropriate interventions [2]. One method that is currently being increasingly used for heart disease prediction is machine learning, specifically through its various algorithms. Machine learning includes various algorithms that can be used to handle complex data that is easily interpreted by humans. These algorithms work by deeply analyzing data to find significant patterns that can be used for classification or prediction [3]. In the context of heart disease prediction, machine learning can be used to model the relationship between various risk factors and the likelihood of heart disease [4].

Prior studies on heart disease prediction have employed various machine-learning models to improve early diagnosis and intervention. Several models have been proposed, such as Support Vector Machines, Neural Networks, and Decision

Trees. For instance, research by Santos et al. [3] utilized decision trees and artificial immune systems for stroke prediction, while Chandrasekhar and Peddakrishna [5] enhanced heart disease prediction accuracy through machine learning techniques. Studies have consistently shown that machine learning models can be valuable tools for predicting heart disease by identifying patterns in patient data that may not be immediately apparent to healthcare professionals.

The novelty in this study is highlighted by its application of a Decision Tree Algorithm to predict heart disease using a dataset of 272 entries, focusing on seven key variables, including age, cholesterol, and blood pressure. The model's primary advantage lies in its simplicity and interpretability, which allows for the identification of critical risk factors associated with heart disease. Unlike previous approaches that consider a wider range of features, this study specifically optimizes the Decision Tree for higher accuracy despite using a relatively small dataset.

The work of Baghdadi et al. [2] on advanced machine learning techniques for cardiovascular disease detection and Bhatt et al. [4] on effective heart disease prediction with machine learning provides a solid foundation for this study. The dataset used in this research is the "Heart Disease Prediction" dataset by Rishi Damarla [6], builds on established findings, incorporating key variables identified in previous studies as critical for heart disease diagnosis, such as cholesterol levels [7], blood pressure [8], and heart rate [9]. Additionally, this study draws from findings by Ozcan and Peker [5], who applied classification and regression tree algorithms for heart disease modeling, further validating the use of decision trees in this context.

In this study, we developed a heart disease prediction model using several machine learning algorithms. The model uses several important variables that have been shown to have a significant relationship with heart disease risk, namely age, gender, cholesterol level, blood pressure, heart rate, and depression ST segment [10]. Age is one of the main risk factors, where the risk of heart disease increases with age. According to the American College of Cardiology, men over 45 years old and women over 55 years old have a higher risk of heart disease [7]. Gender also plays an important role in heart disease risk. Research shows that men have a higher risk of developing heart disease at a younger age than women, although the risk in women increases after menopause [7]. Cholesterol levels are another important indicator, High levels of LDL cholesterol and low HDL cholesterol are strongly associated with an increased risk of atherosclerosis, which can lead to heart disease [8].

High blood pressure or hypertension is another major risk factor for heart disease [11]. Hypertension can damage arteries and accelerate the process of atherosclerosis [9]. Heart rate provides an important indication of heart health. An abnormal heartbeat, either too fast or too slow, can be a sign of serious cardiovascular problems [12]. Using these variables, the machine learning model we developed aims to provide accurate and reliable predictions of a person's likelihood of suffering from heart disease. In addition, the model is expected to assist in identifying the key risk factors that contribute to such predictions, so that it can be used as an aid in clinical decision-making.

A decision tree is a powerful and interpretable machine-learning algorithm used for both classification and regression tasks. It operates by recursively partitioning the dataset into subsets based on the value of input features. This process continues until the algorithm determines that further partitioning would not significantly improve the model's performance [13]. In a decision tree, each internal node represents a decision based on a specific feature, each branch represents the outcome of that decision, and each leaf node represents a class label (for classification) or a continuous value (for regression). The goal of the decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [13].

This paper is divided into four main sections. The first section, the introduction, outlines the research objectives, the issues being tackled, and the proposed solutions. The second section details the Decision Tree algorithm for predicting heart disease. The third section showcases the simulation results based on the previously described algorithm. Lastly, the fourth section concludes the research and suggests potential directions for future work.

## Method:

In this study, we utilized the decision tree algorithm to predict heart disease. The decision tree algorithm is advantageous due to its simplicity and interpretability, making it a popular choice for medical diagnoses. Decision tree learning involves using a decision tree as a predictive model to map observations about an item to conclusions about its target value. It is a popular approach in statistics, data mining, and machine learning. When the target variable has a finite set of possible values, the model is called a classification tree. In these trees, the leaves represent class labels, and the branches represent combinations of features that lead to those labels. When the target variable can take continuous values, usually real numbers, the model is referred to as a regression tree. The goal of our proposal method is to enhance the efficacy and accuracy of heart disease  prediction models. The steps involved in developing our heart disease prediction model are as follows in Fig. 1.

