JURNAL SAINS DAN TEKNOLOGI
E-ISSN 2623-2294
Vol. 6, No.1, February 2026, pp. 29-41
https://journal.utsmakassar.ac.id/index.php/JST
OPEN ACCESS

Application of K-Nearest Neighbors Imputer (KNNI) and Random Forest Methods for Imputation and Prediction of Heart Disease

Penerapan Metode K-Nearest Neighbors Imputer (KNNI) dan Random Forest untuk Imputasi dan Prediksi Penyakit Jantung

Amandasari Dinda Rabbani [a,1,*], Deanita Nur Fauzizah [a,2], Alfan Rizaldy Pratama [a,3]
[a] Program Studi Sains Data, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
[1]22083010093@studentupnjatim.ac.id *; [2] 22083010005@studentupnjatim.ac.id; [3] alfan.fasilkom@upnjatim.ac.id
* corresponding author

ARTICLE INFO

Article history
Received : November 1, 2025
Revised : December 20, 2025
Accepted : January 20, 2026
Published : February 5, 2026

Kata Kunci: Penyakit Jantung; Imputasi Data; *KNN Imputer*; *Random Forest*; *Machine Learning*

Keywords: Heart disease; Data Imputation; KNN Imputer; Random Forest; Machine Learning

ABSTRAK/ABSTRACT

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga diperlukan metode prediksi yang akurat untuk mendukung deteksi dini. Pemanfaatan *machine learning* dalam bidang kesehatan menghadapi tantangan berupa keberadaan nilai hilang (*missing values*) pada data klinis yang dapat menurunkan performa model prediksi. Penelitian ini bertujuan untuk menerapkan metode *K-Nearest Neighbors Imputer* (KNNI) dalam menangani nilai hilang serta membangun model prediksi penyakit jantung menggunakan algoritma *Random Forest*. Dataset yang digunakan adalah *Heart Disease Dataset* dari *UCI Machine Learning Repository* yang terdiri dari 920 data pasien dengan 15 atribut. Proses penelitian meliputi analisis data awal, pra-pemrosesan, imputasi nilai hilang menggunakan KNNI dengan $k = 5$, pembagian data latih dan uji dengan rasio 80:20, serta pemodelan klasifikasi menggunakan *Random Forest* dengan 100 pohon keputusan. Evaluasi model dilakukan menggunakan metrik *accuracy, precision, recall*, *F1-score*, dan Area *Under the Curve* (AUC). Hasil penelitian menunjukkan bahwa model memperoleh nilai *accuracy* sebesar 59,78%, *F1-score* sebesar 56,22%, dan AUC sebesar 0,852, yang mengindikasikan kemampuan diskriminasi model yang baik dalam membedakan pasien berisiko dan tidak berisiko. Analisis *feature importance* menunjukkan bahwa variabel klinis seperti *thalach, oldpeak, age*, dan *ca* memiliki kontribusi signifikan dalam proses prediksi. Secara keseluruhan, kombinasi metode KNN *Imputer* dan *Random Forest* terbukti mampu menghasilkan model prediksi penyakit jantung yang cukup andal dan berpotensi dikembangkan lebih lanjut sebagai sistem pendukung keputusan medis.

Heart disease is one of the leading causes of death worldwide, making accurate prediction methods essential for early detection and prevention. The application of machine learning in healthcare is often challenged by the presence of missing values in clinical data, which can significantly reduce model performance. This study aims to apply the K-Nearest Neighbors Imputer (KNNI) to handle missing values and to develop a heart disease prediction model using the Random Forest algorithm. The dataset used is the Heart Disease Dataset from the UCI Machine Learning Repository, consisting of 920 patient records with 15 attributes. The research process includes initial data analysis, data preprocessing, missing value imputation using KNNI with k = 5, data splitting with an 80:20 ratio, and classification modeling using Random Forest with 100 decision trees. Model performance is evaluated using accuracy,

precision, recall, F1-score, and Area Under the Curve (AUC). The results show that the proposed model achieves an accuracy of 59.78%, an F1-score of 56.22%, and an AUC of 0.852, indicating good discriminative capability in distinguishing between high-risk and low-risk patients. Feature importance analysis reveals that clinical variables such as thalch, oldpeak, age, and ca play significant roles in the prediction process. Overall, the combination of KNN Imputer and Random Forest demonstrates promising potential as a baseline approach for medical decision support systems in heart disease prediction.

1. Pendahuluan

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia dan hingga saat ini masih menjadi permasalahan kesehatan global yang signifikan. *World Health Organization* (WHO) melaporkan bahwa penyakit kardiovaskular menyumbang proporsi terbesar terhadap angka mortalitas global setiap tahunnya [1]. Tingginya prevalensi penyakit jantung dipengaruhi oleh berbagai faktor risiko, seperti usia, tekanan darah, kadar kolesterol, riwayat penyakit, serta pola hidup yang tidak sehat. Faktor-faktor tersebut saling berinteraksi dan membentuk pola risiko yang kompleks sehingga menyulitkan proses diagnosis dini secara konvensional [2]. Seiring dengan pesatnya perkembangan teknologi informasi dan meningkatnya ketersediaan data kesehatan, penerapan teknik analisis berbasis data dan *machine learning* menjadi pendekatan yang menjanjikan dalam mendukung proses diagnosis penyakit jantung. Metode *machine learning* memiliki kemampuan untuk mengolah data dalam jumlah besar dan berdimensi tinggi serta mengidentifikasi pola nonlinier yang sulit dikenali melalui pendekatan statistik tradisional [3]. Berbagai penelitian menunjukkan bahwa algoritma klasifikasi seperti *Random Forest*, *Support Vector Machine*, dan *K-Nearest Neighbors* mampu memberikan performa prediksi yang kompetitif pada dataset penyakit jantung [4], [5]. Oleh karena itu, pemanfaatan *machine learning* di bidang kesehatan berpotensi meningkatkan akurasi prediksi penyakit jantung serta membantu tenaga medis dalam pengambilan keputusan klinis yang lebih objektif dan berbasis data.

