



Komparasi Algoritma KNN dan Random Forest untuk Diagnosa Penyakit Jantung Koroner (Studi Kasus: RSUD Dr. Soeselo Slawi)

Ahmad Miftakhuudin¹, Nugroho Adhi Santoso², Bayu Aji Santoso³

¹Teknik Informatika, STMIK YMI Tegal

^{2,3}Sistem Informasi, STMIK YMI Tegal

ahmadmifta66@gmail.com, nugrohoadhisantoso@stmik-tegal.ac.id, bayu@stmik-tegal.ac.id

Abstrak

Penyakit jantung koroner (PJK) merupakan salah satu penyebab utama kematian di Indonesia, sehingga deteksi dini menjadi aspek krusial dalam upaya mencegah komplikasi dan meningkatkan efektivitas intervensi medis. Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap dua algoritma machine learning, yaitu K-Nearest Neighbors (KNN) dan Random Forest, dalam klasifikasi diagnosis penyakit jantung koroner berdasarkan data rekam medis pasien RSUD Dr. Soeselo Slawi. Metodologi penelitian meliputi tahapan preprocessing data, transformasi fitur kategorikal, normalisasi, pembagian data dengan rasio 80:20, serta pelatihan dan evaluasi model menggunakan metrik akurasi, precision, recall, dan F1-score. Hasil evaluasi menunjukkan bahwa algoritma Random Forest memiliki performa lebih unggul dengan akurasi sebesar 84.00%, precision 84.02%, recall 84.00%, dan F1-score 83.97%. Sementara K-Nearest Neighbors (KNN) memperoleh akurasi 82.00% dengan nilai evaluasi yang sedikit lebih rendah. Pengujian tambahan menggunakan K-Fold Cross Validation menunjukkan bahwa Random Forest lebih stabil terhadap variasi data, dengan akurasi tertinggi 84.80% pada fold ke-4. Penelitian ini juga melakukan pembagian data dengan rasio (60:40, 70:30, dan 90:10). Berdasarkan hasil tersebut, Random Forest direkomendasikan sebagai algoritma terbaik untuk implementasi sistem pendukung keputusan medis dalam diagnosis penyakit jantung koroner. Penelitian ini diharapkan dapat berkontribusi dalam pengembangan teknologi kesehatan berbasis kecerdasan buatan yang lebih akurat, efisien, dan dapat diandalkan di lingkungan rumah sakit.

Kata kunci: Penyakit Jantung Koroner, Machine Learning, K-Nearest Neighbors (KNN), Random Forest, Diagnosis Medis

1. Latar Belakang

Penyakit jantung koroner (PJK) adalah salah satu penyebab utama kematian secara global, dan dampaknya sangat terasa di Indonesia. Penyakit ini secara konsisten menyumbang angka kematian tertinggi, menjadikannya masalah kesehatan masyarakat yang serius dan mendesak untuk ditangani. Berdasarkan studi Global Burden of Disease (GBD), Kematian akibat penyakit jantung dan pembuluh darah di Indonesia mengalami peningkatan dari 292.000 kasus pada tahun 1990 menjadi lebih dari 659.000 kasus pada tahun 2019 [1]. Penyakit Jantung Koroner sebagai bentuk utama dari penyakit kardiovaskular, memiliki prevalensi yang tinggi dan menjadi tantangan serius bagi sistem pelayanan kesehatan nasional. Faktor risiko seperti hipertensi, diabetes, dislipidemia, dan gaya hidup sedentari turut memperburuk tren ini [2]. Selain itu, keterlambatan dalam deteksi dini dan pengobatan optimal turut menyebabkan angka mortalitas dan morbiditas tetap tinggi. Oleh karena itu, penting untuk mengembangkan solusi yang dapat meningkatkan efektivitas diagnosis dan intervensi terhadap Penyakit jantung koroner.

RSUD Dr. Soeselo Slawi sebagai rumah sakit rujukan regional menghadapi tantangan dalam melakukan diagnosis penyakit jantung koroner secara cepat dan akurat. Metode konvensional yang bergantung pada interpretasi klinis dan pengalaman dokter sering kali memerlukan waktu yang lama dan berisiko menimbulkan subjektivitas. Rumah sakit ini menangani banyak pasien dengan beragam gejala dan kondisi komorbid, sehingga membutuhkan alat bantu diagnosis yang dapat mempercepat proses pengambilan keputusan. Data rekam medis pasien yang tersimpan di RSUD memiliki potensi besar untuk dimanfaatkan dalam pengembangan sistem pendukung keputusan berbasis teknologi. Oleh karena itu, dibutuhkan pendekatan berbasis teknologi yang mampu mendukung proses diagnosis secara objektif dan efisien [3]. Selain itu, metode ini mampu mengolah data medis tabular secara efisien dan memberikan output prediktif yang berguna bagi tenaga medis. Penggunaan machine learning juga dapat mengurangi kesalahan diagnosis dan memperkuat akuntabilitas keputusan klinis. Oleh karena itu, pemanfaatan algoritma yang tepat sangat penting dalam mempercepat dan meningkatkan ketepatan diagnosis penyakit jantung koroner.