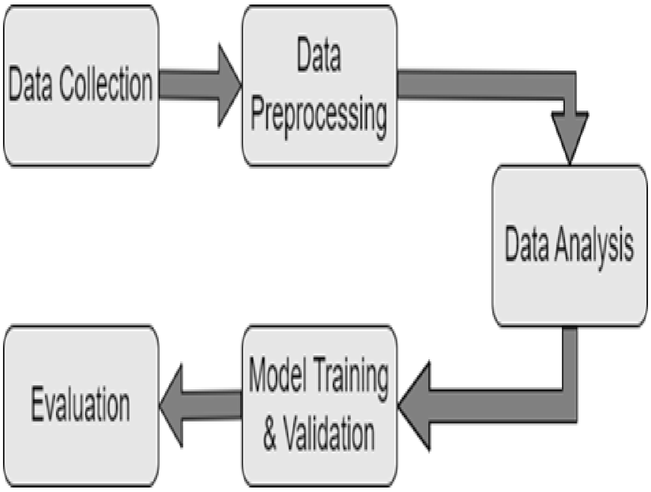

**Fig. 1. Stages of Proposed Method**

A. Data Collection

This study makes use of the "Heart Disease Prediction" dataset, which offers an extensive compilation of patient health characteristics from multiple sources. The dataset was cleaned and data preprocessed, reducing it from its original 13 variables to 7 important variables as shown in Table 1. The dataset provides relevant information for the evaluation and prognosis of heart disease by capturing a broad range of health parameters. Concerning various patient profiles, the gathered data provide a comprehensive picture of the health situations. This extensive dataset provides a strong basis for tasks involving prediction and comparison analysis concerning heart disease, allowing for an in-depth investigation of different risk variables and how they interact.

**Table 1. Variables Decision Tree**

| Name | Description |
|---|---|
| Age | Age in years |
| Sex | 0 = female 1 = male |
| Blood Pressure | Blood pressure (in mm Hg) |
| Cholesterol | Cholesterol in mg/dl |
| Fasting Blood Sugar | Fasting blood sugar>120 mg/dl: 1=True 0=False |
| Max Heart Rate | Maximum heart rate achieved |
| ST Depression | Numeric value measured in depression |

B. Data Preprocessing

An essential step in getting datasets ready for machine learning models is data preparation. Statistical imputation was used to handle missing variables at the start of the processing. In particular, mean values were employed for "Heart Rate", "Blood Pressure", and "Cholesterol". For the categories of "Age" and "ST Depression", median values were used. This tactic reduces data loss and preserves the robustness of the dataset. Label encoding was used to transform categorical columns, like "Gender", into numerical values to make sure the machine learning method would work with them. To improve data quality, duplicate records were also removed. The dataset is made clean, consistent, and ready for training and assessing machine learning models thanks to these preprocessing procedures.

C. Data Analysis

A dataset including 272 rows of data was used to build and train the predictive model in this study. This dataset was split into training and test subnets to do the analysis. The predictive model was constructed using 80% of the data as the training set used for training to recognize the patterns in the data set and the remaining 20% as the test set to confirm the model's accuracy. Using the scaled training set, the 'Decision Tree' model was trained to predict the target variable. 'Heart Disease' (0 for no disease, 1 for disease).

D. Model Training and Validation

Heart disease parameters are predicted using decision tree algorithms. The decision tree is chosen because of its resilience and capacity to handle intricate interactions in a dataset. The Decision Tree model's hyperparameters were adjusted to improve the model's performance even more. A decision tree reduces overfitting and increases accuracy by building several decision nodes and branches for each prediction. This strategy makes use of the Decision Tree algorithm's advantages to produce a reliable and effective model for predicting the characteristics of heart disease.

E. Evaluation

A thorough evaluation is conducted of the results of applying the Decision Tree algorithm to the relevant dataset. To determine how effectively the algorithm performs in the context of data modeling and classification, this evaluation involves a thorough study of the method's performance, including accuracy, precision, recall, and f1-score.

**Result:**

A confusion matrix is an important tool for evaluating the performance of classification models. By giving a summary of forecasts versus actual results, it is possible to identify the right predictions and errors. Also, a confusion matrix is a testing method that is used to assess the performance of a machine learning algorithm. This is done by comparing the predictions generated by the algorithm against the actual data that has been known beforehand. The confusion matrix provides a clear picture of the extent to which the algorithm can correctly predict each class or label in the dataset and helps in identifying several evaluation metrics such as accuracy, precision, recall, and f1-score that provide an in-depth understanding of the algorithm's performance in performing classification [14].

**Table 2. Confusion Matrix**

| | | True Class | |
|---|---|---|---|
| | | True | False |
| Prediction Class | True | TP | FP |
| | False | FN | TN |

Accuracy is a measure that defines how close the predicted value of a model is to the actual value of the observed data Eq. (1). In other words, it shows how well a model or algorithm can identify the true classes or values inside a dataset. This metric compares the forecast and determines how well the resultant prediction matches the facts in the data, indicating the model's predictive accuracy [15].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

A model for precisely identifying every true positive case in the dataset is called recall Eq. (2). Specifically, recall can be considered as the ratio between the number of relevant items identified by the model (true positive) and the total number of relevant items that are present in the dataset (true positive and false negative). The higher the recall value, the fewer relevant items are missed by the model [16].

