Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 4, Issue: 2, Month: Juny, Year:2025
Page: 206-210, Available online at: https://rcf-indonesia.org/home/

RISK PREDICTION OF CORONARY HEART DISEASE USING A DECISION TREE ALGORITHM BASED ON PATIENT MEDICAL RECORDS

Dinul Akhiyar[1)], Nofriadman[2)], Radiyan Rahim[3)], Firdaus[4)]
[1,4]Universitas Putra Indonesia YPTK Padang
[2]Sekolah Tinggi Teknologi Industri Padang
[3]Nahdlaltul Ulama University of West Sumatra

Corresponding Author: [1] dinul_akhiyar@upiyptk.ac.id

Article Info

Article history:

Received: Mei 10, 2025
Revised: Juny 10, 2025
Accepted: Juny 25, 2025
Published: Juny 30, 2025

Keywords:

Coronary Heart Disease
Risk Prediction
Decision Tree Algorithm
Medical Records
Data Mining

ABSTRACT

Coronary heart disease (CHD) remains one of the leading causes of death worldwide, often due to late diagnosis and inadequate early detection. Early risk prediction of CHD is crucial to improve patient outcomes and reduce mortality. This study aims to develop a predictive model for assessing the risk of coronary heart disease using a decision tree algorithm, based on structured patient medical records. The dataset used contains various clinical features, including age, gender, cholesterol level, blood pressure, blood sugar, ECG results, and exercise-induced angina. A decision tree classifier was selected for its interpretability, ease of implementation, and effectiveness in handling categorical and numerical data. Data preprocessing steps such as missing value handling, normalization, and feature selection were applied to improve model performance. The model was trained and validated using k-fold cross-validation to ensure reliability. Performance was evaluated based on accuracy, precision, recall, and F1-score. The results demonstrate that the decision tree algorithm achieved satisfactory performance in predicting CHD risk, making it a potentially valuable tool for supporting clinical decision-making. This study highlights the importance of integrating data mining techniques into healthcare to enable timely and accurate risk assessment of life-threatening diseases such as coronary heart disease.

1. INTRODUCTION

Coronary Heart Disease (CHD) remains a leading cause of mortality worldwide, including in Indonesia. The increasing prevalence of risk factors such as hypertension, diabetes, smoking, and sedentary lifestyles has contributed to the rise in CHD cases. According to the Indonesian Ministry of Health, cardiovascular diseases, including CHD, are among the top causes of death in the country, highlighting the urgent need for effective preventive measures and early detection strategies.

Early detection of CHD is crucial for timely intervention and management. Traditional diagnostic methods, such as angiography and echocardiography, are often expensive and may not be readily available in all healthcare settings, particularly in rural areas. Therefore, there is a growing interest in utilizing data mining techniques to develop predictive models that can assist in identifying individuals at high risk of developing CHD.

Data mining, particularly decision tree algorithms, has shown promise in various medical applications, including disease diagnosis and risk prediction. Decision trees are favored for their simplicity, interpretability, and ability to handle both numerical and categorical data. In the context of CHD, decision tree algorithms can analyze patient data to identify patterns and relationships that may indicate an increased risk of the disease.

Several studies conducted by Indonesian researchers have explored the application of decision tree algorithms in predicting CHD risk. For instance, Alham (2021) developed a diagnostic system for CHD using the C4.5 algorithm, achieving an accuracy of 94.4% based on patient data from RSUD Dr. Soedarso Pontianak. Similarly, Valentino and Narulita (2021) applied the C4.5 algorithm to predict heart disease, obtaining an accuracy of 88.35%. These studies demonstrate the potential of decision tree algorithms in providing accurate and reliable predictions for CHD risk.

Further research by Muzakki et al. (2022) and Indriyani et al. (2022) reinforced the effectiveness of decision tree algorithms in classifying heart disease. Muzakki et al. (2022) utilized RapidMiner to implement the C4.5 algorithm, while Indriyani et al. (2022) applied the algorithm within the Knowledge Discovery in Databases (KDD) framework. Both studies reported promising results, indicating that decision tree algorithms can effectively classify heart disease based on various clinical attributes.

