JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

# Komparasi Model Logistic Regression dan Random Forest pada Prediksi Penyakit Jantung Berdasarkan Analisis Fitur Klinis

**Arsenly Realino[1*], Wahyu Nugraha[2], Rabiatus Sa'adah[3]**

[1,2,3]Fakultas Teknik dan Informatika, Program Studi Informatika, Universitas Bina Sarana Informatika, Pontianak, Indonesia
Email: [1*]renolanino07@gmail.com, [2]wahyu.whn@bsi.ac.id, [3]rabiatus.rbh@bsi.ac.id
(*Email Corresponding Author: renolanino07@gmail.com)
*Received: 1 Juli 2026 | Revision: 8 Juli 2026 | Accepted: 14 Juli 2026*

**Abstrak**

Penyakit jantung merupakan salah satu masalah kesehatan serius yang membutuhkan deteksi dini agar risiko keterlambatan penanganan dapat diminimalkan. Pemanfaatan Machine Learning memberikan peluang untuk membangun model prediksi berbasis fitur klinis pasien, namun pemilihan algoritma yang tepat masih menjadi isu penting karena performa model sangat dipengaruhi oleh karakteristik dataset. Penelitian ini bertujuan untuk membandingkan performa Logistic Regression dan Random Forest dalam memprediksi penyakit jantung serta menganalisis fitur klinis yang paling berkontribusi terhadap hasil klasifikasi. Dataset yang digunakan adalah Heart Disease Prediction Dataset berisi 270 data pasien, 13 fitur prediktor, dan 1 atribut target. Tahapan penelitian meliputi Exploratory Data Analysis, preprocessing, label encoding, pembagian data 80:20, standardisasi data pada Logistic Regression, pemodelan klasifikasi, optimasi Random Forest menggunakan GridSearchCV dengan 5-Fold Cross Validation, serta evaluasi menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil penelitian menunjukkan bahwa Logistic Regression memperoleh performa terbaik dengan accuracy 85,19%, precision 78,57%, recall 91,67%, F1-score 84,62%, dan ROC-AUC 89,86%. Random Forest memperoleh accuracy 81,48% dan meningkat menjadi 83,33% setelah tuning, tetapi belum mampu melampaui Logistic Regression. Fitur terpenting yang ditemukan adalah Chest Pain Type, Max HR, dan ST Depression. Dengan demikian, Logistic Regression dinilai lebih sesuai sebagai model pendukung deteksi dini penyakit jantung pada dataset klinis berskala terbatas.

**Kata Kunci**: Logistic Regression, Random Forest, Machine Learning, Penyakit Jantung, Feature Importance.

***Abstract***

*Heart disease is a serious health problem that requires early detection to reduce the risk of delayed treatment. Machine Learning enables the development of prediction models based on clinical patient features; however, selecting the most appropriate algorithm remains challenging because model performance is strongly affected by dataset characteristics. This study aims to compare the performance of Logistic Regression and Random Forest in predicting heart disease and to analyze the most influential clinical features in the classification process. The dataset used in this study was the Heart Disease Prediction Dataset consisting of 270 patient records, 13 predictor features, and 1 target attribute. The research stages included Exploratory Data Analysis, preprocessing, label encoding, an 80:20 data split, data standardization for Logistic Regression, classification modeling, Random Forest optimization using GridSearchCV with 5-Fold Cross Validation, and evaluation based on accuracy, precision, recall, F1-score, and ROC-AUC. The results show that Logistic Regression achieved the best performance with 85.19% accuracy, 78.57% precision, 91.67% recall, 84.62% F1-score, and 89.86% ROC-AUC. Random Forest achieved 81.48% accuracy and improved to 83.33% after tuning, yet it did not outperform Logistic Regression. The most important features were Chest Pain Type, Max HR, and ST Depression. Therefore, Logistic Regression is considered more suitable as an early heart disease detection support model for limited-scale clinical datasets.*

*Keywords: Logistic Regression, Random Forest, Machine Learning, Heart Disease, Feature Importance.*

## 1. PENDAHULUAN

Transformasi teknologi informasi pada bidang kesehatan telah mendorong pemanfaatan kecerdasan buatan untuk membantu proses deteksi dini penyakit. Salah satu persoalan medis yang membutuhkan dukungan analisis prediktif adalah penyakit jantung, karena penyakit ini berhubungan dengan risiko mortalitas tinggi dan memerlukan penanganan cepat. Pada praktik konvensional, diagnosis penyakit jantung dilakukan melalui pemeriksaan klinis, interpretasi gejala, dan evaluasi rekam medis oleh tenaga medis. Proses tersebut tetap menjadi dasar utama diagnosis, namun pengolahan data klinis dengan pendekatan Machine Learning dapat menjadi instrumen pendukung untuk mengenali pola risiko pasien secara lebih sistematis [1], [2].