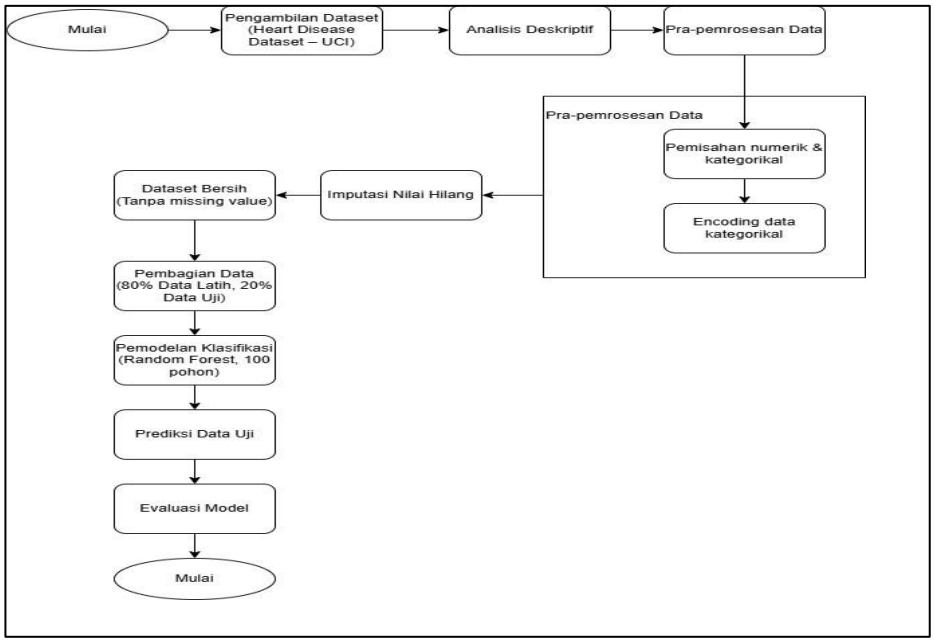
Namun demikian, pengembangan model prediksi berbasis data medis tidak terlepas dari permasalahan kualitas data. Salah satu permasalahan utama yang sering dijumpai adalah keberadaan nilai hilang (*missing values*), yang dapat disebabkan oleh keterbatasan alat pemeriksaan, ketidakkonsistenan prosedur pencatatan data, maupun kondisi klinis pasien yang tidak memungkinkan dilakukan pengukuran tertentu [6]. Keberadaan *missing values* dapat menurunkan performa model *machine learning*, memicu bias, serta mengurangi reliabilitas hasil prediksi apabila tidak ditangani dengan tepat [7]. Oleh karena itu, penanganan nilai hilang menjadi tahap krusial dalam proses *preprocessing* data medis sebelum dilakukan pemodelan. Berbagai teknik imputasi telah dikembangkan untuk mengatasi permasalahan nilai hilang, mulai dari metode sederhana seperti *mean imputation* dan *median imputation* hingga metode yang lebih kompleks berbasis *machine learning*. Metode imputasi sederhana cenderung mengabaikan hubungan antar variabel sehingga dapat mengurangi variasi alami data dan menurunkan kualitas informasi [8]. Sebaliknya, metode *K-Nearest Neighbors Imputation* (KNNI) menawarkan pendekatan yang lebih adaptif dengan mempertimbangkan kedekatan antar observasi dalam ruang fitur, sehingga nilai yang diimputasikan lebih representatif terhadap karakteristik data sebenarnya [9]. Beberapa penelitian terkini menunjukkan bahwa KNNI mampu mempertahankan struktur data, mengurangi distorsi distribusi fitur, serta memberikan performa yang lebih baik dibandingkan metode imputasi konvensional pada dataset kesehatan [6], [10].

Setelah kualitas data ditingkatkan melalui proses imputasi, pemilihan algoritma klasifikasi yang tepat menjadi faktor penting dalam membangun model prediksi yang andal. *Random Forest* merupakan salah satu algoritma *ensemble learning* yang banyak digunakan dalam penelitian medis

