Jurnal Informatika dan Tekonologi Komputer
E-ISSN: 2809-9230
P-ISSN: 2809-9249

Penerapan Algoritma Logistic Regression Untuk Memprediksi Penyakit Jantung

Muhammad Fitra Rhomadon [1*], Wydyanto [2], A. Haidar Mirza [3], dan Nurul Huda [4]

[1] Program Studi Teknik Informatika, Universitas Bina Darma; Kota Palembang, Sumatera Selatan; e-mail : muhammadfitrar1512@gmail.com
[2] Program Studi Teknik Informatika, Universitas Bina Darma; Kota Palembang, Sumatera Selatan; e-mail : Widyanto@binadarma.ac.id
[3] Program Studi Teknik Informatika, Universitas Bina Darma; Kota Palembang, Sumatera Selatan; e-mail : haidarmirza@binadarma.ac.id
[3] Program Studi Teknik Informatika, Universitas Bina Darma; Kota Palembang, Sumatera Selatan; e-mail : nurul_huda@binadarma.ac.id
* Corresponding Author : Muhammad Fitra Rhomadon

Abstract: Heart disease is one of the leading causes of death worldwide, including in Indonesia. Early detection of heart disease risk is crucial to prevent more severe complications and improve patients' quality of life. This study aims to apply the Logistic Regression algorithm to build a data-driven heart disease prediction model. The dataset used is from Kaggle, with 1,025 patient data and 14 attributes covering risk factors such as age, gender, blood pressure, cholesterol, maximum heart rate, and others. The research process was conducted using the CRISP-DM approach, which includes business understanding, data exploration, preprocessing, modeling, evaluation, and model testing. The preprocessing stage includes data cleaning, encoding categorical variables, standardizing numeric data, and dividing the data into training and test data. The model was developed using the Python programming language and the scikit-learn library, then evaluated using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC. The evaluation results showed that the Logistic Regression model was able to provide good prediction results, with an accuracy of 0.93, a precision of 0.93, a recall of 0.96, and an F1-score of 0.95. With this performance, this model can be used as a tool for medical personnel in early detection of heart disease risk and supporting more effective and efficient decision-making.

Keywords: Logistic Regression; Heart Disease; Prediction; Machine Learning; Data Mining.

Received: August 14, 2025
Revised: August 30, 2025
Accepted: November 26, 2025
Published: November 29, 2025
Curr. Ver.: November 29, 2025

Abstrak: Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia, termasuk di Indonesia. Deteksi dini terhadap risiko penyakit jantung sangat penting untuk mencegah komplikasi yang lebih parah dan meningkatkan kualitas hidup pasien. Penelitian ini bertujuan untuk menerapkan algoritma Logistic Regression dalam membangun model prediksi penyakit jantung berbasis data. Dataset yang digunakan berasal dari Kaggle, dengan jumlah 1.025 data pasien dan 14 atribut yang mencakup faktor-faktor risiko seperti usia, jenis kelamin, tekanan darah, kolesterol, denyut jantung maksimum, dan lain-lain. Proses penelitian dilakukan menggunakan pendekatan CRISP-DM yang mencakup tahapan pemahaman bisnis, eksplorasi data, preprocessing, pemodelan, evaluasi, dan uji coba model. Tahapan preprocessing meliputi pembersihan data, *encoding* variabel kategorikal, standarisasi data numerik, dan pembagian data menjadi data latih dan data uji. Model dikembangkan menggunakan bahasa pemrograman Python dan pustaka scikit-learn, lalu dievaluasi menggunakan metrik akurasi, precision, recall, F1-score, confusion matrix, serta ROC-AUC. Hasil evaluasi menunjukkan bahwa model Logistic Regression mampu memberikan hasil prediksi yang baik, dengan nilai akurasi sebesar 0,93, precision 0,93, recall 0,96, dan F1-score 0,95. Dengan performa tersebut, model ini dapat digunakan sebagai alat bantu untuk tenaga medis dalam mendeteksi dini risiko penyakit jantung dan mendukung pengambilan keputusan secara lebih efektif dan efisien.

Kata kunci: Logistik Regresi; Penyakit Jantung; Prediksi; Machine Learning; Data Mining.

1. Pendahuluan

Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Menurut laporan dari World Health Organization (WHO), penyakit kardiovaskular bertanggung jawab atas sekitar 17,9 juta kematian setiap tahunnya, atau sekitar 31% dari total kematian global [1]. Di Indonesia, prevalensi penyakit jantung terus mengalami peningkatan dari tahun ke tahun. Berdasarkan data Riset Kesehatan Dasar (Riskesdas) yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia, prevalensi penyakit jantung di Indonesia pada tahun 2013 tercatat sebesar 1,5% dan meningkat menjadi 1,9% pada tahun 2018 [2]. Kondisi ini menunjukkan bahwa penyakit jantung masih menjadi ancaman serius terhadap kesehatan masyarakat, sehingga diperlukan upaya pencegahan dan deteksi dini yang lebih efektif.

Penyakit jantung dapat menyerang siapa saja tanpa memandang usia, jenis kelamin, maupun latar belakang sosial ekonomi. Faktor risiko penyakit ini secara umum dapat diklasifikasikan menjadi dua, yaitu faktor yang tidak dapat dimodifikasi seperti usia, jenis kelamin, dan faktor genetik, serta faktor yang dapat dimodifikasi seperti dislipidemia, hipertensi, diabetes melitus, kebiasaan merokok, kurangnya aktivitas fisik, dan pola makan yang tidak sehat [3] . Dengan demikian, pengendalian terhadap faktor risiko yang dapat dimodifikasi menjadi sangat penting dalam upaya menekan angka kematian akibat penyakit jantung.