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

Penyakit kardiovaskular (CVD) tetap menjadi penyebab utama kematianglobal, dengan sekitar 17,9 juta kematian setiap tahun atau 32% daritotal kematian di dunia [3]. Di Amerika Serikat saja, penyakit jantung menyebabkan sekitar 252,2 miliar dolar dalam biaya langsung dan tidaklangsung pada periode 2019-2020 [3]. Beban global CVD diperkirakan akanmeningkat sebesar 90% dari tahun 2025 hingga 2050, meningkatkan jumlah kematian dari 20,5 juta pada tahun 2025 menjadi 35,6 juta pada tahun 2050 [3]. Angka-angka ini menyoroti kebutuhan mendesak akan alat diagnostik yang lebih efektif dan efisien.

Machine Learning mampu mempelajari pola dari data historis dan menghasilkan prediksi terhadap data baru berdasarkan atribut yang tersedia. Pada kasus prediksi penyakit jantung, atribut klinis seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, jenis nyeri dada, detak jantung maksimal, depresi ST, dan hasil pemeriksaan thallium dapat digunakan sebagai fitur prediktor. Dengan pendekatan klasifikasi, sistem dapat mengelompokkan pasien ke dalam kelas Presence atau Absence berdasarkan kecenderungan klinis yang dipelajari dari dataset [4], [5].

Berbagai algoritma dapat digunakan dalam pemodelan prediktif, mulai dari algoritma linier yang sederhana hingga algoritma ensemble yang lebih kompleks. Logistic Regression merupakan algoritma klasifikasi yang banyak digunakan karena memiliki struktur model yang relatif sederhana, efisien, dan mudah diinterpretasikan. Sementara itu, Random Forest merupakan algoritma ensemble learning yang membangun banyak decision tree dan menentukan hasil klasifikasi melalui mekanisme majority voting. Secara teoretis, Random Forest mampu menangani hubungan non-linier dan pola data kompleks, tetapi kebutuhan komputasi dan risiko overfitting tetap perlu diperhatikan, terutama ketika ukuran dataset terbatas [6], [7].

Kajian terdahulu menunjukkan bahwa Logistic Regression dan Random Forest sama-sama sering digunakan dalam prediksi penyakit jantung, namun hasil performanya tidak selalu konsisten. Beberapa penelitian menunjukkan algoritma ensemble menghasilkan performa tinggi, sedangkan penelitian lain menegaskan bahwa model sederhana tetap dapat bersaing apabila karakteristik data relatif kecil, bersih, dan tidak terlalu kompleks. Perbedaan hasil tersebut dipengaruhi oleh ukuran dataset, teknik preprocessing, pembagian data, metode validasi, dan konfigurasi parameter yang digunakan [8], [9].

Kesenjangan penelitian yang menjadi dasar artikel ini adalah perlunya komparasi yang lebih terfokus antara Logistic Regression dan Random Forest pada dataset yang sama, tahapan preprocessing yang seragam, serta metrik evaluasi yang identik. Selain membandingkan performa model, penelitian ini juga menambahkan analisis feature importance untuk mengidentifikasi atribut klinis yang paling berperan dalam klasifikasi penyakit jantung. Dengan demikian, hasil penelitian tidak hanya menunjukkan model terbaik berdasarkan angka evaluasi, tetapi juga memberikan gambaran fitur klinis dominan yang dapat menjadi perhatian dalam pengembangan sistem pendukung keputusan medis.[[10][11]

Penelitian terbaru juga menunjukkan bahwa optimasi hyperparameter menggunakan GridSearchCV dapat meningkatkan performa model secara signifikan. Studi oleh Faiqurahman et al. (2025) melaporkan bahwa tuning hyperparameter pada Random Forest berhasil meningkatkan akurasi dari 98,5% menjadi 100% [12]. Penelitian lain oleh Muhamad dan Matin (2023) juga menunjukkan efektivitas GridSearchCV dalam meningkatkan performa Random Forest [13]. Namun, efektivitas tuning sangat bergantung pada karakteristik dataset dan konfigurasi parameter yang diuji.