karena memiliki performa yang tinggi, stabilitas yang baik, serta ketahanan terhadap *overfitting* melalui mekanisme pembentukan banyak *decision tree* secara acak [11]. Selain itu, *Random Forest* mampu menangani hubungan nonlinier dan interaksi kompleks antar variabel, yang umum ditemukan pada data klinis penyakit jantung [12]. Keunggulan lain dari algoritma ini adalah kemampuannya dalam menyediakan informasi *feature importance*, yang dapat membantu mengidentifikasi faktor-faktor klinis paling berpengaruh terhadap risiko penyakit jantung, sehingga meningkatkan interpretabilitas model dalam konteks medis [13]. Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengembangkan model prediksi penyakit jantung dengan mengombinasikan metode *K- Nearest Neighbors Imputation* (KNNI) dalam penanganan nilai hilang dan algoritma *Random Forest* sebagai model klasifikasi utama. Diharapkan kombinasi kedua metode tersebut mampu menghasilkan model prediksi yang akurat, stabil, serta layak digunakan sebagai *decision support system* dalam mendukung deteksi dini penyakit jantung.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen berbasis *machine learning* yang disesuaikan dengan hasil dan pembahasan yang diperoleh. Tujuan utama penelitian adalah melakukan imputasi nilai hilang pada data klinis penyakit jantung menggunakan metode K-Nearest Neighbors Imputer (KNNI) serta membangun model prediksi menggunakan algoritma Random Forest. Alur penelitian meliputi analisis dataset, pra-pemrosesan data, imputasi nilai hilang, pemodelan klasifikasi, dan evaluasi performa model.



Gambar 1. Flowchart

2.1. Material

Material yang digunakan dalam penelitian ini berupa data sekunder dan parameter algoritma. Dataset yang digunakan adalah *Heart Disease Dataset* yang diperoleh dari UCI Machine Learning Repository. Dataset tersebut terdiri dari 920 data pasien dengan 15 atribut, yang mencakup variabel numerik, variabel kategorikal, serta satu variabel target (*num*) yang merepresentasikan tingkat keparahan penyakit jantung (kelas 0–4). Selain dataset, material penelitian juga mencakup parameter eksperimen, yaitu jumlah tetangga terdekat ($k = 5$) pada metode KNN Imputer serta jumlah pohon keputusan sebanyak 100 estimator pada algoritma Random Forest.

2.2. Instrumentasi

Instrumentasi yang digunakan dalam penelitian ini berupa perangkat lunak dan perangkat komputasi. Proses pengolahan dan analisis data dilakukan menggunakan bahasa pemrograman Python. Implementasi metode KNN Imputer dan Random Forest dilakukan dengan bantuan pustaka Scikit-learn. Pengolahan dan manipulasi data dilakukan menggunakan pustaka Pandas dan NumPy, sedangkan visualisasi hasil seperti *confusion matrix*, kurva ROC, dan *feature importance* dilakukan menggunakan Matplotlib. Seluruh eksperimen dijalankan pada perangkat komputer dengan sistem operasi Windows.

2.3. Prosedur

Prosedur penelitian diawali dengan pengambilan Heart Disease Dataset dari UCI Machine Learning Repository. Selanjutnya dilakukan analisis awal data untuk mengidentifikasi struktur dataset, jenis atribut, serta keberadaan nilai hilang. Hasil analisis menunjukkan bahwa beberapa fitur penting, terutama slope, ca, dan thal, memiliki jumlah nilai hilang yang cukup besar. Tahap berikutnya adalah pra-pemrosesan data yang meliputi pemisahan atribut numerik dan kategorikal serta proses encoding pada atribut kategorikal agar seluruh fitur berada dalam bentuk numerik. Setelah itu dilakukan imputasi nilai hilang menggunakan metode K-Nearest Neighbors Imputer (KNNI) dengan jumlah tetangga terdekat (k) sebesar 5 berdasarkan perhitungan jarak Euclidean. Dataset hasil imputasi kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20, menghasilkan 736 data latih dan 184 data uji. Pemodelan klasifikasi dilakukan menggunakan algoritma Random Forest dengan 100 pohon keputusan. Tahap akhir adalah evaluasi performa model menggunakan metrik accuracy, precision, recall, F1-score, dan Area Under the Curve (AUC), serta analisis confusion matrix, kurva ROC, dan feature importance.

3. Hasil dan Pembahasan

Penelitian ini bertujuan untuk menerapkan metode *K-Nearest Neighbors Imputer (KNNI)* sebagai teknik imputasi data hilang serta algoritma *Random Forest* sebagai model prediksi penyakit jantung berdasarkan dataset klinis. Seluruh data terlebih dahulu dianalisis struktur, distribusi, serta keberadaan *missing value* pada beberapa fitur. Proses imputasi dengan *KNNI* dilakukan menggunakan 5 tetangga terdekat, sehingga setiap nilai hilang digantikan dengan rata-rata nilai dari sampel terdekat berdasarkan jarak *Euclidean* pada ruang fitur.

Dataset yang telah terimputasi kemudian melalui proses *encoding* label untuk variabel kategorik, sebelum dipisahkan menjadi data pelatihan dan pengujian dengan rasio 80:20. Pemodelan dilakukan menggunakan algoritma *Random Forest* dengan jumlah estimator sebanyak 100 pohon keputusan. Model kemudian diuji menggunakan data uji untuk memperoleh hasil prediksi yang dibandingkan dengan label sebenarnya. Berikut hasil dari tahapan pada penulisan ini:

3.1. Analisis Missing Value dan Mengencoding Kolom Kategorikal Menjadi Numerik

Tabel 1. Output *Missing Value*

Jumlah *Missing Value* Tiap Kolom	
Nama Kolom	**Hasil *Missing Value***
id	0
Age	0
Sex	0
Dataset	0
Cp	0
Trestbps	59

Chol	30
Fbs	90
Restecg	2
Thalch	55
Exang	55
Oldpeak	62
Slope	309
Ca	611
Thal	486

Hasil analisis missing value menunjukkan bahwa beberapa fitur pada dataset memiliki jumlah nilai hilang yang bervariasi. Beberapa fitur seperti *trestbps*, *chol*, *fbs*, *thalch*, *exang*, dan *oldpeak* memiliki nilai hilang dalam jumlah sedang, sedangkan fitur *slope*, *ca*, dan *thal* memilik jumlah missing value yang sangat tinggi. Kondisi ini menunjukkan bahwa dataset tidak sepenuhnya lengkap dan memerlukan proses imputasi agar dapat digunakan dalam pemodelan. Untuk mengatasi hal tersebut digunakan metode *KNN Imputer*, karena metode ini mampu mengisi nilai hilang berdasarkan kemiripan pola antar sampel sehingga hasil imputasi lebih realistis dibandingkan metode sederhana seperti mean atau median.