Deteksi dini memiliki peran yang sangat krusial dalam mengidentifikasi faktor risiko maupun gejala awal penyakit jantung sebelum berkembang menjadi kondisi yang lebih berat. Langkah ini dapat mencegah komplikasi lanjutan dan meningkatkan efektivitas intervensi medis yang dilakukan [4]. Seiring dengan perkembangan teknologi di bidang kesehatan, pemanfaatan metode analisis data dan kecerdasan buatan mulai banyak diterapkan untuk meningkatkan akurasi dalam deteksi dini penyakit. Salah satu algoritma yang cukup populer dalam memprediksi kejadian biner seperti keberadaan penyakit jantung adalah algoritma *Logistic Regression*. Algoritma ini mampu memodelkan hubungan antara beberapa variabel independen seperti usia, tekanan darah, kolesterol, dan kebiasaan merokok dengan kemungkinan terjadinya penyakit jantung.

Beberapa penelitian telah menunjukkan efektivitas algoritma *Logistic Regression* dalam memprediksi resiko penyakit jantung. Pada penelitian terdahulu, penerapan algoritma *Logistic Regression* terbukti efektif dalam memprediksi risiko penyakit jantung dengan tingkat akurasi yang cukup tinggi [5]. Namun, terdapat tantangan dalam penanganan data yang hilang, seleksi fitur yang tepat, serta integrasi dengan teknologi medis terkini yang perlu diperhatikan untuk meningkatkan akurasi dan efektivitas deteksi penyakit jantung. Selain itu, penelitian lain juga menunjukkan bahwa algoritma Logistic Regression memiliki tingkat sensitivitas sebesar 88,54%

pada data training dan spesifisitas sebesar 87,50% pada data testing [6]. Hal ini menunjukkan bahwa algoritma tersebut cukup andal dalam mendeteksi keberadaan penyakit jantung.

Beberapa penelitian juga membandingkan algoritma *Logistic Regression* dengan algoritma K-Nearest Neighbor (K-NN), dan hasilnya menunjukkan bahwa algoritma Logistic Regression menghasilkan akurasi lebih tinggi, yaitu sebesar 88% dibandingkan dengan algoritma K-NN yang hanya mencapai 83% [7]. Hal serupa ditemukan oleh [8] yang mengembangkan sistem prediksi berbasis website dengan menggunakan algoritma Logistic Regression dan memperoleh akurasi sebesar 90%, dengan presisi sebesar 92%, recall 86%, dan f1-score 89%. Penelitian lain juga mendukung keunggulan algoritma *Logistic Regression* dengan akurasi sebesar 88,52% mengungguli algoritma lain seperti K-NN dan *Random Forest* [9].

Selanjutnya, penelitian lain juga menyebutkan bahwa efektifitas model prediksi penyakit jantung berbasis *Logistic Regression* menghasilkan akurasi sebesar 86% pada data training dan 88% pada data testing, dengan nilai *Area Under Curve* (AUC) sebesar 0,95 [10]. Nilai AUC ini menunjukkan performa model yang sangat baik dalam mengklasifikasikan data. Adapun penelitian dari ref [11] menunjukkan bahwa algoritma Logistic Regression mencapai akurasi sebesar 85% dan direkomendasikan sebagai salah satu metode efektif dalam prediksi penyakit jantung.

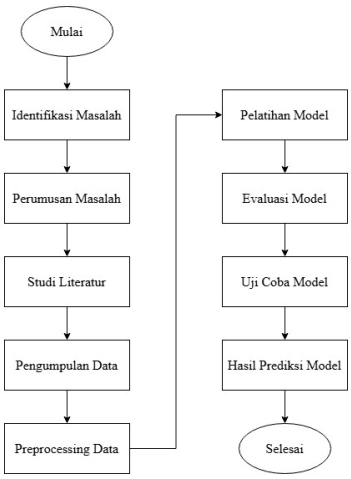
Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa algoritma Logistic Regression memiliki potensi besar dalam mendeteksi dan memprediksi risiko penyakit jantung secara akurat. Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma Logistic Regression dalam membangun model prediksi penyakit jantung guna meningkatkan akurasi identifikasi pasien berisiko serta mendukung upaya pencegahan penyakit jantung secara lebih efektif. Dengan model prediksi yang tepat, diharapkan tenaga medis dapat terbantu dalam pengambilan keputusan dan mampu memberikan intervensi yang lebih cepat serta meningkatkan kualitas hidup pasien.

2. Metode yang Diusulkan

Penelitian ini menggunakan pendekatan kuantitatif yang bertujuan untuk membangun model prediksi penyakit jantung berdasarkan data yang telah ada. Model tersebut dibuat menggunakan algoritma *Logistic Regression*, yang dikategorikan dalam metode klasifikasi biner [12]. Pendekatan kuantitatif dipilih karena hasil yang diinginkan berupa angka dan persentase akurasi dari model prediksi yang dibangun. Penelitian ini bersifat terapan (*applied research*) karena hasil akhirnya diharapkan bisa digunakan untuk membantu pengambilan keputusan di bidang kesehatan, khususnya dalam mendeteksi potensi penyakit jantung secara dini. Teknik analisis data dalam penelitian ini dilakukan melalui pendekatan kuantitatif komputasional, yaitu dengan menggunakan algoritma *machine learning Logistic Regression* untuk melakukan prediksi terhadap kemungkinan seseorang menderita penyakit jantung berdasarkan atribut-atribut yang tersedia dalam dataset. Seluruh proses analisis data dilakukan menggunakan

bahasa pemrograman Python, yang didukung oleh pustaka (*library*) seperti pandas, numpy, scikit-learn, matplotlib, dan seaborn.