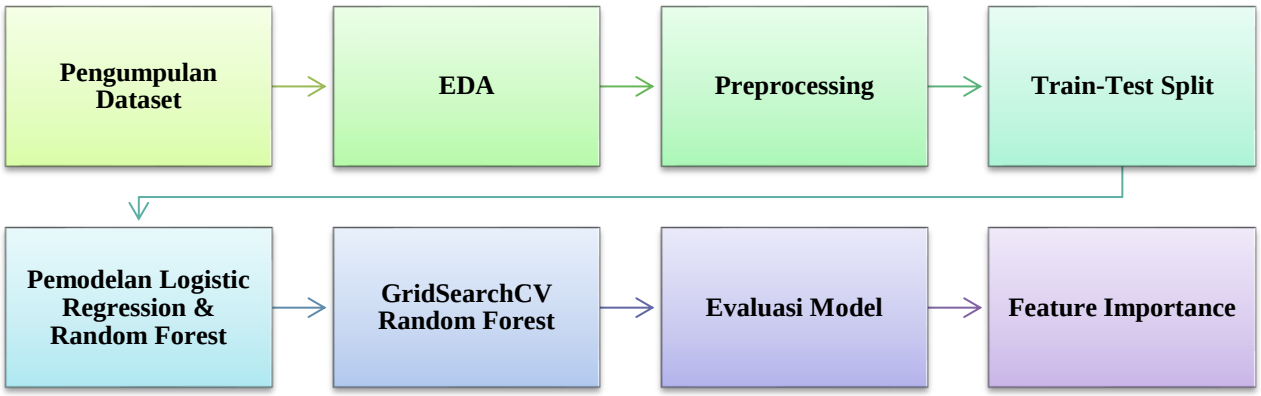
Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis dan mengkomparasi performa Logistic Regression dan Random Forest dalam memprediksi penyakit jantung menggunakan Heart Disease Prediction Dataset. Penelitian juga mengevaluasi pengaruh hyperparameter tuning pada Random Forest menggunakan GridSearchCV dan mengidentifikasi fitur klinis yang memiliki kontribusi tertinggi terhadap prediksi penyakit jantung.[14][15].

## 2. METODOLOGI PENELITIAN

### 2.1 Tahapan Alur Penelitian

Penelitian ini menggunakan pendekatan eksperimen kuantitatif berbasis Machine Learning. Tahapan penelitian disusun secara berurutan mulai dari pengumpulan data, eksplorasi data, prapemrosesan, pembagian data, pemodelan Logistic Regression dan Random Forest, optimasi parameter Random Forest, evaluasi performa, serta analisis feature importance. Alur tersebut dirancang agar kedua model diuji pada kondisi data yang sama sehingga hasil perbandingan dapat dianalisis secara objektif.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400




**Gambar 1.** Alur penelitian komparasi model

### 2.2 Dataset dan Atribut Penelitian

Dataset yang digunakan merupakan data sekunder Heart_Disease_Prediction.csv yang diperoleh dari platform Kaggle. Dataset berisi 270 data pasien dengan 14 atribut, terdiri atas 13 fitur klinis dan 1 atribut target. Atribut target Heart Disease memiliki dua kategori, yaitu Presence untuk pasien yang terindikasi penyakit jantung dan Absence untuk pasien yang tidak terindikasi penyakit jantung. Ringkasan dataset ditunjukkan pada Tabel 1.

**Tabel 1.** Ringkasan dataset

| Komponen | Jumlah/Keterangan |
|---|---|
| Jumlah Data | 270 observasi pasien |
| Jumlah Fitur Prediktor | 13 fitur klinis |
| Atribut Target | Heart Disease |
| Kelas Target | Presence dan Absence |
| Sumber Data | Heart_Disease_Prediction.csv |

Fitur prediktor yang digunakan meliputi Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Max HR, Exercise Angina, ST Depression, Slope of ST, Number of Major Vessels Fluro, dan Thallium. Atribut tersebut merepresentasikan informasi klinis yang relevan dalam proses klasifikasi risiko penyakit jantung.

### 2.3 Prapemrosesan Data

Tahapan prapemrosesan dilakukan untuk memastikan data berada dalam format yang layak digunakan oleh algoritma klasifikasi. Proses yang dilakukan meliputi pemeriksaan nilai kosong, pemeriksaan data duplikat, label encoding pada atribut target, pembagian dataset, dan standardisasi data. Hasil pemeriksaan menunjukkan bahwa seluruh atribut tidak memiliki nilai kosong sehingga proses imputasi tidak diperlukan. Atribut target kemudian dikonversi menjadi bentuk numerik, yaitu Absence = 0 dan Presence = 1.

**Tabel 2.** Pembagian dataset

| Jenis Data | Persentase | Jumlah Data |
|---|---|---|
| Data Latih (Training Data) | 80% | 216 |
| Data Uji (Testing Data) | 20% | 54 |
| Total | 100% | 270 |

Pembagian data dilakukan menggunakan skema train_test_split dengan rasio 80:20 dan random_state = 42. StandardScaler diterapkan pada data yang digunakan oleh Logistic Regression karena model linier berbasis optimasi matematis sensitif terhadap perbedaan skala antarfitur. Random Forest tidak memerlukan standardisasi karena bekerja berdasarkan pemisahan fitur pada struktur decision tree.