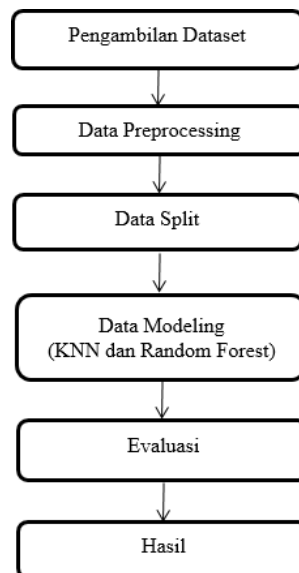
Perkembangan teknologi machine learning memberikan peluang untuk meningkatkan akurasi diagnosis medis [4]. Algoritma machine learning dapat menganalisis pola kompleks dalam data medis yang sulit diidentifikasi secara manual, serta memberikan prediksi berbasis data historis pasien. Dalam konteks diagnosis penyakit jantung koroner, Algoritma K-Nearest Neighbors (KNN) serta Random Forest menjadi pilihan yang relevan karena telah terbukti efektif dalam klasifikasi penyakit berbasis data klinis [5]. Selain itu, metode ini mampu mengolah data medis tabular secara efisien dan memberikan output prediktif yang berguna bagi tenaga medis. Penggunaan machine learning juga dapat mengurangi kesalahan diagnosis dan memperkuat akuntabilitas keputusan klinis. Oleh karena itu, pemanfaatan algoritma yang tepat sangat penting dalam mempercepat dan meningkatkan ketepatan diagnosis penyakit jantung koroner.

Studi terbaru menunjukkan bahwa Random Forest sangat baik dalam mengelola data medis yang rumit menghasilkan akurasi tinggi dalam diagnosis penyakit [6]. Sementara itu, KNN dikenal sebagai algoritma yang sederhana namun cukup efektif dalam klasifikasi berbasis kemiripan antar data pasien [7]. Kedua algoritma ini telah banyak diterapkan dalam studi diagnosis penyakit, termasuk diabetes, hipertensi, dan gangguan kardiovaskular lainnya. Penelitian yang dilakukan Susi Eva Maria Purba bahwa Random Forest mencapai akurasi 100% dalam klasifikasi obat, sementara KNN hanya mencapai 70% [8]. Selain itu, menurut Shivahare et al., kedua algoritma efektif dalam diagnosis berbasis citra medis, namun Random Forest lebih tahan terhadap variasi data [9].

Oleh karena itu, penelitian ini bertujuan untuk membandingkan performa K-Nearest Neighbors (KNN) dan Random Forest, sebagai dua algoritma machine learning, dalam klasifikasi diagnosis penyakit jantung koroner [10]. Kedua algoritma dipilih karena mampu menangani data medis tabular dengan banyak variabel, seperti yang umum ditemukan dalam rekam medis pasien. Berdasarkan studi terkini, KNN dan Random Forest menunjukkan performa tinggi dalam klasifikasi penyakit. Berdasarkan keunggulan tersebut, karena algoritma KNN dan Random Forest dinilai sangat menjanjikan untuk aplikasi di sistem pendukung keputusan klinis, penelitian ini dirancang untuk mengidentifikasi algoritma mana yang memberikan hasil klasifikasi paling akurat.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan tujuan untuk membandingkan performa dua algoritma machine learning, yaitu K-Nearest Neighbors (KNN) dan Random Forest, dalam klasifikasi diagnosis penyakit jantung koroner [11]. Proses penelitian dilakukan secara sistematis melalui beberapa tahapan utama, tahapan penelitian dapat dilihat pada gambar 1.



Gambar 1. Tahapan Penelitian

Gambar 1. Tahapan penelitian ini dimulai dengan pengambilan dataset dari rekam medis pasien di RSUD Dr. Soeselo Slawi. Setelah data terkumpul, langkah selanjutnya adalah preprocessing untuk membersihkan dan menyiapkan data, termasuk penanganan nilai kosong, encoding data kategorikal, normalisasi fitur numerik, serta deteksi dan penanganan outlier. Data yang telah diproses kemudian dibagi menjadi data latih dan data uji dengan

rasio 80:20, guna memastikan model dapat diuji secara objektif. Tahap selanjutnya adalah modeling, di mana dua algoritma machine learning K-Nearest Neighbors (KNN) dan Random Forest diimplementasikan dan dioptimalkan untuk klasifikasi diagnosis penyakit jantung koroner. Model yang telah dibangun kemudian dievaluasi menggunakan metrik klasifikasi seperti accuracy, precision, recall, F1-score, serta confusion matrix dan teknik K-Fold Cross Validation untuk mengukur kestabilan performa [12]. Hasil akhir dari seluruh proses ini disajikan dalam bentuk grafik, tabel, dan narasi akademik yang menggambarkan efektivitas masing-masing algoritma dalam mendeteksi penyakit jantung koroner berdasarkan data yang tersedia.

2.1. Pengambilan Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari rekam medis elektronik RSUD Dr. Soeselo Slawi dengan rentan waktu tertentu. Dataset terdiri dari 500 data pasien dengan 10 fitur dan dua kelas target yaitu fitur diagnosa. Penjelasan mengenai fitur tersebut tercantum pada Tabel 1.