Setelah seluruh nilai hilang berhasil diimputasi, tahap berikutnya adalah melakukan encoding pada fitur kategorik. *Encoding* diperlukan agar variabel non-numerik seperti *sex*, *cp*, *restecg*, *exang*, *slope*, *ca*, dan *thal* dapat diubah menjadi bentuk numerik yang dapat dipahami oleh algoritma pembelajaran mesin. Proses *encoding* memastikan bahwa seluruh fitur berada dalam format numerik sehingga dataset siap digunakan untuk tahap pemodelan selanjutnya dengan *Random Forest*.

Kolom Numerik : ['id', 'age', 'trestbps', 'chol', 'thalch', 'oldpeak', 'ca', 'num']

Kolom Kategorikal : ['sex', 'dataset', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'thal']

Tabel 2. Output *Encoding*

No	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	1	0	3	145.0	233.0	1	0	150.0	0	2.3	0	0.0	0	0
1	2	67	1	0	0	160.0	286.0	0	0	108.0	1	1.5	1	3.0	2	2
2	3	67	1	0	0	120.0	229.0	0	0	129.0	1	2.6	1	2.0	3	1
3	4	37	1	0	2	130.0	250.0	0	2	187.0	0	3.5	0	0.0	2	0
4	5	41	0	0	1	130.0	204.0	0	0	172.0	0	1.4	3	0.0	2	0

Berdasarkan struktur dataset, variabel dibagi menjadi dua jenis, yaitu kolom numerik dan kolom kategorikal. Kolom numerik terdiri dari *id*, *age*, *trestbps*, *chol*, *thalch*, *oldpeak*, *ca*, dan *num*, yang seluruhnya memiliki nilai kontinu atau diskrit yang dapat diolah secara langsung dalam bentuk angka. Sementara itu, kolom kategorikal mencakup *sex*, *dataset*, *cp*, *fbs*, *restecg*, *exang*, *slope*, dan

thal, yaitu fitur yang merepresentasikan kategori tertentu seperti jenis kelamin, tipe nyeri dada, hasil tes elektrokardiografi, dan kondisi thalium scan.

Data menunjukkan bahwa setiap baris merepresentasikan satu pasien dengan kombinasi nilai numerik dan kategorikal yang menggambarkan kondisi klinisnya. Variabel numerik seperti *age*, *chol*, dan *trestbps* menunjukkan karakteristik fisiologis pasien, sedangkan variabel kategorikal seperti *cp*, *restecg*, dan *exang* memberi informasi klinis tambahan terkait gejala dan hasil pemeriksaan jantung. Pembagian fitur menjadi numerik dan kategorikal ini penting untuk menentukan jenis *preprocessing* yang tepat, seperti *encoding* untuk kolom kategorikal dan normalisasi atau imputasi untuk kolom numerik, sebelum dataset digunakan dalam proses pemodelan menggunakan algoritma pembelajaran mesin.

3.2. Imputasi Data Menggunakan KNN Imputer

Tabel 3. Output *KNN Imputer*

No	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1.0	63.0	1.0	0.0	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	0.0	0.0
1	2.0	67.0	1.0	0.0	0.0	160.0	286.0	0.0	0.0	108.0	1.0	1.5	1.0	3.0	2.0	2.0
2	3.0	67.0	1.0	0.0	0.0	120.0	229.0	0.0	0.0	129.0	1.0	2.6	1.0	2.0	3.0	1.0
3	4.0	37.0	1.0	0.0	2.0	130.0	250.0	0.0	2.0	187.0	0.0	3.5	0.0	0.0	2.0	0.0
4	5.0	41.0	0.0	0.0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	3.0	0.0	2.0	0.0

Data yang telah melalui tahap imputasi dan encoding menunjukkan bahwa seluruh variabel kini berada dalam bentuk numerik, termasuk fitur kategorikal yang sebelumnya berbentuk label. Setiap baris mewakili satu sampel pasien dengan karakteristik klinis yang terdiri dari campuran informasi demografis, hasil pemeriksaan medis, serta variabel penentu risiko penyakit jantung. Kolom seperti *age*, *trestbps*, *chol*, *thalch*, dan *oldpeak* menggambarkan nilai fisiologis pasien, sedangkan kolom seperti *cp*, *restecg*, *exang*, *slope*, *ca*, dan *thal* telah berhasil dikonversi ke dalam bentuk angka hasil encoding sehingga dapat diproses oleh algoritma pembelajaran mesin. Seluruh nilai missing telah terisi, terlihat dari konsistensi format numerik tanpa adanya nilai kosong. Kondisi ini menunjukkan bahwa dataset telah dalam keadaan siap digunakan untuk tahap pemodelan, karena setiap fitur telah tersusun dalam format standar yang dibutuhkan oleh metode Random Forest maupun teknik analisis lanjutan lainnya. Dengan demikian, dataset yang semula memiliki nilai hilang dan campuran tipe data kini telah bersih, seragam, dan siap dipakai sebagai input dalam proses prediksi penyakit jantung.