### 2.4 Pemodelan dan Optimasi

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

Model pertama yang dibangun adalah Logistic Regression. Model ini dipilih sebagai representasi algoritma linier yang sederhana, transparan, dan mudah diinterpretasikan. Model kedua adalah Random Forest Classifier yang mewakili algoritma ensemble berbasis decision tree. Konfigurasi awal Random Forest menggunakan n_estimators = 100 dan random_state = 42. Kedua model dilatih menggunakan data latih yang sama dan diuji menggunakan data uji yang sama agar hasil perbandingan bersifat adil.

Setelah evaluasi awal, optimasi parameter hanya diterapkan pada Random Forest menggunakan GridSearchCV. Parameter yang diuji adalah n_estimators dengan nilai 50, 100, dan 200; max_depth dengan nilai None, 5, dan 10; serta min_samples_split dengan nilai 2, 5, dan 10. Proses pencarian dilakukan menggunakan 5-Fold Cross Validation untuk memperoleh konfigurasi parameter terbaik.

### 2.5 Metrik Evaluasi

Evaluasi performa model dilakukan menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC. Accuracy digunakan untuk melihat proporsi prediksi yang benar secara keseluruhan. Precision menunjukkan ketepatan prediksi positif, sedangkan recall atau sensitivity menunjukkan kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya. Pada konteks prediksi penyakit jantung, recall menjadi metrik yang sangat penting karena nilai recall yang tinggi dapat membantu meminimalkan false negative, yaitu kondisi ketika pasien yang sebenarnya menderita penyakit jantung diprediksi sehat. F1-score digunakan untuk menilai keseimbangan precision dan recall, sedangkan ROC-AUC digunakan untuk menilai kemampuan model membedakan kelas positif dan negatif.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Hasil Analisis Eksplorasi Data

Analisis eksplorasi data dilakukan untuk memahami struktur dataset sebelum pemodelan. Hasil analisis menunjukkan bahwa dataset terdiri atas 270 observasi pasien dan memiliki distribusi kelas target yang relatif seimbang antara Presence dan Absence. Kondisi ini penting karena ketidakseimbangan kelas dapat menyebabkan model lebih condong memprediksi kelas mayoritas. Heatmap korelasi menunjukkan bahwa beberapa fitur klinis memiliki hubungan yang cukup relevan dengan target penyakit jantung, terutama Chest Pain Type, ST Depression, dan Number of Major Vessels Fluro.

Analisis outlier melalui boxplot menunjukkan adanya nilai ekstrem pada beberapa atribut numerik seperti Serum Cholesterol dan Resting Blood Pressure. Namun, nilai tersebut tidak dihapus karena masih dapat merepresentasikan kondisi klinis yang mungkin terjadi pada pasien. Dengan demikian, data tetap dipertahankan agar model dapat mempelajari variasi klinis yang lebih utuh.

**a. Gambaran umum Dataset**

Tahap pertama dalam penelitian ini adalah melakukan analisis eksplorasi data (Exploratory Data Analysis/EDA) terhadap dataset Heart Disease Prediction. Analisis ini bertujuan untuk memahami karakteristik data, mengidentifikasi distribusi variabel, mendeteksi adanya nilai kosong (missing value), serta memperoleh gambaran awal mengenai hubungan antarvariabel yang digunakan dalam proses klasifikasi penyakit jantung. Dataset yang digunakan terdiri atas 270 data pasien dengan 14 atribut yang mencakup 13 fitur prediktor dan 1 variabel target yaitu Heart Disease.

**Tabel 3.** Ringkasan Dataset

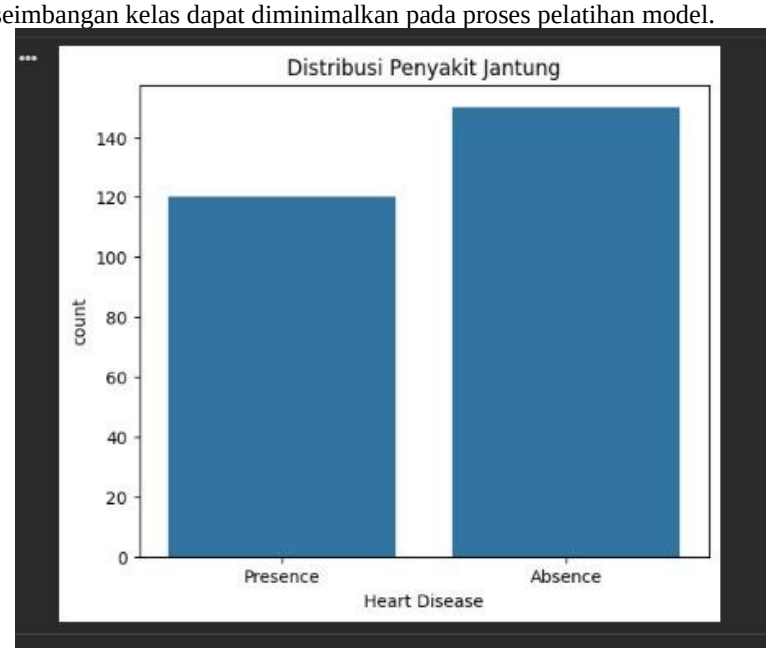
| Komponen | Jumlah |
|---|---|
| Jumlah Data | 270 |
| Jumlah Fitur | 13 |
| Target | Heart Disease |
| Kelas Target | Presence dan Absence |