Table 1. Fitur Data

	Feature	Tipe Data	Keterangan
X1	Umur	Numerik	Usia Pasien dalam tahun
X2	Jenis Kelamin	Kategorik	Jenis Kelamin Pasien (M/F)
X3	Riwayat Keluarga	Kategorik	Status riwayat penyakit jantung dalam keluarga
X4	Chest Pain Type (cp)	Numerik	Tipe nyeri dada
X5	Trestbps (mmHg)	Numerik	Tekanan darah saat istirahat (mmHg)
X6	Chol (mg/dl)	Numerik	Kadar kolesterol total
X7	Fasting Blood Sugar (fbs)	Numerik	Gula darah saat puasa (mg/dL)
X8	BMI	Numerik	Indeks massa tubuh
X9	Aktivitas Fisik	Kategorik	Tingkat aktivitas pasien
X10	Merokok	Kategorik	Status merokok
Y	Diagnosis	Kategorik	Hasil diagnosis penyakit jantung koroner

2.2. Data Preprocessing

Preprocessing dilakukan untuk menyiapkan data diagnosis penyakit jantung koroner agar siap digunakan dalam klasifikasi. Langkah awal meliputi pembersihan data (penghapusan kolom ID dan penanganan missing values), diikuti transformasi fitur kategorikal menggunakan label encoding. Analisis korelasi Pearson digunakan untuk mengidentifikasi hubungan antar fitur dan menghindari multikolinearitas. Fitur numerik dinormalisasi dengan Min-Max Scaler agar skala data seragam. Seluruh proses diimplementasikan menggunakan Python (pandas, scikit-learn) dalam Jupyter Notebook, menghasilkan dataset yang bersih dan siap untuk modeling dengan algoritma KNN dan Random Forest.

2.3. Pembagian Data (Data Split)

Setelah melalui tahap preprocessing, dataset dibagi menjadi dua subset utama: data latih (training set) dan data uji (testing set). Tujuan dari pembagian ini adalah untuk melatih model pada sebagian data, lalu menguji performanya pada data yang belum pernah dilihat sebelumnya. Dalam penelitian ini, pembagian awal dilakukan dengan rasio 80:20, di mana 80% data digunakan untuk pelatihan dan 20% sisanya untuk pengujian. Dengan total 500 entri, sebanyak 400 data digunakan untuk melatih model, dan 100 data digunakan untuk evaluasi.

Selain rasio 80:20, dilakukan eksperimen tambahan dengan variasi rasio pembagian data lainnya, yaitu 60:40, 70:30, dan 90:10. Tujuannya adalah untuk mengamati pengaruh proporsi data terhadap performa model klasifikasi. Pembagian data dilakukan secara acak menggunakan fungsi `train_test_split` dari pustaka scikit-learn dalam Python, dengan parameter `random_state` untuk memastikan reproduktibilitas hasil.

2.4. Pemodelan Data (Data Modeling)

Tahapan pemodelan data bertujuan untuk membangun model klasifikasi diagnosis penyakit jantung koroner menggunakan dua algoritma supervised learning, yaitu K-Nearest Neighbors (KNN) dan Random Forest. Proses modeling dilakukan setelah data melalui tahap preprocessing dan pembagian data latih dan uji.

K-Nearest Neighbors (KNN) merupakan algoritma klasifikasi berbasis instance yang bekerja dengan prinsip kemiripan antar data. Algoritma ini mengklasifikasikan data baru berdasarkan mayoritas kelas dari sejumlah K tetangga terdekat dalam ruang fitur. Penentuan tetangga dilakukan dengan menghitung jarak umumnya menggunakan Euclidean Distance antara data baru dan seluruh data latih. KNN bersifat lazy learning, artinya tidak ada proses pelatihan eksplisit, dan prediksi dilakukan langsung saat data baru diberikan. Dalam penelitian ini, nilai

k = 17 memberikan hasil terbaik dengan akurasi 82.00%. Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, F1-score, serta visualisasi confusion matrix dan K-Fold Cross Validation [13].

Random Forest adalah algoritma ensemble learning yang menggabungkan keluaran dari banyak decision tree untuk menghasilkan prediksi yang lebih akurat dan stabil. Setiap pohon dibangun dari subset data dan subset fitur yang dipilih secara acak, dan hasil akhir ditentukan melalui voting mayoritas. Algoritma ini efektif dalam menangani data non-linear dan mengurangi risiko overfitting. Dalam studi ini, parameter max_depth = 10 menghasilkan performa terbaik dengan akurasi 84.00%. Evaluasi dilakukan dengan metrik yang sama seperti pada KNN, dan hasil menunjukkan bahwa Random Forest memiliki performa lebih konsisten, terutama pada rasio pembagian data 80:20 dan fold ke-4 dalam cross-validation [14].

2.5. Evaluasi

Evaluasi model dilakukan untuk mengukur performa algoritma K-Nearest Neighbors (KNN) dan Random Forest dalam klasifikasi diagnosis penyakit jantung koroner. Proses evaluasi menggunakan empat metrik utama, yaitu accuracy, precision, recall, dan F1-score, yang masing-masing memberikan gambaran kuantitatif terhadap kualitas klasifikasi. Selain itu, digunakan confusion matrix untuk memvisualisasikan distribusi prediksi benar dan salah, serta K-Fold Cross Validation untuk menguji kestabilan model terhadap variasi data [15].