3.3. Pemodelan Menggunakan Random Forest

Tabel 4. Data Latih dan Data Uji

Jenis Data	Jumlah Sampel
Data Latih	(736, 15)
Data Uji	(184, 15)

Pada tahap pemodelan, dataset yang telah melalui proses pembersihan, imputasi, dan encoding kemudian dibagi menjadi dua bagian, yaitu data latih dan data uji. Pembagian dilakukan menggunakan rasio umum 80:20 untuk memastikan bahwa model memperoleh cukup data untuk belajar sekaligus tetap memiliki data independen untuk evaluasi. Hasil pembagian menunjukkan bahwa data latih berjumlah 736 sampel dengan 15 fitur, sementara data uji terdiri dari 184 sampel

dengan jumlah fitur yang sama. Data latih digunakan untuk melatih model *Random Forest* agar mampu mengenali pola yang berkaitan dengan kondisi penyakit jantung, sedangkan data uji digunakan untuk menilai kemampuan generalisasi model terhadap data baru yang tidak pernah dilihat sebelumnya. Dengan komposisi pembagian ini, proses pemodelan dapat dilakukan secara optimal dan menghasilkan evaluasi performa yang lebih objektif.

3.4. Evaluasi Model

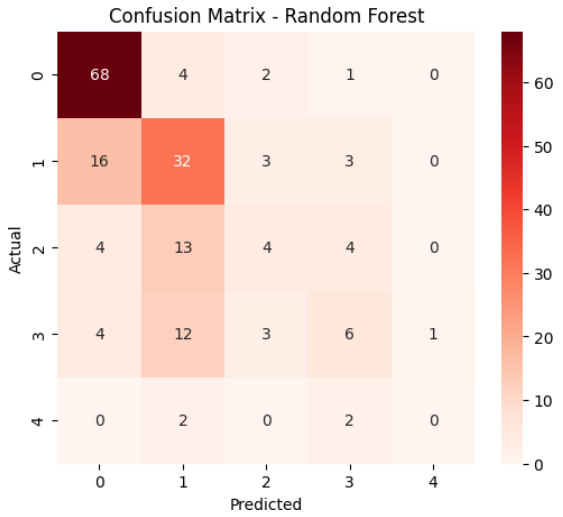
Tabel 5. Hasil Evaluasi Model

Hasil Evaluasi Model
Accuracy : 0.5978
Precision : 0.5486
Recall : 0.5978
F1 Score : 0.5622
AUC : 0.8520

Hasil evaluasi model *Random Forest* menunjukkan bahwa performa model dalam memprediksi penyakit jantung berada pada tingkat sedang. Nilai *accuracy* sebesar 0.5978 mengindikasikan bahwa sekitar 59,78% sampel pada data uji berhasil diprediksi dengan benar oleh model. Nilai *precision* sebesar 0.5486 menunjukkan bahwa dari seluruh prediksi positif yang dibuat model, sekitar 54,86% benar-benar merupakan kasus positif, menandakan bahwa model masih menghasilkan sejumlah *false positive*. Sementara itu, nilai *recall* sebesar 0.5978 menunjukkan kemampuan model dalam mendeteksi kasus positif, yaitu berhasil menemukan sekitar 59,78% dari total kasus yang sebenarnya positif. Nilai *F1-score* sebesar 0.5622 menjadi indikator bahwa keseimbangan antara *precision* dan *recall* masih berada pada kategori cukup, meskipun belum optimal.

Meskipun akurasi dan *F1-score* tidak terlalu tinggi, nilai *AUC* sebesar 0.8520 memberikan gambaran berbeda. *AUC* yang berada di atas 0.85 menunjukkan bahwa kemampuan model dalam membedakan kelas positif dan negatif berada pada kategori baik, serta model memiliki kemampuan diskriminasi yang kuat meskipun prediksi biner akhirnya belum optimal. Perbedaan ini menunjukkan bahwa model sebenarnya mampu mengenali pola dasar antara pasien berisiko dan tidak berisiko, namun proses penentuan ambang batas (*threshold*) atau distribusi kelas yang tidak seimbang kemungkinan mempengaruhi skor evaluasi lainnya. Secara keseluruhan, performa model cukup baik pada aspek pemisahan kelas, namun masih perlu peningkatan pada aspek akurasi prediksi akhir.