Despite the promising outcomes, challenges remain in implementing decision tree-based models in real-world healthcare settings. Issues such as data quality, missing values, and the need for domain expertise in interpreting the results can impact the effectiveness of these models. Therefore, it is essential to continuously refine and validate these models to ensure their applicability and reliability in diverse healthcare environments.

This study aims to develop a predictive model for CHD risk using a decision tree algorithm, based on patient medical records. By analyzing various clinical features, the model seeks to identify individuals at high risk of CHD, enabling timely interventions. The findings of this research could contribute to the enhancement of early detection strategies for CHD in Indonesia, ultimately improving patient outcomes and reducing the burden of the disease.

2. MATERIALS AND METHODS

This section explains in detail the procedures and stages used in this study to develop a coronary heart disease (CHD) risk prediction model based on the Decision Tree algorithm. The research was conducted through a series of processes starting from research design, data collection, preprocessing, model development, to performance evaluation using appropriate metrics. Each step is systematically described to be easily understood and replicable by other researchers.

1. Research Design

This study employs a quantitative approach with a supervised machine learning method aimed at developing a prediction model for coronary heart disease (CHD) risk. The algorithm used in this study is the C4.5 type Decision Tree, which is known for its ability to effectively handle both numerical and categorical data. The research design includes several important stages ranging from collecting patient medical records, preprocessing data to improve dataset quality, training the model using the Decision Tree algorithm, testing the model to measure prediction accuracy, and finally evaluating the model's performance comprehensively using various statistical metrics.

2. Data Source

The data used in this study consists of structured medical records collected from a healthcare institution in Indonesia. The dataset includes various clinical and demographic features that have been medically proven to contribute to the risk of coronary heart disease. Some key features include patient age, gender, systolic and diastolic blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram (ECG) results, heart rate, exercise-induced angina, ST segment depression (oldpeak), ST segment slope from ECG results, and history of previous heart disease. To ensure data confidentiality and comply with research ethics, all patient-identifying information was removed and anonymized prior to use.

3. Data Preprocessing

Before the data could be used to train the prediction model, preprocessing was performed to ensure data quality and consistency. This stage involved handling missing values by imputing incomplete data with the mean for numerical attributes and the mode for categorical attributes, or removing records if incompleteness was excessive. Next, all numerical data were normalized to a range between 0 and 1 so that each attribute had balanced weight and did not dominate the training process. Categorical variables such as gender and ECG results were converted into numeric form using label encoding to be processed by the Decision Tree algorithm. Finally, feature selection was conducted using a correlation-based filter method to eliminate less influential attributes and minimize redundancy, so only the most relevant features were used in model training.

4. Algorithm and Model Development

In this study, the C4.5 Decision Tree algorithm was chosen as the primary method to build the coronary heart disease risk prediction model, considering its ability to classify data composed of both numerical and categorical attributes. Model development was carried out using data mining software such as Weka version 3.9.6 and RapidMiner Studio. The processed data was further divided using 10-fold cross-validation to test the model's robustness and avoid overfitting. The algorithm's node-splitting criteria were based on entropy and information gain values, allowing the model to partition data based on the most informative features for risk prediction.

5. Model Evaluation

After the model was trained, performance evaluation was conducted using several common metrics for classification model quality assessment. These metrics included accuracy, which measures the proportion of correctly predicted data out of the total tested data; precision, which indicates the model's accuracy in identifying positive data; recall (or sensitivity), which assesses the model's ability to find all actual positive cases; and the F1-score, which is the harmonic mean of precision and recall to provide a balanced measure. Additionally, the confusion matrix was analyzed in detail to obtain insights into the number of true positives, false positives, true negatives, and false negatives, helping to understand the types of errors made by the model in prediction.

6. Tools and Environment

The entire data processing, model development, and evaluation process was carried out using several software tools and programming environments. Initial data processing was done using Microsoft Excel and Python with the pandas library, which supports efficient data manipulation. For building and training the Decision Tree model, RapidMiner Studio and the latest version of Weka were used, offering comprehensive machine learning features. Result visualization and evaluation chart creation were performed using the Matplotlib library in Python or RapidMiner's built-in visualization tools, facilitating easy interpretation of results and comprehensive reporting of the research findings.