3. Hasil dan Diskusi

Penelitian ini mengimplementasikan dua algoritma machine learning, yaitu K-Nearest Neighbors (KNN) dan Random Forest, untuk klasifikasi diagnosis penyakit jantung koroner berdasarkan data rekam medis pasien RSUD Dr. Soeselo Slawi. Proses analisis dilakukan melalui tahapan preprocessing, transformasi fitur kategorikal, normalisasi, pembagian data dengan perbandingan 80:20, pelatihan model, dan evaluasi performa.

3.1. Implementasi Program

Implementasi dalam Python melibatkan penggunaan pustaka-pustaka penting seperti NumPy, Seaborn, Matplotlib, Min-Max Scaler, Random Forest Classifier, K-Neighbors Classifier, dan fungsi evaluasi seperti accuracy_score, confusion_matrix, precision_score, recall_score, serta f1_score. Setelah melakukan import library, dilakukan pembacaan dataset dengan membuat fungsi untuk pembacaan dataset. Selanjutnya, fungsi df = pd.read_csv('data.csv', sep=',') digunakan untuk membaca file data.csv dan fungsi df.head() digunakan untuk menampilkan lima baris pertama dari dataset. Tujuan dari langkah ini adalah untuk melakukan inspeksi awal terhadap struktur data, termasuk nama kolom, tipe data, dan potensi adanya nilai kosong atau inkonsistensi format.

Table 2. Hasil Import Dataset

ID	Umur	Jenis Kelamin	Riwayat Keluarga	Chest Pain Type (cp)	Trestbps (mmhg)	Chol (mg/dM)	Fasting Blood Sugar (fbs)	BMI	Aktivitas Fisik	Merokok	Diagnosis
1	53	F	Tidak	3	118	216	87	20.5	Sedikit	Tidak	Tidak Terindikasi
2	61	F	Tidak	2	136	206	107	19.35	Normal	Tidak	Terindikasi
3	73	M	Ya	2	124	175	154	24.55	Aktif	Ya	Terindikasi
4	46	F	Tidak	2	140	120	124	19.53	Sedikit	Tidak	Terindikasi
5	55	F	Tidak	3	153	260	155	19.95	Aktif	Tidak	Terindikasi
496	46	M	Ya	3	149	304	156	26.04	Normal	Ya	Terindikasi
497	49	F	Tidak	1	131	165	119	20.56	Normal	Tidak	Tidak Terindikasi
498	49	M	Tidak	4	126	186	111	24.50	Aktif	Ya	Terindikasi
499	59	M	Ya	3	127	196	164	21.55	Normal	Tidak	Terindikasi
500	50	F	Ya	4	132	180	123	26.44	Normal	Ya	Terindikasi

Tabel 2 tampilan data setelah dilakukan inputan data, dari dataset yang sudah ada terdapat atribut yaitu umur, Jenis Kelamin, Riwayat Keluarga, Chest Pain, Trestbps, Chol, Fasting Blood Sugar, BMI, Aktivitas Fisik, Merokok dan

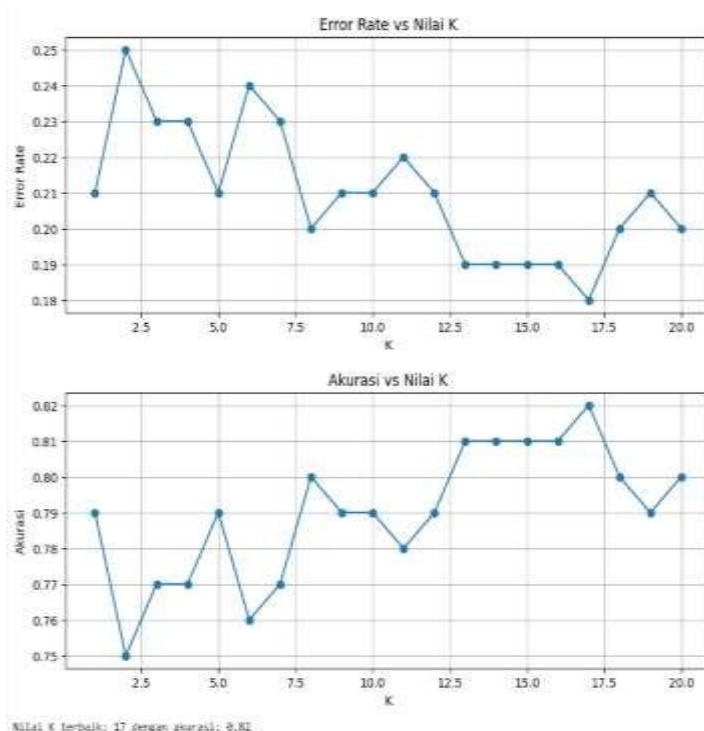
Diagnosis. Jumlah data yang digunakan adalah 500 data pasien setelah melakukan input dataset, dilakukan Preprocessing data untuk memastikan bahwa data yang digunakan bebas dari kesalahan, inkonsistensi, atau informasi yang tidak relevan. Langkah awal meliputi pembersihan data (penghapusan kolom ID dan penanganan missing values), diikuti transformasi fitur kategorikal menggunakan label encoding. Analisis korelasi Pearson digunakan untuk mengidentifikasi hubungan antar fitur dan menghindari multikolinearitas. Fitur numerik dinormalisasi dengan Min-Max Scaler agar skala data seragam. Seluruh proses diimplementasikan menggunakan Python (pandas, scikit-learn) dalam Jupyter Notebook, menghasilkan dataset yang bersih dan siap untuk modeling dengan algoritma KNN dan Random Forest. Hasil Preprocessing data dapat dilihat pada Tabel 3