Sumber : Hasil olahan dataset Heart_Disease_Predicton.csv (2025)

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

**b. Distribusi Kelas Target**

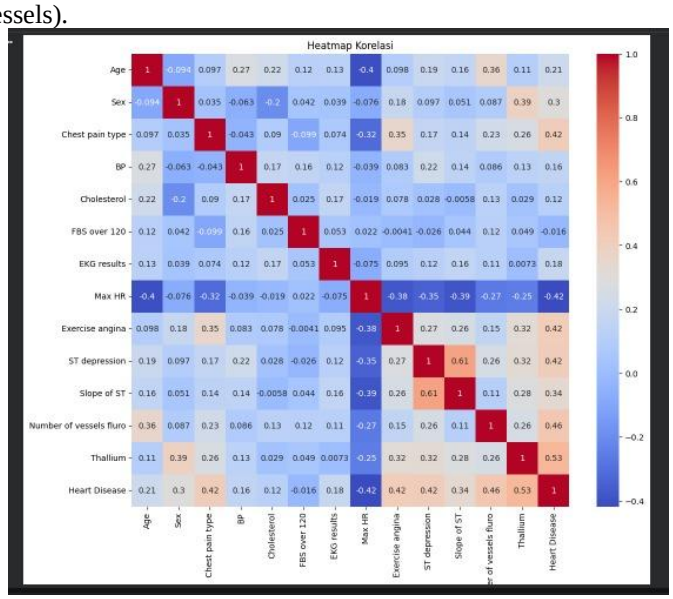
Analisis distribusi kelas dilakukan untuk mengetahui keseimbangan jumlah data pada masing-masing kategori target. Hasil visualisasi menunjukkan bahwa distribusi data antara kelas Presence dan Absence relatif seimbang sehingga risiko bias akibat ketidakseimbangan kelas dapat diminimalkan pada proses pelatihan model.




**Gambar 2.** Grafik Distribusi Target.

**c. Analisis Korelasi Fitur**

Analisis korelasi dilakukan untuk mengetahui tingkat hubungan antar fitur dalam dataset. Heatmap korelasi menunjukkan bahwa beberapa variabel klinis memiliki hubungan yang cukup kuat dengan variabel target penyakit jantung, terutama tipe nyeri dada (Chest Pain Type), depresi segmen ST (ST Depression), dan jumlah pembuluh darah utama (Number of Major Vessels).