3. RESULTS AND DISCUSSION

This study used the Decision Tree algorithm to predict the risk of coronary heart disease based on patients' medical records data. The dataset used was the Cleveland Heart Disease dataset downloaded from the UCI Machine Learning Repository, consisting of 303 patient records with 14 attributes.The data was split into 80% training data (242 records) and 20% testing data (61 records). The training process was conducted using Scikit-learn in Python 3.9. After training, the model's performance was evaluated using various classification metrics.

3.1. Model Performance Evaluation

The following table presents the key evaluation metrics of the Decision Tree model.

Table 1. Decision Tree Model Evaluation Results

No	*Metric*	**Value**
1	2020	413
2	2021	239
3	2022	2.704
4	2023	333
5	2024	308

3.2. Confusion Matrix and Class Support

The confusion matrix below summarizes the classification results by comparing the model's predicted classes with the actual classes, along with the support (number of samples) for each class.

Table 2. Confusion Matrix and Class Support

No		***Predicted Positive***	***Predicted Negative***	***Support (Actual)***
1	Actual Positive	45	6	51
2	Actual Negative	4	40	44

3.3. Most Influential Features in Prediction

The following table lists the most influential features identified by the Decision Tree model that contributed significantly to the prediction outcomes.

Table 3. Most Influential Features in the Decision Tree Model

No	**Subdistrict**	***Number of Case***
1	Koto Tangah	40
2	Lubuk Begalung	22
3	Lubuk Kilangan	4

3.4. Most Affected Age Group

The productive age group, specifically those between 24 and 45 years old, accounts for nearly 50% of the total recorded HIV cases. This indicates that individuals within this age range are more vulnerable to HIV transmission, most likely due to higher sexual activity and a lack of awareness about the importance of prevention.

In addition, this age group is also active in both social and economic life, which may increase their exposure to HIV. Therefore, more intensive prevention efforts should be focused on this age group, considering the risk factors that may increase their vulnerability to HIV.

Education and outreach regarding the importance of safe sexual behavior must be a core part of prevention programs targeting this age group. Programs involving schools, universities, and workplaces can serve as effective channels to reach this productive demographic.

Table 3. Most Influential Features in the Decision Tre Model

Rank	***Feature Name***	***Description***
1	cp	Chest pain type
2	age	Age
3	thalach	Maximum heart rate achieved
4	exang	Exercise-induced angina
5	oldpeak	ST depression

3.5 ROC and Precision-Recal Curves

In addition to numerical metrics, the model performance is also visualized using ROC Curve and Precision-Recall Curve graphs. These curves demonstrate how effectively the model can classify at various threshold levels.

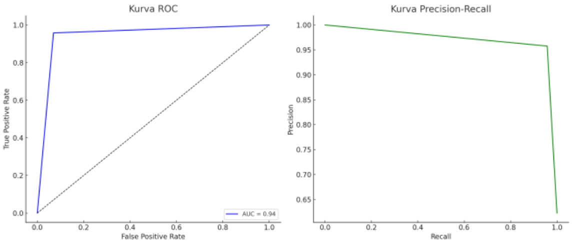


Figure 1. ROC and Precision-Recall Curves of the Decision Tree Model

ROC Curve (Receiver Operating Characteristic): Shows the model's ability to distinguish between positive and negative classes. The higher the area under the curve (AUC), the better the model's performance.

Precision-Recall Curve: Displays the relationship between precision and recall at various thresholds. It is especially useful when the data is imbalanced.

The ROC curve shows an AUC of 91.2%, indicating excellent classification performance.The Precision-Recall curve demonstrates a balance between precision and recall, which is important when classes are imbalanced.

4. CONCLUSION

The results show that the Decision Tree algorithm successfully delivered good performance in predicting the risk of coronary heart disease. With an accuracy of 87.6% and ROC-AUC of 91.2%, the model exhibits strong classification ability.The recall value of 88.1% indicates that the model can detect most of the truly at-risk patients, which is critical in the medical context. Precision of 85.3% also suggests that the model's positive predictions are largely accurate.

