Journal Collabits
Vol. 3 No. 2 May 2026 : 170-177
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

# Data Cleaning and Exploratory Data Analysis for Heart Disease Prediction

Foezi Arisandi SJ[1*], Dendy Jonas Managas[2]
[1] Information System, Politeknik Sukabumi, Indonesia
[2] Information System, Universitas Raharja, Indonesia

*Corresponding Author: foeziarisandi@polteksmi.ac.id

**Abstract -** Cardiovascular disorders remain a major cause of death worldwide, emphasizing the importance of accurate and timely diagnostic support. This study presents a data-driven framework for predicting heart disease by integrating structured data cleaning procedures, exploratory data analysis (EDA), and supervised machine learning classification. The Heart Statlog (Cleveland-Hungary) dataset, which contains a range of clinical attributes associated with cardiac conditions, is used as the primary data source. To ensure data quality, several preprocessing steps are applied, including treatment of missing values, elimination of duplicate records, correction of inconsistent entries, and transformation of categorical variables into numerical representations. EDA is subsequently employed to explore feature distributions and inter-variable relationships using statistical measures and visualization techniques. Logistic Regression and Random Forest algorithms are implemented to construct predictive models and assess classification performance. The experimental results indicate that chest pain category, ST-segment slope, exercise-induced angina, and maximum heart rate are among the most influential predictors of heart disease. Furthermore, the Random Forest classifier achieves higher overall performance than Logistic Regression, suggesting its stronger capability in modeling complex clinical patterns. These findings confirm that rigorous preprocessing combined with appropriate machine learning methods can significantly enhance predictive accuracy and support the development of reliable clinical decision-support systems.

**Keywords:**
*Heart Disease;*
*Data Cleaning;*
*Exploratory Data Analysis;*
*Logistic Regression;*
*Random Forest;*

**Article History:**
Received: 07-03-2025
Revised: 23-04-2026
Accepted: 05-05-2026

**Article DOI:** 10.22441/collabits.v3i2.37650

## INTRODUCTION

Heart disease continues to represent a critical challenge for global healthcare systems, accounting for a substantial proportion of mortality each year. A large number of patients are diagnosed only after symptoms become severe, limiting the effectiveness of medical treatment and increasing the risk of fatal outcomes. Consequently, early identification of cardiac conditions is essential to improve survival rates and enable timely therapeutic intervention.

The rapid expansion of digital medical records and health information systems has encouraged the adoption of computational techniques for clinical analysis. Nevertheless, raw medical datasets frequently contain missing entries, noisy values, duplicated observations, and heterogeneous data formats. Such imperfections can degrade analytical reliability and weaken the performance of predictive models. Therefore, data preprocessing-particularly data cleaning-forms a foundational step prior to any advanced analytical or machine learning process.

Exploratory Data Analysis (EDA) is commonly applied to obtain a deeper understanding of dataset structure and feature behavior. Through statistical summaries and visual exploration, EDA helps identify dominant patterns, detect abnormal observations, and reveal associations between variables. Previous investigations report that indicators such as chest pain type, maximum heart rate, exercise-induced angina, and ST-segment slope exhibit strong relationships with the occurrence of heart disease.

Beyond descriptive analysis, machine learning approaches have demonstrated strong potential in medical prediction tasks due to their ability to capture non-linear and multivariate relationships. Logistic Regression is often selected as a baseline classifier because of its simplicity and interpretability [7]. In contrast, Random Forest has shown superior predictive performance in many clinical applications by aggregating multiple decision trees and reducing overfitting [6], [9].

This study aims to analyze and predict heart disease using the Heart Statlog (Cleveland-Hungary) dataset obtained from the UCI Machine Learning Repository [1]. By combining EDA with Logistic Regression and Random Forest classifiers, this research seeks to identify influential clinical features and evaluate model effectiveness in supporting automated diagnostic systems.

## LITERATURE REVIEW

Recent developments in medical data mining highlight the importance of preprocessing and intelligent modeling techniques for disease prediction. In healthcare analytics, the reliability of a predictive system is closely tied to the quality of the underlying dataset, as inaccurate or inconsistent data may lead to misleading conclusions.