Table 3. Hasil Preprocessing data

Umur	Jenis Kelamin	Riwayat Keluarga	Chest Pain Type (cp)	Trestbps (mmHg)	Chol (mg/dl)	Fasting Blood Sugar (fbs)	BMI	Aktivitas Fisik	Merokok
0.333333	0.0	0.0	0.5	0.066667	0.391829	0.426103	0.114130	0.053548	1.0
0.153846	0.0	0.0	0.5	0.333333	0.467088	0.377193	0.222086	0.043268	0.0
0.697436	1.0	1.0	0.5	0.133333	0.391829	0.412281	0.074923	0.077024	0.0
0.153846	0.0	0.0	0.5	0.333333	0.142366	0.000000	0.315217	0.077108	0.0
0.256410	0.0	0.5	0.0	0.066667	0.599282	0.614035	0.483696	0.047900	0.0

Setelah melakukan preprocessing data, dilakukan split data untuk membagi menjadi dua yaitu data pelatihan dan data pengujian. Data pelatihan berfungsi untuk melatih model, sementara data pengujian digunakan untuk mengevaluasi kinerjanya. Pembagian dataset dilakukan dengan rasio 80% untuk pelatihan dan 20% untuk pengujian, sehingga 400 entri dialokasikan untuk pelatihan dan 100 entri untuk pengujian. Setelah melakukan split data, dilakukan tahapan pemodelan data bertujuan untuk membangun model klasifikasi diagnosis penyakit jantung koroner menggunakan dua algoritma supervised learning, yaitu K-Nearest Neighbors (KNN) dan Random Forest.

Algoritma KNN merupakan metode klasifikasi yang bekerja berdasarkan kedekatan jarak antar data, di mana suatu data akan diklasifikasikan ke dalam kelas yang paling umum di antara k tetangga terdekatnya. Sebagai lanjutan dari eksperimen pemilihan parameter k pada algoritma K-Nearest Neighbors (KNN), dilakukan visualisasi terhadap dua metrik utama yaitu error rate dan akurasi untuk tiap nilai k dalam rentang 1 hingga 20. Gambar error rate dan akurasi dapat dilihat pada Gambar 2



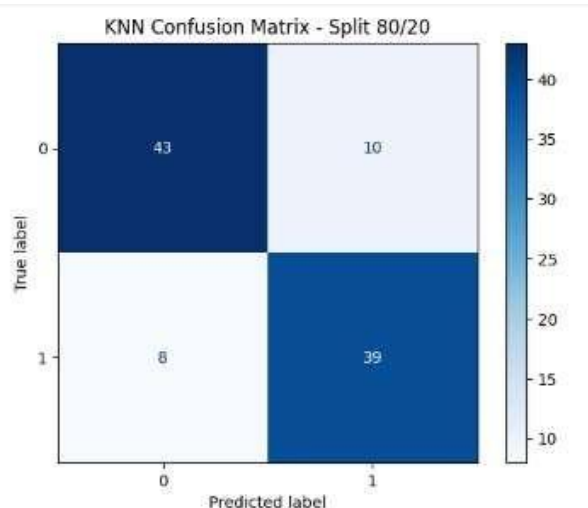
Gambar 2. Hasil error rate dan akurasi

Berdasarkan Gambar 2 menunjukkan bahwa nilai $K = 17$ menghasilkan akurasi tertinggi (82.00%) dan error rate terendah, sehingga dipilih sebagai konfigurasi optimal untuk model KNN pada klasifikasi diagnosis penyakit jantung koroner. Setelah model K-Nearest Neighbors (KNN) selesai dilatih dan diuji, tahap selanjutnya adalah melakukan evaluasi performa klasifikasi menggunakan metrik kuantitatif. Evaluasi dilakukan dengan menghitung empat metrik utama, yaitu accuracy, precision, recall, dan F1-score, yang masing-masing memberikan gambaran kualitas prediksi model terhadap data uji. Langkah pertama dalam proses evaluasi adalah mengimpor fungsi-fungsi evaluasi dari pustaka scikit-learn, yaitu `accuracy_score`, `precision_score`, `recall_score`, dan `f1_score`. Fungsi-fungsi ini digunakan untuk menghitung nilai metrik berdasarkan hasil prediksi model dan label sebenarnya dari data uji.