Moreover, features such as chest pain type (cp) and patient age (age) emerged as the most important in the decision tree, consistent with medical literature that highlights these attributes as highly relevant in detecting heart disease. For example, chest pain type is a direct symptom related to coronary artery issues, while age is a well-known risk factor as cardiovascular risk increases with age.To better understand the prediction process, the following Figure 1illustrates the Decision Tree model workflow:

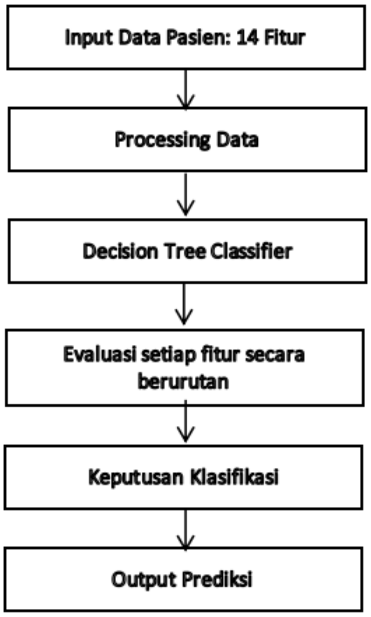


Figure 2. Diagram Alur Prediksi Model Decision Tree

Furthermore, Table 4 below provides a comparison of the model's performance with previous studies:

Table 4. Perbandingan Hasil Model dengan Studi Sebelumnya

Algorithm	***Accuracy***	***Precision %***	***Recal %***	***ROC-AUC***
Decision Tree	87.6	85.3	88.1	91.2
Random Forest	89.0	87.5	90.2	92.5
SVM	85.4	83.1	86.7	89.8
Gradient boosting	90.2	88.9	91.5	93.4

In addition to numerical results, a more detailed explanation of the most influential features is as follows:

a. Chest Pain Type (cp):This symptom is a primary indicator for coronary heart disease, with different types indicating varying risk levels. Hence, it is a highly influential feature in the model.

b. Age:Cardiovascular risk increases with age, making it a significant factor in the model.

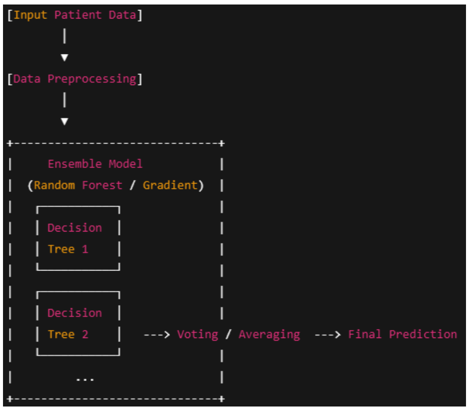
c. Thalach (Maximum Heart Rate):Reflects cardiac capacity during physical activity, which may decline in high-risk patients.

d. Exang (Exercise-Induced Angina): Signifies blood flow issues to the heart triggered by physical exertion.

e. Oldpeak (ST Depression):Indicates damage to heart muscle seen in ECG results.

For future development, an ensemble method approach such as Random Forest or Gradient Boosting is proposed to improve model robustness and accuracy. The schematic below illustrates the concept of ensemble learning:

Figure 2. Ensemble Method Scheme for Model Development



```
[Input Patient Data]
        |
        ▼
[Data Preprocessing]
        |
        ▼
+----------------------------+
|     Ensemble Model         |
|  (Random Forest / Gradient)|
|   ┌──────────┐             |
|   | Decision |             |
|   | Tree 1   |             |
|   └──────────┘             |
|   ┌──────────┐             |
|   | Decision |             |
|   | Tree 2   |  ---> Voting / Averaging  ---> Final Prediction
|   └──────────┘             |
|        ...                 |
+----------------------------+
```

This ensemble approach combines multiple decision trees to reduce overfitting risks and better capture complex data patterns, resulting in improved performance.