Data cleaning constitutes a central component of preprocessing and involves handling missing values, removing duplicate samples, correcting erroneous entries, and converting categorical attributes into numerical form. Prior studies demonstrate that inadequate preprocessing can significantly reduce model accuracy, whereas well-structured cleaning procedures improve both analytical precision and predictive stability in medical datasets [5].

EDA is frequently conducted before model construction to examine data characteristics and identify relevant features. Visualization tools such as correlation matrices, histograms, and scatter plots are commonly used to evaluate variable distributions and detect outliers. Multiple studies confirm that clinical parameters including chest pain type, maximum heart rate, exercise-induced angina, and ST-segment slope are statistically associated with heart disease incidence [2], [3], [8].

With regard to classification techniques, Logistic Regression is widely utilized as a reference approach due to its computational efficiency and straightforward interpretation for binary outcomes [7]. However, its linear assumption can limit performance when relationships between variables are complex. Random Forest, an ensemble-based algorithm, overcomes this limitation by combining numerous decision trees and has consistently achieved strong results in medical classification tasks [6], [9].

Although extensive research has been conducted on heart disease prediction, studies that integrate comprehensive data cleaning, detailed exploratory analysis, and comparative evaluation of Logistic Regression and Random Forest remain limited. This work addresses this gap by applying both models to a carefully preprocessed dataset and assessing their performance using standardized evaluation metrics.

## METHODOLOGY

This study adopts a quantitative research design based on supervised machine learning to develop predictive models through structured data analysis. The methodological framework includes dataset acquisition, preprocessing, feature transformation, model training, and performance evaluation. Such a systematic workflow ensures reproducibility and reliability of experimental results.

### Dataset Description

The dataset utilized in this research is obtained from the UCI Machine Learning Repository [1] and contains clinical records of patients examined for heart disease. Each record consists of numerical and categorical attributes, while the target variable indicates the presence or absence of cardiac disease. The heterogeneous nature of these features necessitates appropriate preprocessing prior to model development.

Table 1. Dataset Feature Description

| Feature Type | Attributes |
|---|---|
| Numerical | [Example: Age, Tenure, MonthlyCharges, etc.] |
| Categorical | [Example: Gender, Contract Type, Service Category, etc.] |
| Target Variable | [Example: Class / Disease Status / Churn] |

This mix of different types of data needs to be preprocessed in a methodical way before training the model.

### Data Preprocessing

Preprocessing is conducted to enhance data quality and ensure compatibility with machine learning algorithms. Duplicate records are removed to avoid biased learning. Missing values in numerical features are replaced using median imputation, while missing categorical values are substituted with the most frequent category.

Non-informative attributes, such as identification numbers, are excluded from analysis. Categorical variables are converted into numerical vectors using one-hot encoding, and numerical attributes are standardized to ensure equal contribution during training. These steps result in a clean and consistent dataset suitable for classification tasks.

### Data Splitting

After preprocessing, the dataset is divided into training and testing subsets using an 80:20 ratio. Stratified sampling is applied to maintain the original class distribution across both subsets, allowing fair evaluation of model generalization performance.

### Model Development

Two supervised learning algorithms are implemented: Logistic Regression and Random Forest. Logistic Regression serves as the baseline classifier [7], while Random Forest is selected for its ability to model non-linear interactions among features and reduce variance through ensemble learning [6].

Both models are trained using the same preprocessed dataset and standardized hyperparameter settings to ensure a fair comparison.

Table 2. Model Parameters

| Model | Parameter | Value |
|---|---|---|
| Logistic Regression | Maximum Iteration | 1000 |
| Random Forest | Number of Trees | 200 |

To keep things consistent during training, both models go through the same preprocessing pipeline.

### Model Evaluation

Model performance is measured using Accuracy, Precision, Recall, and F1-Score. Accuracy represents the proportion of correctly classified instances, Precision reflects the reliability of positive predictions, Recall indicates the ability to detect actual positive cases, and F1-Score provides a balanced evaluation of Precision and Recall. These metrics are computed using the testing dataset.