Alur penelitian merupakan tahapan sistematis yang menggambarkan proses yang dilakukan peneliti dari awal hingga akhir dalam menyelesaikan penelitian. Dalam penelitian ini, alur dimulai dari identifikasi masalah hingga evaluasi dan interpretasi hasil dari model prediksi penyakit jantung menggunakan algoritma *Logistic Regression*.



Gambar 1. Alur Penelitian

Penjelasan dari gambar diatas adalah sebagai berikut:

2.1 Identifikasi Masalah

Peneliti mengamati fenomena bahwa penyakit jantung merupakan salah satu penyebab utama kematian di dunia, termasuk di Indonesia. Di sisi lain, banyak kasus penyakit jantung yang baru terdeteksi saat sudah dalam tahap lanjut. Oleh karena itu, dibutuhkan metode prediksi dini untuk membantu deteksi lebih awal, salah satunya dengan algoritma Logistic Regression.

2.2 Perumusan Masalah

Peneliti merumuskan masalah utama yaitu Bagaimana penerapan algoritma Logistic Regression dapat digunakan untuk memprediksi kemungkinan seseorang terkena penyakit jantung berdasarkan data riwayat kesehatan?

2.3 Studi Literatur

Peneliti mengkaji sumber-sumber ilmiah yang relevan, baik jurnal maupun laporan penelitian sebelumnya. Kajian ini meliputi:

a. Teori penyakit jantung dan faktor risikonya
b. Konsep Logistic Regression
c. Penerapan *machine learning* di bidang medis

Tujuannya adalah memperkuat landasan teori serta menemukan celah penelitian sebelumnya yang bisa dikembangkan.

2.4 Pengumpulan Data

Data yang digunakan diperoleh dari data sekunder yaitu dengan cara mengunduh dataset "*Heart Disease Dataset*" dari situs Kaggle. Dataset ini berisi 1025 data pasien dengan 14 atribut seperti usia, tekanan darah, kolesterol, kadar gula darah, detak jantung maksimal, hasil tes EKG, dll. Dataset ini menjadi sumber utama dalam pelatihan dan pengujian model prediksi.

2.5 Preprocessing Data

Sebelum data digunakan, perlu dilakukan proses pembersihan dan transformasi data. Tahapan ini mencakup:

a. Menangani data kosong (*missing values*)
b. Menormalisasi data numerik
c. Mengubah data kategorikal menjadi numerik (*encoding*)
d. Menghapus data duplikat atau *Outlier*
e. Memisahkan data menjadi fitur (x) dan target (y)

Hasil preprocessing ini akan menjadi input untuk tahap pelatihan model.

2.6 Pelatihan Model

Pada tahap ini, data yang telah dibersihkan digunakan untuk melatih algoritma Logistic Regression. Tujuannya adalah membuat model yang bisa mempelajari pola-pola dari data sehingga dapat digunakan untuk memprediksi apakah seseorang berisiko terkena penyakit jantung atau tidak.Langkah umum yang dilakukan adalah dengan membagi data menjadi data latih dan data uji (80:20) dan menggunakan *scikit-learn* untuk membangun dan melatih model

2.7 Evaluasi Model

Setelah model dilatih, dilakukan pengujian terhadap data uji. Tujuannya untuk mengetahui seberapa baik model dalam melakukan prediksi. Evaluasi dilakukan menggunakan metrik seperti Akurasi, Precision, Recall, F1-Score, ROC-AUC Score, Confusion Matrix. Evaluasi ini penting untuk memastikan model tidak hanya menghafal data latih, tetapi mampu melakukan generalisasi.

2.8 Uji Coba Model

Setelah model dievaluasi, model digunakan untuk memprediksi risiko penyakit jantung pada data baru. Algoritma Logistic Regression menghasilkan probabilitas antara 0 dan 1. Misalnya, jika probabilitasnya lebih besar dari ambang batas tertentu (misalnya 0.5), maka pasien dikategorikan memiliki risiko tinggi terkena penyakit jantung.

2.9 Hasil Prediksi Model

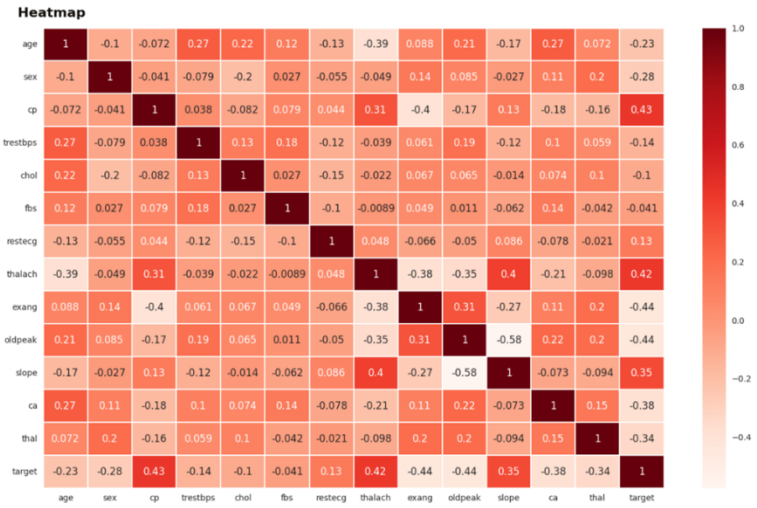
Setelah model di uji coba, model akan memberikan hasil prediksi berupa output apakah seseorang tersebut terkena penyakit jantung atau tidak.