Selanjutnya, hasil prediksi terbaik dari model KNN disimpan dalam variabel `best_prediction_knn`, yang merupakan hasil dari proses klasifikasi sebelumnya. Evaluasi dilakukan dengan membandingkan `best_prediction_knn` terhadap `y_test`, yaitu label sebenarnya dari data uji. Perhitungan dilakukan dengan pendekatan rata-rata tertimbang (`average='weighted'`) untuk precision, recall, dan F1-score, guna mengakomodasi distribusi kelas yang tidak seimbang. Nilai-nilai metrik kemudian dikonversi ke dalam bentuk persentase dan dibulatkan hingga dua angka di belakang koma untuk memudahkan interpretasi. Hasil evaluasi KNN dapat dilihat pada tabel 4

Matrik	Niai (%)
Accuracy	82.00
Precision	82.09
Recall	82.00
F1-Score	82.01

Dalam penelitian ini pada algoritma KNN mendapatkan nilai accuracy sebesar 82.00%, precision sebesar 82.09%, recall sebesar 82.00% dan f1-score sebesar 82.01%. Gambar Hasil confusion matrix K-Nearest Neighbors dapat dilihat pada Gambar 3



Gambar 3. Hasil Confusion Matrik KNN

Berdasarkan Gambar 5 dapat dijelaskan bahwa Hasil visualisasi confusion matrix KNN menunjukkan bahwa model berhasil memprediksi 43 sampel sebagai Tidak Terindikasi dan benar (True Negative), serta 39 sampel sebagai Terindikasi dan benar (True Positive). Sementara itu, terdapat 10 kasus yang salah diprediksi sebagai Terindikasi padahal sebenarnya Tidak Terindikasi (False Positive), dan 8 kasus yang seharusnya Terindikasi, namun diprediksi sebagai Tidak Terindikasi (False Negative).

Selain confusion matrix, kinerja algoritma KNN juga dievaluasi menggunakan K-fold cross-validation. Metode ini menguji performa model KNN dengan cara membagi data menjadi k subkelompok (fold) yang lebih kecil, evaluasi teknik K-Fold Cross Validation untuk mengukur kestabilan akurasi terhadap variasi data. Rentang jumlah fold yang diuji adalah dari 2 hingga 9. Pada setiap iterasi, data dibagi secara acak, dan model KNN dilatih serta diuji menggunakan fungsi `cross_val_score` dari pustaka `scikit-learn`.

Table 5. Hasil K-fold Cross Validation KNN

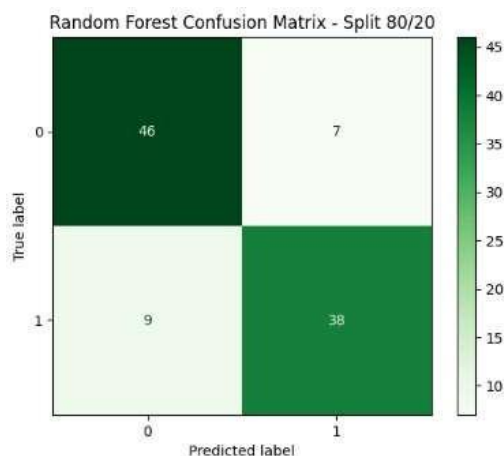
Folds	Accuracy
2	80.88%
3	81.01%
4	80.40%
5	80.66%
6	80.82%
7	80.43%
8	80.64%
9	81.04%

Berdasarkan Tabel 5 dapat disimpulkan bahwa nilai akurasi terbaik K-fold cross validation KNN diperoleh pada `n_fold = 9` sebesar 81.04%.

Selain algoritma K-Nearest Neighbor (KNN), penelitian ini juga menerapkan algoritma Random Forest digunakan untuk melakukan klasifikasi. Model Random Forest dilatih dan diuji pada setiap nilai `max_depth` menggunakan data yang telah dibagi menjadi training dan testing dengan proporsi 80:20. Setelah model Random Forest selesai dilatih dan diuji menggunakan data diagnosis penyakit jantung koroner, dilakukan evaluasi akhir untuk menilai kinerja klasifikasi secara kuantitatif. Evaluasi dilakukan dengan mengukur akurasi, precision, recall, dan f1-score. Nilai-nilai metrik kemudian dikonversi ke dalam bentuk persentase dan dibulatkan hingga dua angka di belakang koma untuk memudahkan interpretasi. Hasil evaluasi KNN dapat dilihat pada tabel 6

Matrik	Niai (%)
Accuracy	84.00
Precision	84.02
Recall	84.00
F1-Score	83.97

Dalam penelitian pada algoritma Random Forest mendapatkan nilai accuracy sebesar 84.00%, precision sebesar 84.02%, recall sebesar 84.00% dan f1-score sebesar 83.97%. Gambar Hasil confusion matrix Random Forest dapat dilihat pada Gambar 4



Gambar 4. Hasil Confusion Matrix Random Forest

Berdasarkan Gambar 7 dapat dijelaskan bahwa algoritma Random Forest dengan diperoleh Hasil visualisasi confusion matrix menunjukkan bahwa model berhasil memprediksi 46 sampel sebagai Tidak Terindikasi dan benar (True Negative), serta 38 sampel sebagai Terindikasi dan benar (True Positive). Sementara itu, terdapat 7 kasus yang salah diprediksi sebagai Terindikasi padahal sebenarnya Tidak Terindikasi (False Positive), dan 9 kasus yang seharusnya Terindikasi, namun diprediksi sebagai Tidak Terindikasi (False Negative).