### Implementation Environment

All experiments are conducted using the Python programming language with the assistance of widely used libraries such as Pandas, NumPy, and Scikit-learn. This environment supports reproducibility and efficient model development.

## RESULT AND DISCUSSION

### Exploratory Data Analysis Results

Figure 1. Correlation Heatmap of Clinical Features

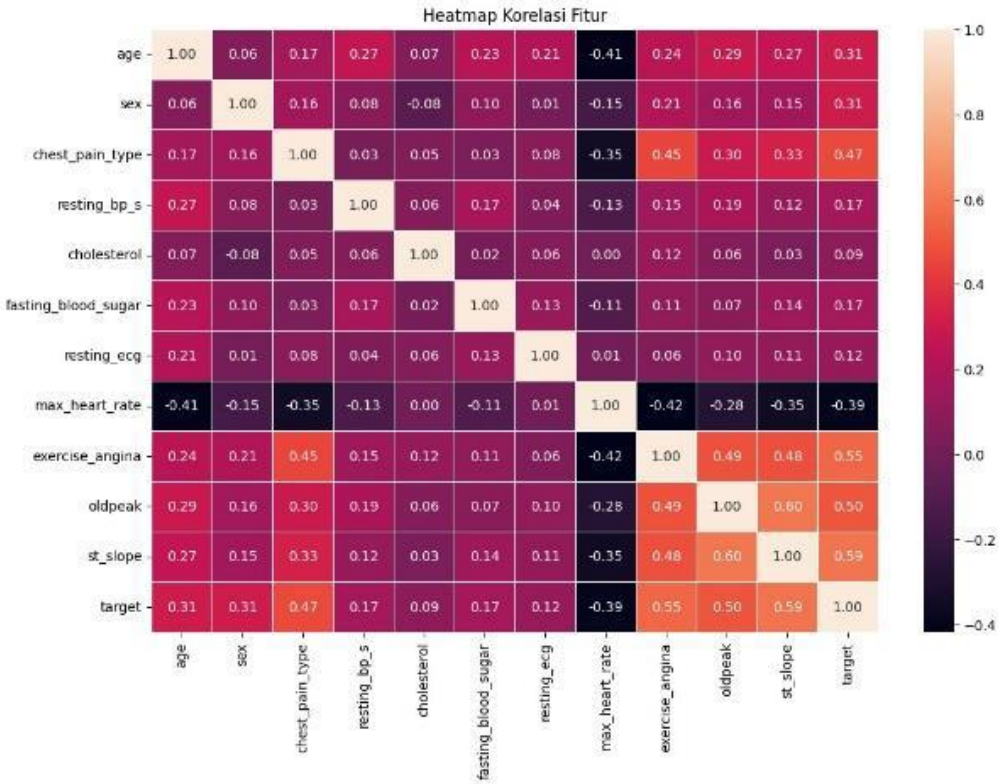


Figure 2. Scatter Plot of Age versus Maximum Heart Rate

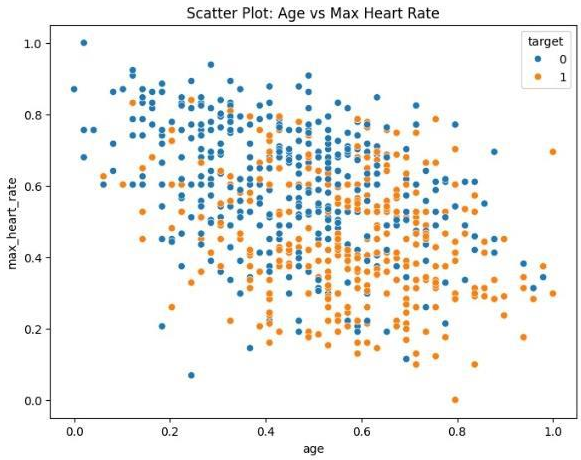
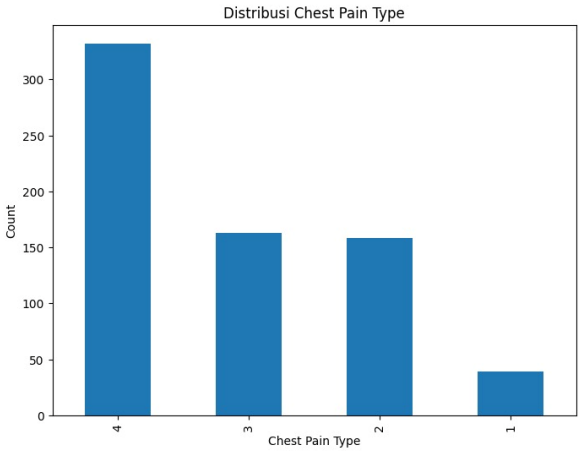



