Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Penerapan Machine Learning Algoritma Random Forest Untuk Prediksi Penyakit Jantung

Shafira Amanda Putri[1], Nabilah Selayanti[2], Mirechelin Kristanaya[3],
Melinda Putri Azzahra[4], Muhammad Ghinan Navsih[5], Kartika Maulida Hindrayani[6]
[1,2,3,4,5] *Program Studi Sains Data, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur*
[1]22083010008@student.upnjatim.ac.id, [2]22083010013@student.upnjatim.ac.id, [3]22083010032@student.upnjatim.ac.id
[4]22083010035@student.upnjatim.ac.id, [5]22083010057@student.upnjatim.ac.id, [6]kartika.maulida.ds@upnjatim.ac.id
Corresponding author email: 22083010013@student.upnjatim.ac.id

Abstract: *Heart diseases are the leading cause of death around the world. More than 17 million deaths occur each year due to the same, says WHO. In this context, the paper presents the study regarding the improvement in the precision of heart disease prediction, which uses the Random Forest algorithm, noted for handling complex data and protecting against overfitting. This present study adopts the Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression models in classifying a Kaggle dataset on heart failure prediction with 918 records and 12 attributes. Preprocessing steps include missing data processing, categorical encoding, and normalization. For balancing classes, the SMOTE oversampling technique will be used. Model evaluation is done based on accuracy, recall, and F1-score. Performance in the model based on Random Forest is 87.7%, with test data and 92.63% on validation data. These results indicate the extent to which a random forest can be used in clinical decision-making toward predicting heart diseases.*
Keywords: *Heart disease, Random Forest, machine learning, prediction, SMOTE.*

Abstrak: Menurut perkiraan Organisasi Kesehatan Dunia (WHO), penyakit jantung diproyeksikan tetap menjadi penyebab kematian utama di dunia pada tahun 2030.. Lebih dari 17 juta jiwa dilaporkan kehilangan nyawa setiap tahun karena penyakit kardiovaskular, menurut WHO. Penelitian ini bertujuan meningkatkan akurasi deteksi penyakit jantung dengan memanfaatkan algoritma Random Forest untuk menangani kompleksitas data dan mencegah overfitting. Penelitian ini mengevaluasi kinerja empat algoritma pembelajaran mesin, yaitu Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Regresi Logistik, dalam memprediksi kegagalan jantung. Evaluasi dilakukan dengan menggunakan dataset "Heart Failure Prediction Dataset" yang diunduh dari Kaggle yang terdiri dari 918 observasi dengan 12 atribut. Langkah-langkah persiapan datanya termasuk pengelolaan nilai yang hilang, pemrosesan variabel kategorik, dan normalisasi. Dalam pengalaman nyata, Hasil menunjukkan bahwa Random Forest menghasilkan hasil terbaik dengan akurasi 87.7% di data pengujian dan 92.63% di data validasi. Oleh karena itu, penekanan penelitian ini adalah pada kemampuan Random Forest dalam.
Kata kunci: Penyakit Jantung, Random Forest, Machine Learning, Prediksi, SMOTE

I. PENDAHULUAN

Penyakit kardiovaskular, seperti penyakit jantung koroner, stroke, dan berbagai kondisi lainnya, merupakan salah satu penyebab kematian tertinggi di dunia. Menurut data dari WHO, lebih dari 17 juta orang meninggal setiap tahun akibat penyakit-penyakit ini [1]. Hal ini menjadikan penyakit jantung sebagai beban berat bagi kesehatan global dan mendorong fokus penelitian untuk mengembangkan strategi pencegahan yang efektif.

Dengan kemajuan teknologi dalam bidang data mining dan machine learning, berbagai pendekatan telah dimanfaatkan untuk memperoleh tingkat prediksi yang tinggi dalam mendiagnosis penyakit jantung. Salah satu algoritma yang telah memperlihatkan tingkat keberhasilan yang baik adalah Hutan Acak. Pendekatan Hutan Acak adalah proses pembelajaran ensemble di mana akurasi dan stabilitasnya ditingkatkan dengan merata-rata prediksi dari sejumlah besar [2]. Algoritma ini sejak lama digunakan karena kemampuannya dalam menangani data yang rumit, tetapi juga dalam mengatasi sebagian dari masalah overfitting yang muncul dalam algoritma lain.