3. Hasil dan Pembahasan

Dalam penelitian ini, pengumpulan Dataset berasal dari kompilasi data pasien yang berasal dari Cleveland, Hungary, Switzerland, and Long Beach V dan digunakan untuk tujuan analisis prediktif dalam ranah medis. Dataset telah banyak digunakan dalam publikasi ilmiah dan kompetisi analisis data, sehingga cukup kredibel untuk dijadikan objek penelitian. Dataset yang diperoleh berisi 76 atribut, termasuk atribut yang diprediksi, tetapi semua eksperimen yang dipublikasikan merujuk pada penggunaan subset yang terdiri dari 14 atribut. Kolom "target" mengacu pada keberadaan penyakit jantung pada pasien. Dataset ini terdiri dari 1025 data pasien dan 14 Variabel / Kolom Prediksi Penyakit Jantung. Dari dataset yang telah dikumpulkan maka data akan diolah sedemikian rupa sehingga memperoleh suatu prediksi atau keputusan.

3.1. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) merupakan tahap awal dalam proses analisis data yang bertujuan untuk memahami karakteristik data melalui teknik visualisasi, statistik deskriptif, dan deteksi nilai ekstrem atau tidak normal. Berdasarkan distribusi data yang telah dilakukan terhadap variabel penelitian, representasi hasil EDA dapat di gambarkan Heatmap sebagai berikut.



Gambar 2. Heatmap

Heatmap ini menunjukkan hubungan antara masing-masing variabel (fitur) dalam dataset prediksi penyakit jantung. Fokus utama adalah melihat seberapa besar hubungan (korelasi) antara fitur-fitur tersebut dengan variabel target, yaitu apakah seseorang menderita penyakit jantung atau tidak (1 = sakit, 0 = tidak sakit). Berdasarkan hasil heatmap diatas diketahui fitur-fitur yang memiliki pengaruh kuat terhadap target, baik dalam arah positif maupun negatif dapat dijelaskan sebagai berikut.

3.1.1.Fitur yang Paling Berpengaruh (Korelasi Tinggi)

a. cp (*Chest Pain* / Jenis Nyeri Dada) – Korelasi: 0.43

Artinya, jenis nyeri dada memiliki hubungan kuat dengan kemungkinan seseorang terkena penyakit jantung. Semakin "serius" tipe nyeri dada (misalnya *typical angina* atau *asymptomatic*), semakin besar peluang adanya penyakit jantung.

b. thalach (Maximum Heart Rate / Detak Jantung Maksimum) – Korelasi: 0.42

Semakin tinggi detak jantung maksimum yang dicapai saat tes olahraga, semakin besar kemungkinan seseorang menderita penyakit jantung.

c. slope (Kemiringan ST Segmen Setelah Olahraga) – Korelasi: 0.35

Menggambarkan bentuk segmen ST setelah olahraga. Nilai tertentu pada slope berhubungan dengan risiko lebih besar terkena penyakit jantung.

d. oldpeak (Depresi ST) – Korelasi: -0.44

Menunjukkan seberapa besar penurunan segmen ST dibandingkan kondisi istirahat. Semakin besar nilai oldpeak, semakin besar kemungkinan ada gangguan jantung.

e. exang (Exercise Induced Angina / Nyeri Dada Saat Olahraga) – Korelasi: -0.44

Jika seseorang mengalami nyeri dada saat olahraga (*exang = 1*), maka besar kemungkinan ia menderita penyakit jantung.

f. ca (Jumlah Pembuluh Darah Utama yang Terlihat) – Korelasi: -0.38

Semakin banyak pembuluh darah yang terlihat saat pemeriksaan (biasanya lewat fluoroskopi), semakin besar indikasi masalah jantung.

g. thal (Jenis Kelainan Aliran Darah) – Korelasi: -0.34

Tipe hasil dari tes thal (misalnya fixed defect, reversible defect) berpengaruh terhadap diagnosis penyakit jantung.

3.1.2. Fitur yang Kurang Berpengaruh (Korelasi Lemah)

a. *sex* (Jenis Kelamin) – Korelasi: -0.28

Meski sedikit berpengaruh, jenis kelamin pria cenderung memiliki risiko lebih tinggi terkena penyakit jantung.

b. *age* (Usia) – Korelasi: -0.23

Usia sedikit mempengaruhi, namun tidak sepenting faktor lain seperti detak jantung atau nyeri dada.

***c. trestbps* (Tekanan Darah Saat Istirahat) – Korelasi: 0.14**

Korelasinya lemah. Tekanan darah saat istirahat bukan penentu utama penyakit jantung dalam dataset ini.

d. restecg (Hasil EKG Saat Istirahat) – Korelasi: 0.13

Pemeriksaan EKG saat istirahat punya pengaruh kecil terhadap diagnosis penyakit jantung.

e. *chol* (Kolesterol Total) – Korelasi: -0.10

Meski kolesterol sering dikaitkan dengan penyakit jantung, pada dataset ini korelasinya lemah.