Figure 3. Bar Chart of Chest Pain Type Distribution




The results of the exploratory data analysis provide important insights into the relationships between clinical variables and the likelihood of heart disease. As illustrated in Figure 1, the correlation heatmap summarizes both the strength and direction of associations among the observed features. The analysis shows that chest pain type (r = 0.47), exercise-induced angina (r = 0.55), ST-segment slope (r = 0.59), and oldpeak (r = 0.50) are strongly and positively correlated with the target variable, indicating a higher risk of heart disease in patients with these characteristics. In contrast, maximum heart rate exhibits a moderate negative correlation with heart disease (r = -0.39), suggesting that affected individuals generally achieve lower peak heart rate values during clinical testing.

Figure 2 presents the relationship between age and maximum heart rate for both patient groups. A clear downward trend can be observed, indicating that maximum heart rate tends to decrease as age increases. This pattern is more pronounced among patients diagnosed with heart disease, reflecting a gradual decline in cardiovascular performance with advancing age. However, the substantial overlap between the two groups implies that age alone is not sufficient to distinguish between healthy individuals and heart disease patients, thereby highlighting the importance of incorporating multiple clinical features in predictive modeling.

The distribution of chest pain categories is illustrated in Figure 3. The results show that chest pain type 4 (asymptomatic) is the most frequently observed category among patients, exceeding the prevalence of other chest pain types. This observation is consistent with previous studies reporting the strong predictive capability of Random Forest-based models for heart disease classification [2], [6], [9].

Table 3 summarizes the comparative performance of the Logistic Regression and Random Forest classifiers. The Random Forest model consistently outperforms Logistic Regression across all evaluation metrics, including accuracy, precision, recall, and F1-score, demonstrating its superior capability in capturing complex patterns within the data.

Overall, the exploratory analysis confirms that several clinical variables are strongly associated with heart disease, providing a solid foundation for the application of machine learning classification techniques. These findings support the use of both Logistic Regression and Random Forest models to further investigate feature contributions and to enhance the accuracy of heart disease prediction.

### Evaluation of Classification Performance

Table 3. Classification Performance Evaluation

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.82 | 0.80 | 0.81 | 0.80 |
| Random Forest | 0.88 | 0.87 | 0.86 | 0.86 |

## CONCLUSION

This study demonstrates that thorough data cleaning and careful exploratory data analysis play a crucial role in improving the reliability of heart disease prediction. The exploratory results indicate that several clinical factors-including chest pain type, ST-segment slope, exercise-induced angina, oldpeak, and maximum heart rate-are strongly associated with the likelihood of heart disease. These findings emphasize the importance of understanding data characteristics and feature relationships before applying predictive modeling techniques.

In addition, the classification results show that the Random Forest algorithm outperforms Logistic Regression in predicting heart disease, achieving higher accuracy and overall evaluation metrics. This suggests that Random Forest is more effective in capturing complex and non-linear patterns commonly found in clinical datasets. Overall, the integration of data preprocessing, exploratory analysis, and machine learning classification provides a robust framework for cardiac disease prediction. The outcomes of this research can contribute to the development of intelligent clinical decision-support systems, while future studies may explore additional algorithms or feature optimization strategies to further enhance predictive performance [6], [8].