3.5. Confusion Matrix, ROC Curve, dan Top 10 Feature Importance

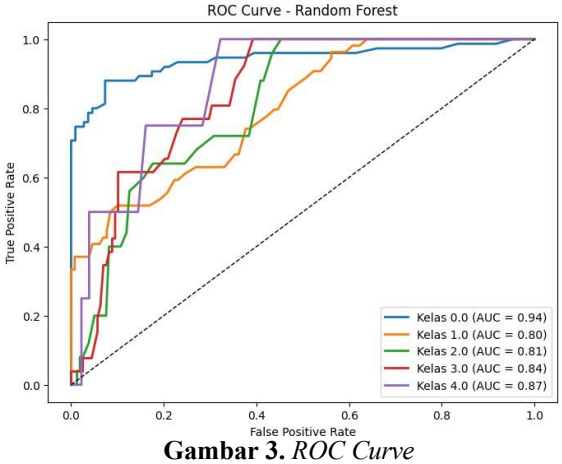


Gambar 2. *Confusion Matrix*

Confusion matrix menunjukkan bahwa model *Random Forest* lebih baik dalam memprediksi kelas 0, terlihat dari tingginya jumlah prediksi benar pada kelas tersebut yaitu sebanyak 68 sampel. Hal ini mengindikasikan bahwa model mampu mengenali kasus tanpa penyakit jantung dengan cukup baik. Namun, kemampuan model menurun pada kelas lainnya. Pada kelas 1, model berhasil memprediksi dengan benar sebanyak 32 sampel, tetapi masih terjadi kesalahan prediksi yang cukup besar, terutama terhadap kelas 0 sebanyak 16 sampel, yang berarti banyak kasus berisiko justru diklasifikasikan sebagai tidak berisiko (*false negative*).

Untuk kelas 2, model hanya memprediksi benar sebanyak 13 sampel, sementara sisanya tersebar salah ke kelas lain, menunjukkan bahwa kelas ini sulit dibedakan karena kemiripan pola dengan kelas lain. Kesalahan serupa juga ditemukan pada kelas 3, dengan hanya 6 prediksi benar dan banyaknya prediksi yang bergeser ke kelas 1 dan 2. Kelas 4 menjadi kelas yang paling jarang terdeteksi, dengan prediksi benar yang sangat sedikit, yang kemungkinan disebabkan oleh sedikitnya sampel kelas tersebut dalam dataset atau pola fitur yang kurang jelas.

Secara keseluruhan, *confusion matrix* menunjukkan bahwa model memiliki performa yang baik dalam menangani kelas mayoritas namun kesulitan dalam mengidentifikasi kelas yang lebih kecil atau memiliki pola yang mirip, yang berdampak pada turunnya nilai *precision* dan *recall* di beberapa kelas. Hal ini mengindikasikan perlunya perbaikan, misalnya melalui penyeimbangan data atau tuning parameter model. Berikut adalah gambar dari *ROC Curve.*

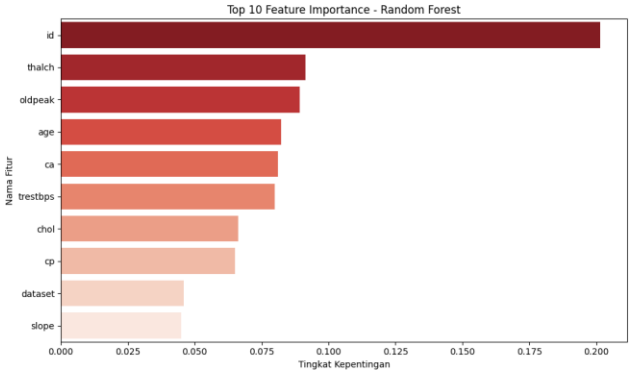


Gambar 3. *ROC Curve*

Grafik *ROC Curve* menunjukkan kemampuan model *Random Forest* dalam membedakan masing-masing kelas pada prediksi penyakit jantung. Setiap kelas memiliki nilai *AUC* yang berbeda, mencerminkan tingkat performa model pada tiap kategori risiko. Kelas 0 menunjukkan performa terbaik dengan nilai *AUC* 0.94, menandakan bahwa model sangat efektif dalam memisahkan kelas ini dari kelas lainnya. Hal ini sejalan dengan hasil *confusion matrix* sebelumnya yang juga menunjukkan akurasi tinggi pada kelas 0.

Untuk kelas 1, nilai *AUC* sebesar 0.80 mengindikasikan performa yang cukup baik, meskipun masih terdapat beberapa kesalahan prediksi. Kelas 2 dan 3 masing-masing memiliki nilai *AUC* 0.81 dan 0.88, menunjukkan bahwa kemampuan model dalam mengidentifikasi kedua kelas ini berada pada kategori baik, meskipun tidak sekuat kelas 0. Sementara itu, kelas 4 memiliki nilai *AUC* 0.87, menunjukkan bahwa meskipun jumlah sampelnya sedikit, model masih mampu membedakan kelas ini dengan cukup baik.

Secara keseluruhan, *ROC Curve* menegaskan bahwa kemampuan diskriminasi model *Random Forest* cukup solid, terutama pada kelas mayoritas. Nilai *AUC* yang berada pada rentang 0.80-0.94 menunjukkan bahwa model memiliki kemampuan klasifikasi yang baik meskipun akurasi keseluruhan tidak terlalu tinggi. Hal ini menunjukkan bahwa model sebenarnya mampu mengenali pola dasar pada data, namun perlu diperbaiki pada aspek *threshold* atau keseimbangan kelas agar prediksi final lebih akurat dan konsisten. Berikut adalah gambar dari Top 10 *Feature Importance.*



Gambar 4. Top 10 *Feature Importance*

Tabel 6. Nilai *Importance* dari 10 fitur terpenting

No	Fitur	Importance
0	id	0.201460
9	thalch	0.091284
11	oldpeak	0.089130
1	age	0.082378
13	ca	0.081101
5	trestbps	0.079959
6	chol	0.066280
4	cp	0.064942
3	dataset	0.045795
12	slope	0.044909

Hasil analisis *feature importance* dari model *Random Forest* menunjukkan bahwa fitur id memiliki nilai kepentingan tertinggi sebesar 0.201, jauh lebih besar dibanding fitur lainnya. Hal ini mengindikasikan adanya kemungkinan bahwa fitur *id* tidak hanya berfungsi sebagai penanda unik, tetapi juga mengandung pola berurutan atau informasi tidak sengaja terkait dengan label, sehingga perlu diperiksa lebih lanjut karena dapat menyebabkan *data leakage*. Setelah fitur *id*, fitur *thalch*, *oldpeak*, *age*, dan *ca* muncul sebagai kontributor penting dalam proses prediksi. Fitur-fitur ini memiliki relevansi klinis yang kuat, misalnya *thalch* (detak jantung maksimum), *oldpeak* (depresi ST saat uji stres), serta jumlah pembuluh darah yang mengalami penyempitan (*ca*).