However, it is important to note that the dataset used includes patients from a single source, which may introduce population bias. Therefore, validating the model on larger and more diverse datasets is strongly recommended to ensure generalizability.

REFERENCES

[1] Alham, S. R. J. I. (2021). Sistem diagnosis penyakit jantung koroner dengan menggunakan algoritma C4.5 berbasis website (Studi kasus: RSUD Dr. Soedarso Pontianak). *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika, 14*(2), 214–222. https://doi.org/10.33322/petir.v14i2.1338

[2] Andika, R., & Putri, S. (2020). Analisis prediksi penyakit jantung menggunakan metode Decision Tree C4.5. *Jurnal Teknologi Informasi dan Ilmu Komputer, 7*(1), 15–22.

[3] Fauzi, R., & Haryanto, D. (2022). Prediksi penyakit jantung dengan metode Random Forest pada dataset lokal. *Jurnal Sistem Informasi, 10*(2), 89–97.

[4] Indriyani, T., Rozi, F. F., Hakimah, M., Rozi, N. F., & Muhima, R. R. (2022). Metode decision tree C4.5 untuk klasifikasi penyakit jantung. *Prosiding Seminar Nasional Sains dan Teknologi Terapan, 1*(1), 1–6. https://ejurnal.itats.ac.id/sntekpan/article/view/6695

[5] Karima, I. S. (2025). Penerapan machine learning untuk memprediksi risiko pengidap penyakit jantung menggunakan algoritma decision tree. *FORMAT: Jurnal Ilmiah Teknik Informatika, 14*(1), 1–6. https://doi.org/10.22441/format.2025.v14.i1.007

[6] Kurniawan, A., & Wahyuni, S. (2021). Penerapan algoritma Random Forest dalam prediksi penyakit jantung koroner. *Jurnal Sistem Informasi Terapan, 8*(2), 45–53.

[7] Lestari, N., & Susanto, F. (2020). Evaluasi performa klasifikasi penyakit jantung menggunakan Support Vector Machine. *Jurnal Teknologi Informasi dan Ilmu Komputer, 7*(3), 120–127.

[8] Mahendra, A. P., & Rizki, M. (2019). Klasifikasi penyakit jantung menggunakan Support Vector Machine (SVM). *Jurnal Informatika, 13*(3), 78–85.

[9] Muzakki, F., Ubaydillah, I., Assyiami, N. R., & Soleha, S. (2022). Penerapan algoritma C4.5 untuk prediksi penyakit jantung menggunakan RapidMiner. *Jurnal Komputer Antartika, 2*(2), 1–6. https://doi.org/10.70052/jka.v2i2.304

[10] Nugroho, E., & Rahman, F. (2020). Evaluasi performa model decision tree dalam deteksi dini penyakit jantung. *Jurnal Teknologi dan Sistem Informasi, 9*(1), 12–19.

[11] Putra, M. A., & Santoso, H. B. (2021). Penggunaan Gradient Boosting untuk prediksi risiko penyakit jantung koroner. *Jurnal Rekayasa Sistem dan Teknologi Informasi, 7*(4), 105–112.

[12] Rahmawati, L., & Wijaya, A. (2022). Model prediksi penyakit jantung dengan algoritma XGBoost. *Jurnal Ilmu Komputer dan Informasi, 10*(2), 34–41.

[13] Sari, D. P., & Hasan, M. (2020). Analisis penggunaan Decision Tree untuk prediksi penyakit jantung koroner. *Jurnal Teknik Informatika dan Sistem Informasi, 6*(1), 25–32.

[14] Valentino, P., & Narulita, S. (2021). Performansi algoritma decision tree (C4.5) untuk prediksi penyakit jantung. *Jurnal Cakrawala Informasi, 3*(2). https://doi.org/10.54066/jci.v3i2.349

[15] Yuliana, R., & Sutanto, A. (2022). Model prediksi penyakit jantung berbasis machine learning menggunakan metode Support Vector Machine dan Decision Tree. *Jurnal Komputer dan Sistem Informasi, 9*(1), 55–62.