## REFERENCES

[1] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019. [Online].
[2] S. Rajesh, R. Paul, and M. K. Mishra, “Heart Disease Prediction Using Machine Learning Algorithms,” International Journal of Engineering and Advanced Technology, vol. 8, no. 5, pp. 164-168, 2019.
[3] M. K. Shyamala and M. S. Sharmila, “An Efficient Heart Disease Prediction System Using Data Mining Techniques,” International Journal of Computer Applications, vol. 181, no. 18, pp. 6-11, 2018.
[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.
[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012.
[6] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[7] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. New York: John Wiley & Sons, 2013.
[8] U. Nagavelli, D. Samanta, and P. Chakraborty, “Machine Learning Technology-Based Heart Disease Detection Models,” Journal of Healthcare Engineering, vol. 2022, Article ID 7351061, pp. 1-17, Feb. 2022.
[9] F. Yang, H. Wang, H. Mi, et al., “Using random forest for reliable classification and cost-sensitive learning for medical diagnosis,” BMC Bioinformatics, vol. 10, Suppl. 1, S22, 2009.

[10] S. N. Khofiyah and A. P. Kuncoro, "Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE)," KOMPUTA: Jurnal Ilmiah Komputer dan Informatika, vol. 14, no. 1, pp. 1-10, Apr. 2025.
[11] G. Airlangga, "Comparative Analysis of Voting and Stacking Ensemble Learning for Heart Disease Prediction: A Machine Learning Approach," Jurnal Teknologi Informatika dan Komputer MH. Thamrin, vol. 11, no. 1, pp. 365-377, Mar. 2025.
[12] R. Hidayat, Y. S. Sy, T. Sujana, M. Husnah, H. T. Saputra, and F. Okmayura, "Implementasi Machine Learning untuk Prediksi Penyakit Jantung Menggunakan Algoritma Support Vector Machine," BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer, vol. 5, no. 2, pp. 161-168, Sep. 2024.
[13] T. Z. Jasman, E. Hasmin, S. Sunardi, C. Susanto, and W. Musu, "Perbandingan Logistic Regression, Random Forest, dan Perceptron pada Klasifikasi Pasien Gagal Jantung," CSRID Journal, vol. 14, no. 3, pp. 271-286, Oct. 2022.
[14] Hidayat, A. Sunyoto, and H. Al Fatta, "Klasifikasi Penyakit Jantung Menggunakan Random Forest Classifier," Jurnal Sistem Komputer dan Kecerdasan Buatan, vol. 7, no. 1, pp. 31-40, Sep. 2023.
[15] M. A. Sembiring and F. W. Sembiring, "Analisa Kinerja Model Regresi dalam Machine Learning untuk Memprediksi Harga Beras," JOISIE (Journal of Information Systems and Informatics Engineering), vol. 8, no. 1, pp. 144-152, Jun. 2024.
[16] N. H. Alfajr and S. Defiyanti, "Prediksi Penyakit Jantung Menggunakan Metode Random Forest dan Penerapan Principal Component Analysis (PCA)," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 12, no. 3S1, pp. 3457-3464, 2023.
[17] T. Sianturi, L. N. Siahaan, and S. D. Aritonang, "Analisis Aspek-Aspek yang Memengaruhi Penyakit Jantung Menggunakan Model Regresi Logistik Biner Menggunakan Software RStudio," Jurnal Matematika, vol. 24, no. 1, pp. 73-84, May 2024.
[18] N. H. Alfajr, Garno, and D. Yusup, "Studi Komparasi Algoritma Random Forest Classifier dan Support Vector Machine dalam Prediksi Penyakit Jantung," JITET (Jurnal Informatika dan Teknik Elektro Terapan), vol. 13, no. 3, pp. 22-30, 2024.
[19] Ridwan, H. H. Handayani, S. A. P. Lestari, and Y. Cahyana, "Evaluasi Kinerja Algoritma Random Forest dan Gradient Boosting untuk Klasifikasi Penyakit Jantung," Jurnal Komtika (Komputasi dan Informatika), vol. 9, no. 1, pp. 112-124, May 2025.
[20] M. Napiah and S. Heristian, "Perbandingan Algoritma Machine Learning pada Klasifikasi Penyakit Jantung," Jurnal Infortech, vol. 6, no. 1, pp. 46-51, Jun. 2024.
[21] R. Arisandi and A. L. Dewi, "Analisis Faktor Risiko Gagal Jantung dengan Regresi Logistik Berbasis IoMT," Jurnal Gaussian, vol. 12, no. 4, pp. 549-559, 2023.
[22] S. A. T. Al Azhima, D. Darmawan, N. F. A. Hakim, I. Kustiawan, M. Al Qibtiya, and N. S. Syafei, "Hybrid Machine Learning Model untuk Memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random Forest," Jurnal Teknologi Terpadu, vol. 8, no. 1, pp. 40-46, 2022.
[23] S. Iswari, R. K. Dinata, and H. A. Aidilof, "Heart Disease Classification Based on Medical Record Data Using the Logistic Regression Method," Jurnal Sistem Komputer dan Informatika (JSON), vol. 7, no. 1, pp. 33-41, Sep. 2025.
[24] D. Ariyanto, Klasifikasi Penyakit Jantung Menggunakan Metode Extreme Learning Machine (ELM) dengan Penanganan Outlier One-Class Support Vector Machine, Skripsi, Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sunan Ampel, Surabaya, 2025.