Fitur fisiologis lainnya seperti *trestbps* dan *chol* juga memberikan kontribusi signifikan, meskipun tidak sebesar fitur utama. Variabel kategorikal seperti *cp* (jenis nyeri dada) serta *dataset* dan *slope* memiliki nilai kepentingan lebih rendah, namun tetap berperan dalam membentuk keputusan model. Secara keseluruhan, grafik ini menunjukkan bahwa model banyak bergantung pada kombinasi variabel klinis utama yang memang relevan dengan kondisi kardiovaskular, namun dominasi fitur *id* menunjukkan bahwa proses pra-pemrosesan perlu diperbaiki agar model tidak mempelajari pola yang bersifat non-klinis dan tidak valid secara medis.

3.6. Penggunaan Rumus dan Persamaan

a. Imputasi KNN

$$x^i = {}_1 \sum_{j \in K(i)} x_j \qquad (1)$$

Keterangan:

Rumus ini digunakan untuk menghitung nilai imputasi ($\hat{x}_i$) dari sebuah data yang hilang dengan menggunakan rata-rata nilai dari k tetangga terdekat. Tetangga ditentukan berdasarkan jarak Euclidean dalam ruang fitur. Dengan demikian, nilai yang diimputasikan bukan ditentukan oleh rata-rata global, tetapi oleh sampel yang memiliki karakteristik paling mirip. Metode ini sangat berguna untuk dataset medis karena mampu mempertahankan pola dan variasi alami dalam data[16].

b. Random Forest (Voting Ensemble)

$$\hat{y} = \text{mode}(h_1(x),\ h_2(x),\ \dots,\ h_T(x)) \quad (2)$$

c. Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Keterangan :

Akurasi menunjukkan proporsi prediksi benar (positif dan negatif) dari keseluruhan sampel. Ini merupakan metrik umum yang digunakan untuk evaluasi model, meskipun sensitif terhadap ketidakseimbangan kelas. Pada prediksi penyakit jantung, akurasi harus dilihat bersama metrik lain untuk memperoleh gambaran yang lebih lengkap[18].

d. Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Keterangan :

Precision menggambarkan seberapa banyak prediksi positif yang benar-benar positif. Nilai ini penting untuk konteks medis karena menunjukkan seberapa sering model salah memberi label "berisiko" pada pasien sehat[18].

e. Recall

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Keterangan :

Recall menunjukkan seberapa baik model mendeteksi kasus yang benar-benar positif (pasien yang benar-benar berisiko). Pada prediksi penyakit jantung, recall sangat penting karena kesalahan (false negative) dapat berdampak pada kegagalan mendeteksi masalah serius[18].

f. F1-Score

$$F1 = \frac{T2.Precision.Recall}{Precision + Recall} \quad (6)$$

Keterangan :

F1-score menggabungkan *precision* dan *recall* dalam satu nilai. Rumus ini baik digunakan ketika data tidak seimbang, seperti kasus kesehatan yang biasanya memiliki lebih banyak kelas tertentu dibanding yang lain. F1 memastikan bahwa model tidak hanya presisi, tetapi juga sensitif[18].

g. AUC (Area Under the ROC Curve)

$$AUC = \int_0^1 TPR\ (FPR)\ d\ (FPR) \quad (7)$$

Keterangan:

AUC menggambarkan kemampuan model membedakan kelas positif dan negatif pada berbagai ambang batas (*threshold*). Semakin tinggi nilai *AUC*, semakin baik kemampuan model membedakan pasien berisiko dan tidak berisiko. Dalam pemodelan medis, *AUC* sangat penting karena tidak bergantung pada distribusi kelas[19].

h. Jarak *Euclidean*

$$d\ (x_i, x_j) = \sum_{k=1}^{p} (\ x_{ik} - \ x_{jk}\ )^2 \quad (8)$$

Keterangan

Rumus ini menentukan jarak antara dua sampel data dalam ruang fitur. Jarak *Euclidean* dipakai untuk mengukur seberapa mirip dua sampel. Semakin kecil jarak antar sampel, semakin dekat (semakin mirip) kedua data tersebut, sehingga lebih relevan dijadikan tetangga untuk melakukan imputasi. Hal ini membuat imputasi lebih akurat karena mempertimbangkan struktur lokal data[16].