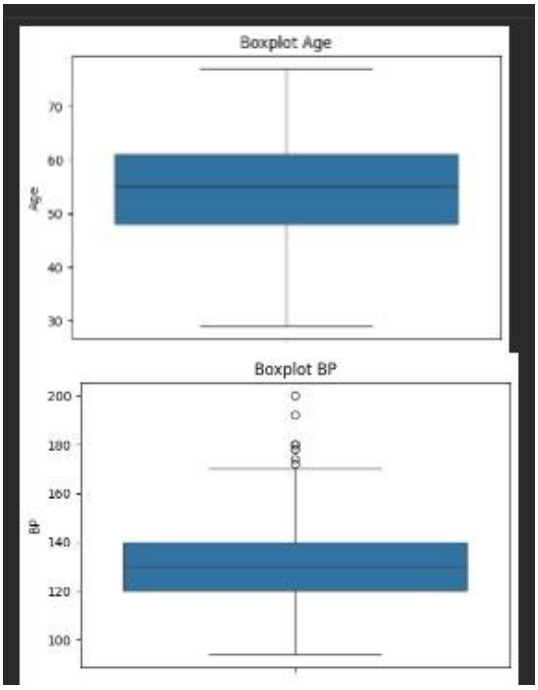



**Gambar 3.** Heatmap Korelasi

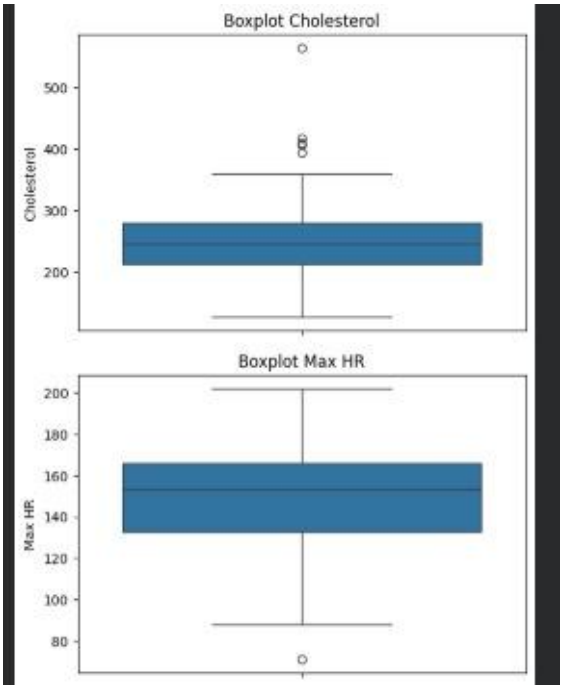
JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

**d. Analisis Outlier**

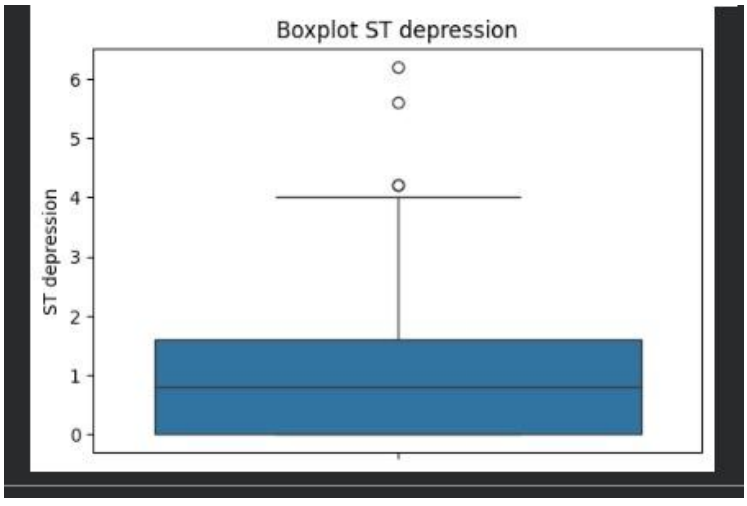
Hasil visualisasi boxplot menunjukkan bahwa beberapa atribut numerik seperti Serum Cholesterol dan Resting Blood Pressure memiliki nilai ekstrem (outlier). Namun demikian, data tersebut tidak dihapus karena masih berada dalam rentang yang secara klinis memungkinkan terjadi pada pasien penyakit jantung sehingga tetap dipertahankan dalam proses analisis.




(a)




(b)




(c)

**Gambar 4.** (a),(b)dan(c)Analisis Outlier Boxplot

### 3.2 Hasil Prapemrosesan Data

Pemeriksaan missing value menunjukkan bahwa seluruh atribut tidak memiliki nilai kosong. Variabel target Heart Disease yang semula berbentuk kategori teks dikonversi menjadi nilai numerik menggunakan Label Encoding, dengan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

Absence sebagai 0 dan Presence sebagai 1. Dataset kemudian dibagi menjadi 216 data latih dan 54 data uji. Tahapan ini memastikan data siap digunakan dalam proses pelatihan dan pengujian model klasifikasi.

### 3.3 Evaluasi Performa Logistic Regression dan Random Forest

Hasil pengujian menunjukkan bahwa Logistic Regression memperoleh performa terbaik dibandingkan Random Forest konfigurasi awal. Logistic Regression menghasilkan accuracy sebesar 85,19%, precision 78,57%, recall 91,67%, F1-score 84,62%, dan ROC-AUC 89,86%. Nilai recall yang tinggi menunjukkan bahwa model mampu mendeteksi sebagian besar pasien yang benar-benar memiliki indikasi penyakit jantung.

Random Forest memperoleh accuracy sebesar 81,48%, precision 76,92%, recall 83,33%, F1-score 80,00%, dan ROC-AUC 86,74%. Meskipun Random Forest tetap menunjukkan performa yang baik, hasil tersebut masih berada di bawah Logistic Regression. Perbedaan ini menunjukkan bahwa pada dataset klinis dengan jumlah observasi terbatas, model yang lebih sederhana dapat menghasilkan generalisasi yang lebih baik dibandingkan model ensemble yang lebih kompleks.

**Tabel 3.** Hasil evaluasi model klasifikasi

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 85,19% | 78,57% | 91,67% | 84,62% | 89,86% |
| Random Forest | 81,48% | 76,92% | 83,33% | 80,00% | 86,74% |

### 3.4 Hasil Optimasi Hyperparameter Random Forest

Optimasi Random Forest dilakukan menggunakan GridSearchCV dengan 5-Fold Cross Validation. Proses pencarian parameter menghasilkan konfigurasi terbaik berupa max_depth = 5, min_samples_split = 2, dan n_estimators = 200. Konfigurasi tersebut menunjukkan bahwa pembatasan kedalaman pohon dan peningkatan jumlah pohon dapat membantu model meningkatkan stabilitas prediksi.