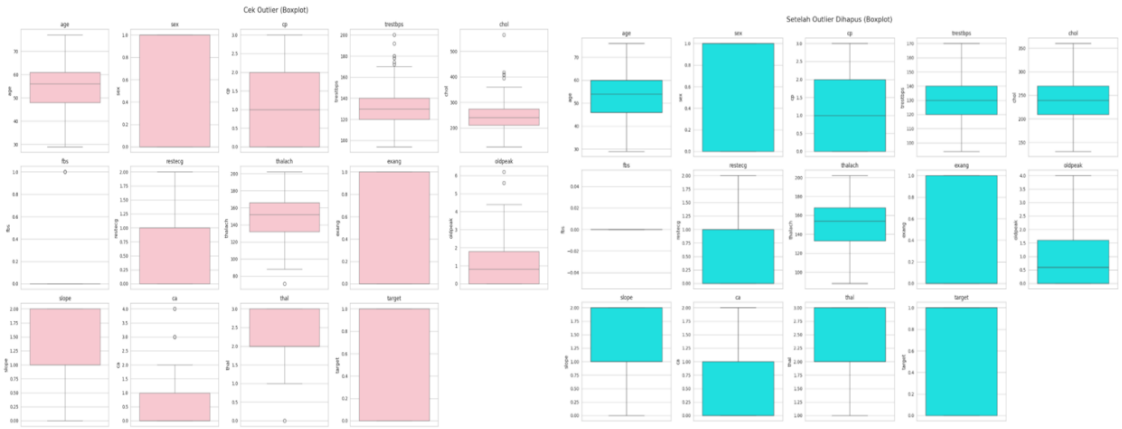
f. *fbs* (Gula Darah Puasa) – Korelasi: -0.041

Tidak memiliki pengaruh yang signifikan terhadap penyakit jantung dalam data ini.

3.2. Preprocessing data

Preprocessing data (pra-pemrosesan data) adalah proses pengolahan data yang bertujuan untuk membersihkan, merapikan, dan menyiapkan data mentah agar siap digunakan dalam analisis atau model machine learning [13]. Preprocessing data dimulai dari memeriksa *Outlier* guna nilai data yang menyimpang jauh dari sebagian besar data lainnya dalam sebuah dataset. *Outlier* dapat disebabkan oleh kesalahan pencatatan, variasi alami, maupun kesalahan alat ukur, dan berpotensi memengaruhi hasil analisis atau performa model prediksi. Oleh karena itu, deteksi dan penanganan *Outlier* menjadi langkah awal yang penting.

Metode yang digunakan untuk mendeteksi *Outlier* dalam penelitian ini adalah *Interquartile Range (IQR)*, yaitu rentang antara kuartil ketiga (Q3) dan kuartil pertama (Q1). Data yang berada di luar batas bawah (*lower bound* = $Q1 - 1.5 \times IQR$) dan batas atas (*upper bound* = $Q3 + 1.5 \times IQR$) dikategorikan sebagai *Outlier* dan dihapus dari dataset. Hasil visualisasi boxplot sebelum dan sesudah penghapusan *Outlier* menunjukkan penyebaran data yang lebih merata serta hilangnya nilai-nilai ekstrem. Langkah ini menghasilkan pengurangan jumlah data dari 1025 baris menjadi 918 baris. Proses ini meningkatkan kualitas data yang akan digunakan dalam pemodelan prediktif agar hasilnya lebih akurat dan dapat diandalkan. Gambar berikut menunjukkan boxplot sebagai hasil dari proses deteksi dan penghapusan *Outlier* sebagai berikut:



Sebelum Sesudah

Gambar 3. Hasil Boxplot Sebelum dan Sesudah Penghapusan *Outlier*

Dari visualisasi di atas, terlihat bahwa nilai-nilai pencilan pada beberapa fitur berhasil diidentifikasi dan dihilangkan. Setelah proses ini, jumlah data berkurang dari 1025 menjadi 918 baris.

Langkah selanjutnya adalah mengecek keberadaan data duplikat. Hasil analisis menunjukkan bahwa terdapat 541 baris data duplikat yang berpotensi menurunkan kualitas model dan menyebabkan overfitting. Oleh karena itu, data duplikat tersebut dihapus, sehingga jumlah akhir data menjadi 377 baris unik. Tahap terakhir dalam preprocessing adalah pemeriksaan terhadap missing value. Pemeriksaan ini menunjukkan bahwa tidak terdapat nilai kosong (missing) pada dataset, sehingga tidak diperlukan proses imputasi data. Setelah melakukan seluruh tahapan tersebut, maka dapat dipastikan bahwa seluruh data yang digunakan bersih, valid, dan optimal untuk proses pemodelan prediksi penyakit jantung menggunakan algoritma *Logistic Regression*.

3.3. Modeling

Modeling adalah proses membangun model matematis atau algoritma yang dapat mempelajari pola dari data dan kemudian digunakan untuk membuat prediksi atau keputusan [14]. Modeling pada penelitian ini dimulai dari membagi dataset menjadi fitur (X) dan target (y). Fitur (X) terdiri dari seluruh kolom kecuali kolom "target", sedangkan target (y) adalah variabel target yang menunjukkan status penyakit jantung (1 = ada, 0 = tidak ada). Selanjutnya membagi data menjadi data latih dan data uji menggunakan rasio 80:20. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa model terhadap data yang belum pernah dilihat sebelumnya. Kemudian dilakukan standarisasi fitur numerik agar skala setiap fitur menjadi seragam. Tujuan utama dari standarisasi ini adalah untuk mencegah dominasi fitur tertentu dalam proses pelatihan, serta meningkatkan performa dan konvergensi algoritma seperti Logistic Regression dan SVM. Standarisasi dilakukan dengan rumus:

$$\boldsymbol{z = \frac{x - \mu}{\sigma}}$$

Keterangan:
x = nilai asli
μ = rata-rata dari data
σ = standar deviasi dari data
z = nilai setelah distandarisasi