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The percentage that indicates how many relevant items were chosen out of all the items chosen is known as precision. Precision is an indicator of how precise a system is in identifying relevant items out of the total items selected Eq. (3). This measure shows how well the system responds to requests for information by providing precise and pertinent responses [16].

$$Precision = \frac{TP}{TP + FP} \qquad (3)$$

F1-score is a metric that combines precision and recall in a single value Eq. (4). The harmonic mean of precision and recall is calculated to achieve this. The smaller of the two values is highlighted by the harmonic mean. A balanced representation of precision and recall can be obtained using the f1-score. The better the model in classification that balances precision and recall, the higher the f1-score value [17].

$$F1 - score = \frac{2(Precision + Recall)}{Precision + Recall} \qquad (4)$$

The developed decision tree model was trained and tested using a dataset containing various health parameters of individuals. The dataset included the following key variables: age, gender, cholesterol, blood pressure, heart rate, and depression ST. The model’s performance was evaluated using metrics such as accuracy, precision, recall, and f1-score with a confusion matrix as shown in Table 3.

**Table 3. Confusion Matrix Decision Tree**

| | True Yes | True No |
|---|---|---|
| Prediction Yes | 34 | 3 |
| Prediction No | 7 | 10 |

Based on the confusion matrix results in Table 3, it can be seen that as many as 34 data are classified as “Yes” with valid, then as many as 10 data are classified as “No” validly. After knowing the results of the classification of heart disease using the Decision Tree Algorithm, then can be known the value of accuracy, recall, precision, and f1-score.

The results of the application of the Decision Tree algorithm in the classification of heart disease obtained an accuracy of 81.48%, a recall of 82.93%, a precision of 91.89%, and an f1-score of 87.18%

**Table 4. Decision Tree Results**

| | |
|---|---|
| Accuracy | 0.8148 |
| Recall | 0.8293 |
| Precision | 0.9189 |
| F1-Score | 0.8718 |

## Summary

Based on a dataset of 272 entries with 7 relevant variables: age, sex, blood pressure, cholesterol, fasting blood sugar, maximal heart rate, and ST depression, we used a Decision Tree Algorithm in this work to predict heart disease. To handle missing values and transform categorical data into numerical representation, the dataset underwent preprocessing. Of the total data, 80% were utilized for training the Decision Tree model and the remaining 20% were used for testing. Our evaluation metrics included accuracy, precision, recall, and f1-score, with results showing an accuracy of 81.48%, a recall of 82.93%, a precision of 91.89%, and an f1-score of 87.18%. Although there are several limitations to the study, such as a very small dataset size that may impair generalizability, these results show how well the Decision Tree Algorithm predicts heart disease. To further improve prediction accuracy, future studies should think about enlarging the dataset and investigating different machine learning techniques to enhance generalizability, ultimately making it a more practical and scalable method for heart disease prediction. A deeper understanding and increased predictability of heart disease may be obtained by expanding the variables or applying this methodology to different datasets.

## References:

[1] World Heart Organization, “Cardiovascular diseases (CVDs).” Accessed: Aug. 04, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis,

and E. Atlam, “Advanced machine learning techniques for cardiovascular disease early detection and diagnosis,” *J Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00817-1.

[3] L. I. Santos *et al.*, “Decision tree and artificial immune systems for stroke prediction in imbalanced data,” *Expert Syst Appl*, vol. 191, Apr. 2022, doi: 10.1016/j.eswa.2021.116221.

[4] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective Heart Disease Prediction Using Machine Learning Techniques,” *Algorithms*, vol. 16, no. 2, Feb. 2023, doi: 10.3390/a16020088.

[5] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[6] Rishi Damarla, “Heart Disease Prediction,” https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction/data.

[7] D. K. Arnett *et al.*, “2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines,” Sep. 10, 2019, *NLM (Medline)*. doi: 10.1161/CIR.0000000000000678.

[8] T. G. Richardson *et al.*, “Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis,” *PLoS Med*, vol. 17, no. 3, Mar. 2020, doi: 10.1371/JOURNAL.PMED.1003062.

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

[10] N. Chandrasekhar and S. Peddakrishna, “Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization,” *Processes*, vol. 11, no. 4, Apr. 2023, doi: 10.3390/pr11041210.

[11] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[12] V. Chole, M. Thawakar, M. Choudhari, S. Chahande, S. Verma, and A. Pimpalkar, “Enhancing heart disease risk prediction with GdHO fused layered BiLSTM and HRV features: A dynamic approach,” *Biomed Signal Process Control*, vol. 95, Sep. 2024, doi: 10.1016/j.bspc.2024.106470.

[13] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” *Comput Biol Med*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.

[14] F. Handayani *et al.*, “JEPIN (Jurnal Edukasi dan Penelitian Informatika) Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung,” vol. 7, 2021.

[15] M. M. Ali *et al.*, “A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients,” *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100182.

[16] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, "Classification models combined with Boruta feature selection for heart disease prediction," *Inform Med Unlocked*, vol. 44, Jan. 2024,

doi: 10.1016/j.imu.2023.101442.

[17] D. Deepika and N. Balaji, "Effective heart disease prediction using novel MLP-EBMDA approach," *Biomed Signal Process Control*, vol. 72, Feb. 2022,

doi: 10.1016/j.bspc.2021.103318.