**Tabel 4.** Parameter terbaik Random Forest

| Parameter | Nilai Terbaik |
|---|---|
| max_depth | 5 |
| min_samples_split | 2 |
| n_estimators | 200 |

Setelah tuning, accuracy Random Forest meningkat dari 81,48% menjadi 83,33%. Peningkatan ini menunjukkan bahwa tuning berhasil memperbaiki performa model. Namun, hasil tersebut tetap belum mampu melampaui Logistic Regression yang memperoleh accuracy 85,19%. Oleh karena itu, optimasi parameter Random Forest belum cukup signifikan untuk mengubah kesimpulan model terbaik pada dataset ini.

**Tabel 5.** Perbandingan accuracy sebelum dan sesudah tuning

| Model | Accuracy |
|---|---|
| Random Forest Awal | 81,48% |
| Random Forest Tuning | 83,33% |

### 3.5 Analisis Komparasi Model

Berdasarkan hasil keseluruhan, Logistic Regression menjadi model paling unggul karena memiliki accuracy, recall, F1-score, dan ROC-AUC tertinggi. Dalam konteks kesehatan, recall menjadi perhatian utama karena berkaitan dengan kemampuan model mendeteksi pasien yang benar-benar menderita penyakit jantung. Nilai recall Logistic Regression sebesar 91,67% lebih tinggi dibandingkan Random Forest sebesar 83,33%, sehingga model ini lebih efektif dalam meminimalkan risiko false negative.

**Tabel 6.** Rekapitulasi komparasi model

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Logistic Regression | 85,19% | 78,57% | 91,67% | 84,62% |
| Random Forest | 81,48% | 76,92% | 83,33% | 80,00% |
| RF Tuning | 83,33% | - | - | - |

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

Temuan ini memperkuat asumsi bahwa kompleksitas algoritma tidak selalu menghasilkan performa yang lebih baik. Pada dataset berukuran kecil, model sederhana dapat lebih stabil karena tidak terlalu banyak membentuk pola kompleks yang berpotensi tidak tergeneralisasi pada data baru. Logistic Regression juga memiliki keunggulan interpretabilitas yang lebih baik, sehingga lebih mudah dijelaskan sebagai dasar sistem pendukung keputusan.

### 3.6 Analisis Tingkat Kepentingan Fitur

Analisis feature importance dilakukan untuk mengidentifikasi atribut klinis yang paling berkontribusi dalam proses klasifikasi. Hasil analisis menunjukkan bahwa Chest Pain Type memiliki nilai importance tertinggi sebesar 0,129. Hal ini menunjukkan bahwa jenis nyeri dada menjadi indikator penting dalam membedakan pasien yang terindikasi penyakit jantung dan yang tidak.

**Tabel 7.** Lima fitur klinis dengan tingkat kepentingan tertinggi

| Ranking | Fitur | Importance |
|---|---|---|
| 1 | Chest Pain Type | 0,129 |
| 2 | Max HR | 0,118 |
| 3 | ST Depression | 0,110 |
| 4 | Number of Major Vessels Fluro | 0,108 |
| 5 | Thallium | 0,101 |

Selain Chest Pain Type, fitur Max HR dan ST Depression juga memiliki kontribusi yang tinggi. Kedua fitur ini berkaitan dengan respons fisiologis jantung, terutama saat aktivitas atau pemeriksaan tertentu. Temuan tersebut menunjukkan bahwa gejala klinis yang berhubungan langsung dengan respons jantung lebih dominan dibandingkan beberapa indikator umum lainnya. Hasil ini dapat menjadi masukan awal bagi pengembangan sistem pendukung keputusan, meskipun validasi medis lanjutan tetap diperlukan sebelum model diterapkan dalam praktik klinis.