Langkah selanjutnya adalah melatih model menggunakan algoritma Logistic Regression. Algoritma ini dipilih karena sederhana namun efektif dalam menyelesaikan masalah klasifikasi biner. Setelah pelatihan, model diuji menggunakan data uji untuk melihat sejauh mana akurasi prediksinya. Evaluasi model dilakukan menggunakan beberapa metrik evaluasi seperti akurasi, precision, recall, dan F1-score. Akurasi model Logistic Regression dapat ditampilakan sebagai berikut.

```
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Akurasi Pada Data Pelatihan : ', training_data_accuracy)
```

```
Akurasi Pada Data Pelatihan :  0.8681318681318682
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

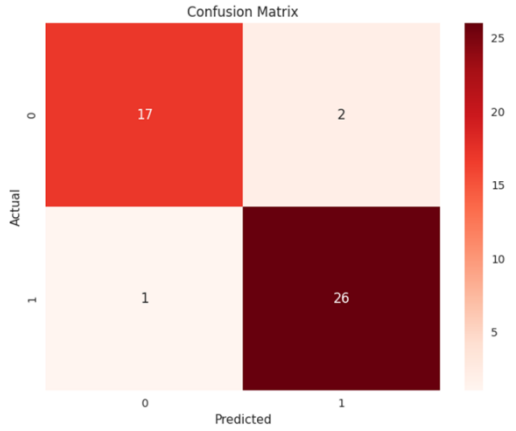
```
print('Akurasi Pada Data Uji : ', test_data_accuracy)
```