4. Kesimpulan

Penelitian ini menunjukkan bahwa penerapan metode *K-Nearest Neighbors Imputer* (*KNNI*) dan algoritma *Random Forest* dapat digunakan secara efektif dalam proses imputasi dan prediksi penyakit jantung. Hasil analisis *missing value* memperlihatkan bahwa dataset memiliki sejumlah nilai hilang yang cukup tinggi pada beberapa fitur, terutama *slope*, *ca*, dan *thal*, sehingga proses imputasi menjadi langkah penting sebelum pemodelan. Penggunaan *KNN Imputer* berhasil mengisi seluruh nilai hilang secara konsisten dengan mempertahankan karakteristik lokal antar sampel, sehingga menghasilkan dataset yang bersih dan siap digunakan dalam tahap analisis berikutnya. Tahap *encoding* pada variabel kategorikal menghasilkan representasi numerik yang sesuai dengan kebutuhan algoritma pembelajaran mesin, dan pembagian data dengan rasio 80:20 memastikan model memperoleh data latih yang mencukupi sekaligus data uji yang objektif. Evaluasi model *Random Forest* menunjukkan bahwa meskipun nilai akurasi dan *F1-score* berada pada tingkat sedang, nilai *AUC* yang tinggi (0.8520) memberikan indikasi kuat bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan pasien berisiko dan tidak berisiko. Analisis *confusion matrix* dan *ROC Curve* juga memperlihatkan bahwa performa model paling kuat terdapat pada kelas mayoritas, sementara kelas minoritas memerlukan penanganan lebih lanjut, seperti penyeimbangan data atau tuning parameter.

Hasil *feature importance* mengungkapkan bahwa beberapa fitur klinis seperti *thalch*, *oldpeak*, *age*, dan *ca* berperan signifikan dalam proses prediksi, meskipun ditemukannya dominasi fitur *id* menunjukkan perlu adanya evaluasi ulang terhadap proses pra-pemrosesan agar menghindari potensi *data leakage*. Secara keseluruhan, penelitian ini membuktikan bahwa kombinasi *KNN Imputer* dan *Random Forest* merupakan pendekatan yang dapat digunakan sebagai *baseline* untuk sistem pendukung keputusan medis dalam prediksi penyakit jantung, dengan potensi peningkatan performa melalui optimasi lanjutan di masa mendatang.

5. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada semua pihak yang telah membantu terselesaikannya penelitian ini, khususnya kepada dosen pembimbing, instansi terkait yang menyediakan data, serta keluarga dan rekan-rekan yang memberikan dukungan moral maupun bantuan dalam proses analisis. Semoga segala bantuan yang diberikan mendapat balasan kebaikan.

6. Referensi

[1] World Health Organization, *Cardiovascular Diseases (CVDs) Fact Sheet*, Geneva: WHO, 2023.

[2] A. Handayani *et al*., "Predictive Analysis Heart Disease Based on Machine Learning Using the Random Forest Algorithm," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 3, pp. 1980–1986, 2025.

[3] N. Nasution *et al*., "Predicting Heart Disease Using Machine Learning," *IT Journal Research and Development*, vol. 9, no. 2, pp. 140–150, 2025.

[4] Hirmayanti and E. Utami, "Enhanced Heart Disease Diagnosis Using Machine Learning Algorithms," *Jurnal RESTI*, vol. 9, no. 2, pp. 385–392, 2025.

[5] A. Miftakhudin *et al*., "Komparasi Algoritma KNN dan Random Forest untuk Diagnosa Penyakit Jantung Koroner," *RIGGS Journal*, vol. 4, no. 3, pp. 2962–2971, 2025.

[6] Y. Yilmaz *et al*., "A Comparative Study of Imputation Techniques for Missing Values in Healthcare Diagnostic Datasets," *International Journal of Data Science and Analytics*, 2025.

[7] N. H. Alfajr and S. Defiyanti, "Prediksi Penyakit Jantung Menggunakan Metode Random Forest," *Jurnal Informatika dan Teknik Elektro Terapan*, 2025.

[8] J. Brown *et al*., "Impact of Simple Imputation Methods on Medical Data Analysis,"*Healthcare Analytics*, 2021.

[9] Y. Yilmaz *et al*., "KNN-Based Imputation for Medical Datasets," *International Journal of Data Science and Analytics*, 2025.

[10] A. Silva *et al*., "Machine Learning-Based Imputation Methods in Healthcare," *IEEE Access*, 2022.

[11] L. Breiman, "Random Forests," *Machine Learning*, revisited in healthcare applications, 2021

[12] S. Defiyanti *et al*., "Random Forest for Nonlinear Medical Data," *Journal of Medical Informatics*, 2023.

[13] A. Firdausiyah *et al*., "Feature Importance Analysis Using Random Forest in Heart Disease Prediction," *JAIEA*, 2025.

[14] L. Oluwaseye, J. Wesley, D. Babu, and S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *Int. J. Data Sci. Anal.*, vol. 20, no. 7, pp. 6357–6373, 2025.

[15] Random forest, Wikipedia: The Free Encyclopedia. Available: https://en.wikipedia.org/wiki/Random_forest. Accessed Dec. 11, 2025.

[16] E. Sahelvi, P. Cikita, and R. M. Sapitri, "Comparison of K-Nearest Neighbors and Random Forest Algorithms for Recommendations for a Healthy Lifestyle in Prevent Heart Disease Perbandingan Algoritma K-Nearest Neighbors dan Random Forest untuk Rekomendasi Gaya Hidup Sehat dalam Mencegah Penyakit Jan," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. July, pp. 830–840, 2025.

[17] Partial Area Under the ROC Curve, Wikipedia: The Free Encyclopedia. Available: https://en.wikipedia.org/wiki/Partial_Area_Under_the_ROC_Curve. Accessed Dec. 11, 2025.