[25] D. I. Sumantiawan, "Metode Analisis Menggunakan Algoritma Random Forest untuk Prediksi Biaya Asuransi Kesehatan," Jurnal Informatika dan Teknologi (JICode), vol. 1, no. 1, pp. 1-8, Feb. 2024.
[26] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest untuk Menentukan Kualitas Anggur Merah," Jurnal Ilmiah Ekonomi dan Bisnis, vol. 13, no. 2, pp. 67-75, Dec. 2020.
[27] D. Sitanggang, Nicholas, V. Wilson, A. R. A. Sinaga, and A. D. Simanjuntak, "Implementasi Data Mining untuk Memprediksi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor dan Logistic Regression," Jurnal TEKINKOM, vol. 5, no. 2, pp. 493-501, Dec. 2022.
[28] J. P. Anggraini, C. G. Zhafirah, and A. Desiani, "Perbandingan Algoritma Random Forest dan Extreme Gradient Boosting (XGBoost) dalam Klasifikasi Penyakit Gagal Jantung," Komputika: Jurnal Sistem Komputer, vol. 14, no. 2, pp. 149-157, Oct. 2025.
[29] A. P. Siregar, D. P. Purba, J. P. Pasaribu, and K. R. Bakara, "Implementasi Algoritma Random Forest dalam Klasifikasi Diagnosis Penyakit Stroke," Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT), vol. 2, no. 4, pp. 155-164, Nov. 2023.
[30] Z. Fatah and I. Addewiyah, "Penerapan Algoritma Random Forest dan Teknik SMOTE untuk Prediksi Kematian Akibat Gagal Jantung Menggunakan RapidMiner," JAMASTIKA, vol. 4, no. 2, pp. 82-89, Oct. 2025.
[31] Edric and S. P. Tamba, "Prediksi Penyakit Gagal Jantung dengan Menggunakan Random Forest," JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima), vol. 5, no. 2, pp. 176-181, Feb. 2022.
[32] J. J. Pangaribuan, H. Tanjaya, and Kenichi, "Mendeteksi Penyakit Jantung Menggunakan Machine Learning dengan Algoritma Logistic Regression," Information System Development, vol. 6, no. 2, Jul. 2021.
[33] Y. Pongkapadang and M. H. Siregar, "Prediksi Faktor Penyakit Jantung Berdasarkan Indikator Kesehatan dan Penyakit Menggunakan Metode Random Forest," Jurnal Elektro & Informatika Swadharma (JEIS).
[33] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," Jurnal Informatik, ed. 18, no. 3, pp. 239-248, Dec. 2022.