Selain menggunakan evaluasi confusion matrix, algoritma Random Forest juga melakukan evaluasi K-fold cross validation. Evaluasi tambahan terhadap model klasifikasi bertujuan untuk menguji kestabilan akurasi terhadap variasi jumlah fold, rentang fold yang diuji adalah dari 2 hingga 9. Pada setiap iterasi, data dibagi secara acak menggunakan KFold dengan parameter shuffle=True dan random_state=42 untuk memastikan replikasi hasil. Model Random Forest dikonstruksi dengan parameter max_depth terbaik dari tuning sebelumnya dan random_state=1. Evaluasi dilakukan menggunakan fungsi cross_val_score dengan metrik akurasi. Rata-rata akurasi dari setiap fold disimpan dan ditampilkan dalam format persentase. Hasil evaluasi menunjukkan bahwa akurasi model bervariasi tergantung jumlah fold, dan proses ini membantu mengidentifikasi konfigurasi cross-validation yang paling stabil untuk model Random Forest.

Folds	Accuracy
2	81.20%
3	83.61%
4	84.80%
5	83.60%
6	84.20%
7	83.41%
8	84.03%
9	83.23%

Berdasarkan Tabel 7 dapat disimpulkan bahwa nilai akurasi terbaik K-fold cross validation Random Forest diperoleh pada $n_fold = 4$ sebesar 84.80%.

Berikut hasil yang diperoleh dari perbandingan Algoritma KNN dan Random Forest berdasarkan pembagian data (60:40, 70:30, 80:20, dan 90:10) yang ditunjukkan pada Tabel 8

Table 8. Hasil Semua Split Data

Model	Split	Best K / Depth	Best Fold	K-Fold Accuracy (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNN	60/40	8	8	80.36	82.00	82.00	82.00	82.00
Random Forest	60/40	10	6	82.33	83.00	83.01	83.00	82.97
KNN	70/30	13	6	79.42	81.33	81.39	81.33	81.35
Random Forest	70/30	20	2	83.43	82.00	82.00	82.00	81.97
KNN	80/20	17	4	80.50	82.00	82.09	82.00	82.01
Random Forest	80/20	10	4	84.50	84.00	84.02	84.00	83.97
KNN	90/10	20	8	80.24	82.00	82.63	82.00	82.02
Random Forest	90/10	10	6	85.56	80.00	81.03	80.00	80.00

3.2. Evaluasi Model

Hasil perhitungan yang dilakukan pada penelitian ini memperoleh hasil tingkat akurasi algoritma Random Forest dalam mendiagnosa penyakit jantung koroner yaitu mendapatkan nilai accuracy 84.00%, precision sebesar 84.02%, recall sebesar 84.00% dan f1-score sebesar 83.97%. Sedangkan evaluasi menggunakan K-fold cross validation memperoleh nilai akurasi sebesar 84.80% pada nilai $n_folds = 4$. Sedangkan untuk tingkat akurasi algoritma KNN dalam mendiagnosa penyakit jantung koroner yaitu mendapatkan nilai accuracy sebesar 82.00%, precision sebesar 82.09%, recall sebesar 82.00% dan f1-score sebesar 82.01%. Sedangkan evaluasi menggunakan K-fold cross validation sebesar 81.04% pada nilai $n_folds = 9$. Sehingga berdasarkan hasil evaluasi tersebut dapat disimpulkan bahwa algoritma Random Forest mendapatkan akurasi terbaik dibandingkan dengan algoritma KNN dalam mendiagnosa penyakit jantung koroner.

Table 9. Komparasi Tingkat Akurasi Algoritma KNN dan Random Forest

Metrik Evaluasi	K-Nearest Neighbors (KNN)	Random Forest	Keterangan
Akurasi Umum	82.00%	84.00%	Random Forest lebih unggul dalam akurasi klasifikasi secara keseluruhan.
Metrik Lainnya	Kurang Konsisten	Unggul & Konsisten	Precision, Recall, dan F1-Score menunjukkan performa Random Forest yang lebih baik.
Confusion Matrix	TP: 39, TN: 43	TP: 38, TN: 46	Random Forest memiliki jumlah False Positive dan False Negative yang lebih rendah.
K-Fold Cross Validation	Akurasi tertinggi 81.04% (Fold 9)	Akurasi tertinggi 84.80% (Fold 4)	Random Forest menunjukkan stabilitas dan ketahanan yang lebih baik terhadap variasi data.
Tuning Parameter Optimal	$k = 17$	$max_depth = 10$	Parameter ini menghasilkan performa terbaik untuk masing-masing algoritma.
Rekomendasi Akhir	Kurang Direkomendasikan	Direkomendasikan	Karena keunggulannya dalam akurasi, konsistensi, dan stabilitas, Random Forest menjadi pilihan utama.