```
Akurasi Pada Data Uji :  0.9347826086956522
```

Gambar 4. Akurasi Model Logistic Regression

Berdasarkan gambar diatas diketahui hasil evaluasi awal menunjukkan bahwa model memiliki akurasi sebesar 93% pada data uji dan 87% pada data pelatihan, sehingga menunjukkan kinerja generalisasi yang baik.

Setelah mengetahui akurasi dari modelnya, tahap selanjutnya yaitu membuat Confusion Matrix. Confusion Matrix adalah sebuah tabel evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, khususnya pada masalah klasifikasi biner. Confusion Matrix dapat ditampilkan sebagai berikut.



Gambar 5. Confusion Matrix

Hasil evaluasi model menunjukkan bahwa terdapat 26 prediksi benar positif (TP) dan 17 prediksi benar negatif (TN). Sementara itu, kesalahan klasifikasi terdiri dari 2 False Positive (FP) dan 1 False Negative (FN). Hasil ini menunjukkan bahwa model memiliki kinerja yang baik dengan tingkat kesalahan yang rendah, meskipun tetap perlu diwaspadai terutama pada kesalahan tipe False Negative yang berisiko dalam konteks medis.

Tahap selanjutnya yaitu membuat Classification Report untuk mengetahui akurasi, *precision*, *recall* dan f1-*score*. *Classification report* dilakukan dengan ringkasan metrik evaluasi yang digunakan untuk menilai kinerja model klasifikasi, terutama dalam masalah klasifikasi biner atau multi-kelas [15]. Hasil *Classification Report* dapat direpresentasikan sebagai berikut.

Classification Report:

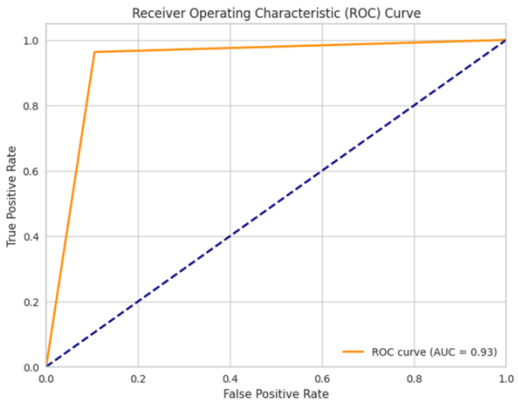
	precision	recall	f1-score	support
0	0.94	0.89	0.92	19
1	0.93	0.96	0.95	27
accuracy			0.93	46
macro avg	0.94	0.93	0.93	46
weighted avg	0.94	0.93	0.93	46

Gambar 6. Classification Report

Dari hasil Classification Report ringkasan evaluasi untuk masing-masing kelas dapat dijelaskan sebagai berikut.

a. Untuk kelas 1 (positif/sakit jantung), model memiliki precision = 0.93, recall = 0.96, dan F1-score = 0.95, yang menunjukkan bahwa model sangat andal dalam mendeteksi kasus penyakit jantung.
b. Untuk kelas 0 (negatif/sehat), precision = 0.94, recall = 0.89, dan F1-score = 0.92, yang juga menunjukkan performa baik meskipun sedikit lebih rendah dari kelas 1.

Tahap terakhir dalam evaluasi model adalah ROC-AUC. Evaluasi ini merupakan salah satu metrik evaluasi model klasifikasi, khususnya untuk model yang menghasilkan probabilitas. ROC (*Receiver Operating Characteristic*) adalah grafik AUC (*Area Under the Curve*) adalah angka yang menunjukkan luas area di bawah grafik ROC [16]. Hasil grafik ROC Curve yang telah dirancang dapat ditampil sebagai berikut.



Gambar 7. ROC Curve

Dari Gambar tersebut menunjukkan ROC Curve dengan nilai AUC sebesar 0.93. Grafik ini menggambarkan kemampuan model dalam membedakan kelas positif dan negatif pada berbagai *threshold*. Semakin dekat kurva ke sudut kiri atas, semakin tinggi akurasi model. Dengan AUC yang tinggi, dapat disimpulkan bahwa model memiliki performa klasifikasi yang sangat baik.

3.4. Membangun Sistem Prediksi

Membangun sistem prediksi adalah proses membuat sebuah aplikasi atau sistem cerdas yang bisa memprediksi hasil atau kejadian tertentu berdasarkan data input. Langkah – Langkah Membangun Sistem Prediksi dapat dilihat pada tabel berikut:

Tabel 1. Langkah - Langkah Membangun Sistem Prediksi

No	Tahapan	Penjelasan Singkat
1	Pengumpulan Data	Kumpulkan data historis (misal: data pasien, hasil pemeriksaan, dll)
2	Preprocessing Data	Bersihkan dan siapkan data (hilangkan *Outlier*, isi nilai kosong, normalisasi, dsb)
3	Pelatihan Model	Gunakan algoritma (misalnya: logistic regression) untuk belajar dari data
4	Evaluasi Model	Uji apakah model sudah cukup bagus (pakai akurasi, confusion matrix, dll)
5	Implementasi Sistem	Buat aplikasi (misalnya dengan Python, Streamlit, atau web) yang bisa menerima input dan memberi hasil prediksi.
6	Pengujian & Validasi	Pastikan sistem bekerja dengan benar sebelum digunakan secara nyata.

3.5. Deployment Ke Streamlit & Uji Coba Model

Deployment adalah proses mengubah model yang sudah selesai dilatih menjadi aplikasi nyata yang bisa digunakan oleh pengguna. Dalam penelitian ini, model prediksi penyakit jantung diimplementasikan menggunakan Streamlit, yaitu framework Python yang sederhana dan efisien untuk membangun aplikasi web interaktif. Proses umum deployment dapat dilihat pada tabel berikut

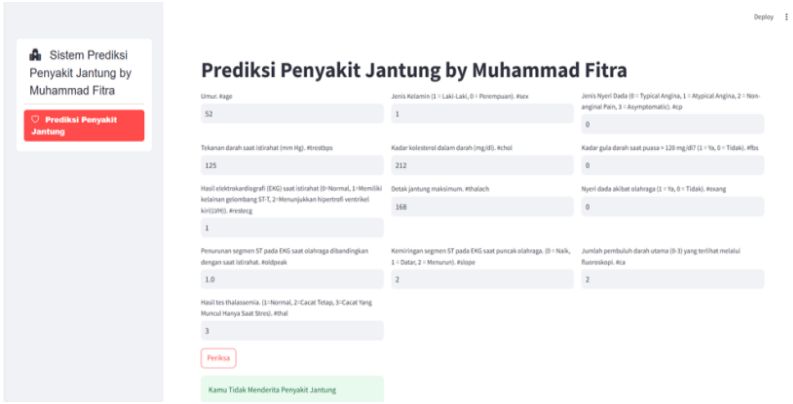
Tabel 2. Proses Umum Deployment

No	Tahap	Penjelasan
1	Latih Model	Buat model dan simpan ke file (misalnya: model.pkl)
2	Siapkan Interface	Buat tampilan input (contoh: form web, Streamlit, REST API)
3	Hubungkan Model ke Aplikasi	Aplikasi akan memuat model dan menghasilkan prediksi
4	Hosting/Publikasi	Pasang di server atau platform (contoh: Heroku, Render, HuggingFace, Streamlit Cloud)
5	Testing & Maintenance	Uji apakah model bekerja dan perbaiki jika ada masalah

Berdasarkan proses deploment diketahui streamlit merupakan alat yang membantu mengubah skrip Python menjadi aplikasi web hanya dalam beberapa baris kode. Untuk mendukung pemahaman, berikut di tampilankan streamlit project prediksi penyakit jantung.

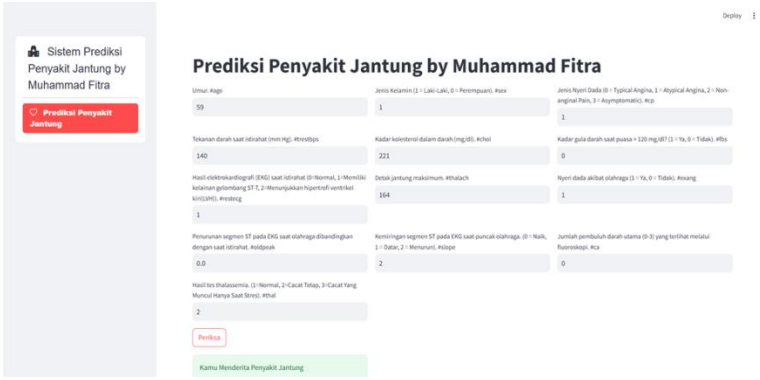
Gambar 9. Tampilan Streamlit Uji Coba Model