## 4. KESIMPULAN

Penelitian ini berhasil melakukan komparasi performa Logistic Regression dan Random Forest pada prediksi penyakit jantung berdasarkan analisis fitur klinis pasien. Dataset yang digunakan terdiri atas 270 observasi pasien dengan 13 fitur prediktor dan 1 atribut target. Tahapan penelitian mencakup EDA, preprocessing, pembagian data 80:20, pemodelan klasifikasi, optimasi Random Forest menggunakan GridSearchCV, evaluasi metrik klasifikasi, dan analisis feature importance. Hasil penelitian menunjukkan bahwa Logistic Regression memperoleh performa terbaik dengan accuracy 85,19%, precision 78,57%, recall 91,67%, F1-score 84,62%, dan ROC-AUC 89,86%. Random Forest memperoleh accuracy 81,48% dan meningkat menjadi 83,33% setelah dilakukan tuning dengan parameter terbaik max_depth = 5, min_samples_split = 2, dan n_estimators = 200. Meskipun tuning mampu meningkatkan performa Random Forest, hasilnya belum mampu melampaui Logistic Regression. Dalam konteks prediksi penyakit jantung, Logistic Regression dinilai lebih layak digunakan sebagai model pendukung deteksi dini karena menghasilkan recall tertinggi. Nilai recall yang tinggi menunjukkan kemampuan model dalam meminimalkan false negative, yaitu kesalahan ketika pasien yang sebenarnya berisiko diprediksi tidak berisiko. Analisis feature importance menunjukkan bahwa Chest Pain Type, Max HR, dan ST Depression merupakan tiga fitur klinis paling dominan dalam proses klasifikasi. Penelitian selanjutnya disarankan menggunakan dataset dengan jumlah observasi lebih besar, mengeksplorasi algoritma lain seperti XGBoost, LightGBM, atau SVM, serta melakukan validasi bersama tenaga medis agar model lebih aplikatif dalam lingkungan klinis.

## REFERENCES

[1] S. Albahra *et al.*, “Seminars in Diagnostic Pathology Artificial intelligence and machine learning overview in pathology & laboratory medicine : A general review of data preprocessing and basic supervised concepts,” vol. 40, no. February, pp. 71–87, 2023, doi: 10.1053/j.semdp.2023.02.002.

[2] A. M. Hernández and I. Van Nieuwenhuyse, *A survey on multi - objective hyperparameter optimization algorithms for machine learning*, vol. 56, no. 8. Springer Netherlands, 2023. doi: 10.1007/s10462-022-10359-2.

[3] S. D. Goals, *World health statistics 2025*. 2025.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 2 September 2026
Hal: 1713-1721
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i2.1400

[4] A. Cuevas-chávez *et al.*, "A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases," pp. 1–50, 2023.

[5] S. Hossain, M. K. Hasan, M. O. Faruk, N. Aktar, and R. Hossain, "Machine learning approach for predicting cardiovascular disease in Bangladesh : evidence from a cross - sectional study in 2023," *BMC Cardiovasc. Disord.*, pp. 1–28, 2024, doi: 10.1186/s12872-024-03883-2.

[6] A. Möller, "Predictive Power of Logistic Regres- sion versus Random Forest : A simu- lation study Matematiska institutionen".

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Springer Texts in Statistics An Introduction to Statistical Learning*.

[8] A. P. Jawalkar, P. Swetcha, N. Manasvi, P. Sreekala, and S. Aishwarya, "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," *J. Eng. Appl. Sci.*, pp. 1–18, 2023, doi: 10.1186/s44147-023-00280-y.

[9] H. Koffijberg, "Health Economic Research Assessing the Value of Early Detection of Cardiovascular Disease : A Systematic Review," *Pharmacoeconomics*, vol. 41, no. 10, pp. 1183–1203, 2023, doi: 10.1007/s40273-023-01287-2.

[10] H. R. Maier *et al.*, "On how data are partitioned in model development and evaluation : Confronting the elephant in the room to enhance model generalization ☆," *Environ. Model. Softw.*, vol. 167, no. June, p. 105779, 2023, doi: 10.1016/j.envsoft.2023.105779.

[11] Y. Mao, B. Legesse, and T. Belsty, "Current Problems in Cardiology Machine learning algorithms for heart disease diagnosis : A systematic review," *Curr. Probl. Cardiol.*, vol. 50, no. 8, p. 103082, 2025, doi: 10.1016/j.cpcardiol.2025.103082.

[12] N. Putri *et al.*, "PENERAPAN FEATURE ENGINEERING DAN HYPERPARAMETER TUNING UNTUK MENINGKATKAN AKURASI MODEL RANDOM FOREST PADA APPLICATION OF FEATURE ENGINEERING AND HYPERPARAMETER TUNING TO IMPROVE THE ACCURACY OF RANDOM FOREST MODELS ON CREDIT RISK," vol. 12, no. 2, pp. 251–262, 2025, doi: 10.25126/jtiik.2025128472.

[13] I. Muhamad and M. Matin, "Hyperparameter Tuning menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," vol. 9, no. 1, 2023.

[14] S. Rasheed, G. K. Kumar, D. M. Rani, M. V. V. P. Kantipudi, and M. Anila, "EAI Endorsed Transactions Heart Disease Prediction Using GridSearchCV and Random Forest," vol. 10, pp. 1–8, 2024, doi: 10.4108/eetpht.10.5523.

[15] L. Techniques, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," 2023.