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan dalam melakukan komparasi tingkat akurasi algoritma K-Nearest Neighbors (KNN) dan Random Forest untuk mendiagnosis penyakit jantung koroner menggunakan data dari RSUD Dr. Soeselo Slawi, diperoleh bahwa algoritma Random Forest menunjukkan performa yang lebih unggul dibandingkan KNN. Random Forest berhasil mencapai nilai akurasi sebesar 84.00%, precision sebesar 84.02%, recall sebesar 84.00%, dan F1-score sebesar 83.97%. Evaluasi tambahan menggunakan metode K-Fold Cross Validation dengan $n_folds = 4$ menghasilkan akurasi sebesar 84.80%. Sebaliknya, algoritma KNN memperoleh akurasi sebesar 82.00%, precision sebesar 82.09%, recall sebesar 82.00%, dan F1-score sebesar 82.01%, dengan hasil K-Fold Cross Validation sebesar 81.04% pada $n_folds = 9$. Hasil ini menunjukkan bahwa Random Forest lebih akurat dan stabil dalam mendiagnosis penyakit jantung koroner. Selain itu, pengujian dengan variasi rasio pembagian data (60:40, 70:30, 80:20, dan 90:10) menunjukkan bahwa rasio 80:20 merupakan konfigurasi paling optimal dalam menghasilkan evaluasi kinerja model yang seimbang dan akurat. Pada rasio ini, Random Forest mencapai akurasi tertinggi sebesar 84.00% dengan nilai precision, recall, dan F1-score yang konsisten berada di atas 83%. Algoritma KNN juga menunjukkan performa kompetitif pada rasio yang sama, meskipun lebih sensitif terhadap perubahan distribusi data. Evaluasi menggunakan K-Fold Cross Validation pada berbagai rasio pembagian data turut memperkuat bahwa Random Forest lebih stabil dalam menghadapi variasi distribusi data latih dan uji. Dengan demikian, algoritma Random Forest direkomendasikan sebagai pendekatan paling efektif dalam membangun model machine learning untuk diagnosis penyakit jantung koroner, dengan rasio pembagian data 80:20 sebagai konfigurasi yang optimal.

Referensi

- [1] F. R. Muharram *et al.*, "The 30 Years of Shifting in The Indonesian Cardiovascular Burden—Analysis of The Global Burden of Disease Study," *J Epidemiol Glob Health*, vol. 14, no. 1, pp. 193–212, Mar. 2024, doi: 10.1007/s44197-024-00187-8.
- [2] N. Nasution, M. A. Hasan, and F. Bakri Nasution, "Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset," *IT Journal Research and Development*, vol. 9, no. 2, pp. 140–150, Apr. 2025, doi: 10.25299/itjrd.2025.17941.
- [3] M. Hajiarbabi, "Heart disease detection using machine learning methods: a comprehensive narrative review," Jun. 30, 2024, *AME Publishing Company*. doi: 10.21037/jmai-23-152.
- [4] Petra Valentino and Siska Narulita, "Performansi Algoritma Decision Tree (C4.5) untuk Prediksi Penyakit Jantung," *Jurnal Cakrawala Informatika*, vol. 3, no. 2, pp. 18–24, Dec. 2023, doi: 10.54066/jci.v3i2.349.
- [5] M. A. Khan *et al.*, "Optimal feature selection for heart disease prediction using modified Artificial Bee colony (M-ABC) and K-nearest neighbors (KNN)," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-78021-1.
- [6] D. Haganta Depari *et al.*, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *JURNAL INFORMATIK Edisi ke*, vol. 18, p. 2022, doi: <https://doi.org/10.52958/iftk.v18i3.4694>.
- [7] N. Marito Putry and B. Nurina Sari, "KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS," *Jurnal Sains dan Manajemen*, vol. 10, no. 1, 2022, doi: <https://doi.org/10.31294/evolusi.v10i1.12514>.
- [8] S. E. M. Purba, "A Comparative Study of Drug Prediction Models using KNN, SVM, and Random Forest," *Journal of Information Systems and Informatics*, vol. 7, no. 1, pp. 378–392, Mar. 2025, doi: 10.51519/journalisi.v7i1.1013.
- [9] B. D. Shivahare *et al.*, "Delving into Machine Learning's Influence on Disease Diagnosis and Prediction," *Open Public Health J*, vol. 17, no. 1, May 2024, doi: 10.2174/0118749445297804240401061128.
- [10] A. Carolina Wibowo, S. Ardi Lestari, S. Informasi, F. Ilmu Komputer, and U. Duta Bangsa Surakarta, "ANALISIS PENGGUNAAN MACHINE LEARNING DALAM KLASIFIKASI PENENTUAN PENYAKIT JANTUNG," vol. 9, no. 2, 2024, doi: <https://doi.org/10.51876/simtek.v9i2.395>.
- [11] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Inform Med Unlocked*, vol. 41, Jan. 2023, doi: 10.1016/j.imu.2023.101316.
- [12] H. Kamal *et al.*, "Heart Disease Prediction Using Machine Learning," in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry, IDICAIEI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IDICAIEI61867.2024.10842908.
- [13] M. J. J. Ghrabat *et al.*, "Utilizing Machine Learning for the Early Detection of Coronary Heart Disease," *Engineering, Technology and Applied Science Research*, vol. 14, no. 5, pp. 17363–17375, Oct. 2024, doi: 10.48084/etasr.8171.
- [14] A. A. Ahmad and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," *Diagnostics*, vol. 13, no. 14, Jul. 2023, doi: 10.3390/diagnostics13142392.
- [15] M. Radhi, D. Ryan Hamonangan Sitompul, S. Hamonangan Sinurat, and E. Indra, "ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK," *Jurnal Sistem Informasi dan Ilmu Komputer Prima*, vol. 4, no. 2, 2021, doi: <https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475>.