Tampilan halaman utama aplikasi prediksi penyakit jantung berbasis Streamlit. Pengguna cukup mengisi form berisi data medis seperti usia, tekanan darah, kolesterol, dll.



Gambar 10. Uji Coba Model Pada Data Orang Yang Tidak Ada Penyakit Jantung

Hasil prediksi menunjukkan bahwa pasien tidak menderita penyakit jantung (prediksi = sehat). Ini sesuai dengan kondisi aktual pasien.



Gambar 11. Uji Coba Model Pada Data Orang Yang Ada Penyakit Jantung

Hasil prediksi menunjukkan bahwa pasien menderita penyakit jantung (prediksi = positif). Ini membuktikan bahwa model mampu mengenali kasus dengan akurasi tinggi. Ketiga gambar di atas menunjukkan bahwa sistem tidak hanya memiliki antarmuka yang ramah pengguna, tetapi juga mampu memberikan prediksi yang tepat, baik untuk pasien yang

sehat maupun yang memiliki penyakit jantung. Ini menegaskan bahwa sistem siap digunakan sebagai alat bantu deteksi awal berbasis data.

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan algoritma Logistic Regression untuk memprediksi penyakit jantung, maka dapat disimpulkan bahwa Algoritma Logistic Regression berhasil diterapkan untuk memprediksi kemungkinan seseorang mengidap penyakit jantung berdasarkan 13 variabel input, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, detak jantung maksimal, dan hasil tes EKG. Secara keseluruhan, algoritma Logistic Regression terbukti efektif dalam membangun sistem prediksi penyakit jantung berbasis data. Meskipun demikian, model ini belum bisa digunakan sebagai alat diagnosa medis utama, melainkan sebagai sistem bantu (decision support) untuk mendukung pengambilan keputusan oleh tenaga medis. Faktor yang paling berpengaruh terhadap penyakit jantung berdasarkan dataset ini adalah *cp, thalach, slope, exang, oldpeak, ca,* dan *thal.* Faktor seperti *chol, fbs, restecg,* dan *trestbps* kurang berpengaruh, meskipun tetap bisa digunakan sebagai pelengkap informasi.

Daftar Pustaka

[1] World Health Organization, "Cardiovascular diseases (CVDs)," *World Health Organization*, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Kementerian Kesehatan Republik Indonesia, "Penyakit jantung penyebab utama kematian," *Kementerian Kesehatan Republik Indonesia*, 2019. https://p2ptm.kemkes.go.id/informasi-p2ptm/penyakit-jantung

[3] J. P. Husada, A. Handayani, and R. W. Nasution, "Hubungan Tingkat Pendidikan Dan Status Sosial Ekonomi Terhadap Tingkat Pengetahuan Tentang Faktor Risiko Penyakit Jantung Koroner Pada Pasien Penyakit Jantung Koroner," *J. Pandu Husada*, vol. 5, 2024, doi: https://jurnal.umsu.ac.id/index.php/JPH.

[4] Rumah Sakit PELNI, "Peran Penting Deteksi Dini dalam Menjaga Kesehatan Jantung," *Rumah Sakit PELNI*, 2022. https://www.rspelni.co.id/peran-penting-deteksi-dini-dalam-menjaga-kesehatan-jantung/

[5] D. Kurniawan Saputro, M. Fiko Rastio Ajie, S. Azizah, and D. Hartanti, "Penerapan Logistic Regression untuk Mendeteksi Penyakit Jantung pada Pasien," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2023, pp. 666–671.

[6] J. J. Pangaribuan, H. Tanjaya, and Kenichi, "Mendeteksi Penyakit Jantung Menggunakan Machine Learning Dengan Algoritma Logistic Regression," *J. Inf. Syst. Dev.*, vol. 6, no. 2, pp. 1–10, 2021.

[7] D. Sitanggang, V. Nicholas, N.; Wilson, A. R. A. Sinaga, and A. D. Simanjuntak, "Implementasi Data Mining untuk Memprediksi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor dan Logistic Regression," *J. TEKINKOM*, vol. 5, no. 2, pp. 493–501, 2022, doi: https://doi.org/10.37600/tekinkom.v5i2.698.

[8] Widiawati, L. Nurazizah, and I. R. Yunita, "Implementasi Algoritma Logistic Regression pada Pembuatan Website Sederhana untuk Prediksi Penyakit Jantung," *J. TEKINKOM*, vol. 15, no. 1, 2024.

[9] I. S. B. Azhar and W. K. Sari, "Penerapan Data Mining dan Teknologi Machine Learning pada Klasifikasi Penyakit Jantung," *JSI J. Sist. Inf.*, vol. 4, no. 1, pp. 2560–2568, 2022, doi: http://ejournal.unsri.ac.id/index.php/jsi/index.

[10] G. R. U. Asyafiiyah and R. M. Akbar, "Prediksi Pasien Terindikasi Penyakit Jantung Menggunakan Metode Logistic Regression," *SUBMIT J. Ilm. Teknol. Inf. dan Sains*, vol. 4, no. 1, pp. 19–23, 2024.

[11] M. Napiah and S. Heristian, "Perbandingan Algoritma Machine Learning pada Klasifikasi Penyakit Jantung," *J. Infortech*, vol. 6, no. 1, pp. 46–51, 2024.

[12] A. Y. Agusyul and F. Firmansyah, "Prediksi Penyakit Jantung Menggunakan Algoritma Random Forest," *J. Minfo Polgan*, vol. 12, no. 2, 2023, doi: https://doi.org/10.33395/jmp.v12i2.13214.

[13] W. Lestari and S. Sumarlinda, "Studi Komparatif Model Klasifikasi Kerentanan Penyakit Jantung Menggunakan Algoritma Machine Learning," *SATIN - Sains Dan Teknol. Inf.*, vol. 9, no. 1, pp. 107–115, 2023, doi: https://doi.org/10.33372/stn.v9i1.918.

[14] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 5, no. 1, pp. 97–100, 2023.

[15] G. Gunawan, S. A. Wibowo, and W. Andriani, "Evaluasi Model Deep Learning pada Pola Dataset Biomedis," *J. SAINTEKOM*, vol. 14, no. 2, pp. 195–207, 2024, doi: https://doi.org/10.33020/saintekom.v14i2.738.

[16] M. Fadli and R. A. Saputra, "Klasifikasi dan Evaluasi Performa Model Random Forest untuk Prediksi Stroke," *J. Tek.*, vol. 12, no. 2, pp. 72–80, 2023.