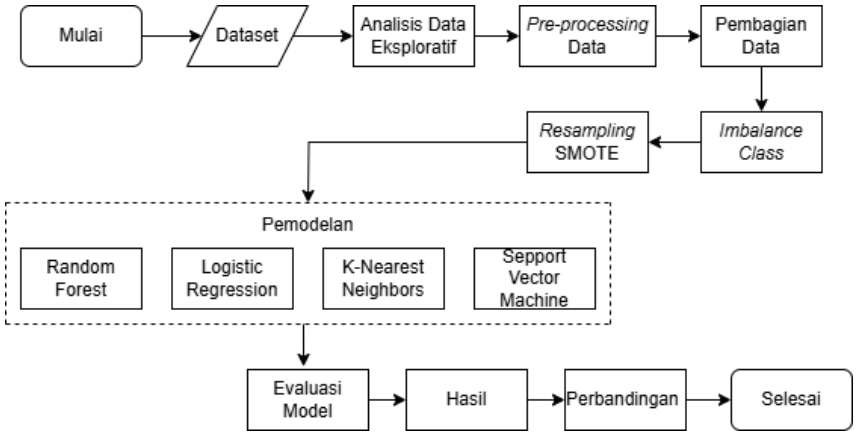
Penelitian sebelumnya yang dilakukan oleh Devina Larassati, Ati Zaidah, Sarika Afrizal pada tahun 2022 bertujuan untuk memprediksi risiko penyakit jantung koroner. Penelitian ini menggunakan data

Heart Attack Analysis & Prediction dari Kaggle dan algoritma Naïve Bayes. Data yang digunakan dalam analisis ini hanyalah 303 baris informasi dengan 13 fitur, termasuk usia, jenis kelamin. Hasil eksperimen menunjukkan bahwa dengan skema data latih dan data uji yang berbeda, akurasi maksimum yang diperoleh adalah 83.1%. Selain itu, penulis juga mengembangkan sistem prediksi sederhana yang mungkin digunakan oleh dokter dalam mendiagnosis penyakit jantung dari data pasien yang diberikan [3].

Salah satu isu yang dihadapi saat mengembangkan model pembelajaran mesin adalah ketidakseimbangan. Solusi dari masalah ini adalah dengan menerapkan SMOTE *(Synthetic Minority Over-Sampling Technique)* untuk menjadikan data minoritas [4]. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja beberapa algoritma klasifikasi, termasuk Random Forest, Support Vector Machine, K-Nearest Neighbors, dan Logistic Regression.

II. METODE PENELITIAN

Prosedur penelitian ini diilustrasikan dalam sebuah diagram alur, yang memperlihatkan rangkaian langkah-langkah yang sistematis dan logis, mulai dari awal hingga akhir penelitian.



Gambar 1. *Flowchart* Alur Penelitian

II.1. Pengumpulan Data

Penelitian ini menggunakan dataset *"Heart Failure Prediction Dataset"* yang diunduh dari Kaggle https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. Dataset ini mencakup 918 data dengan 12 atribut. Atribut-atribut tersebut meliputi *Age*, *Sex*, *ChestPainType*, *RestingBP*, *Cholesterol*, *FastingBS*, *RestingECG*, *MaxHR*, *ExerciseAngina*, *OldPeak*, *ST_Slope*, dan *HeartDisease*. Dataset ini tersedia secara bebas untuk digunakan di bawah lisensi *Open Database License* (ODbL). Sesuai dengan ketentuan lisensi tersebut, penggunaan dataset ini dalam penelitian telah memenuhi persyaratan lisensi dan tidak memerlukan izin khusus tambahan.

II.2. Analisis Data Eksploratif

Eksplorasi data dilakukan untuk memahami karakteristik dan hubungan antar fitur dalam *dataset*. Pada tahapan ini, berbagai visualisasi seperti distribusi target, korelasi antar fitur, dan distribusi fitur-fitur akan digunakan untuk mendapatkan wawasan yang lebih dalam tentang data. Tujuan dari eksplorasi data adalah untuk membantu pemahaman tentang pola-pola yang ada dalam data dan memandu proses pemodelan selanjutnya.

II.3. *Preprocessing Data*

Tahap preprocessing data sangat penting dalam penelitian ini untuk memastikan data berkualitas sebelum digunakan dalam pemodelan. Langkah-langkah preprocessing mencakup penanganan *missing values*, *encoding* data kategorik dengan LabelEncoder, dan normalisasi data menggunakan *MinMaxScaler*. Penanganan *missing values* mengatasi nilai kosong yang bisa mempengaruhi akurasi model, *encoding* data kategorik mengubah data menjadi numerik untuk digunakan dalam pemodelan, dan normalisasi data menyesuaikan skala fitur untuk mencegah dominasi satu fitur atas lainnya.

II.4. Pembagian Data

Pembagian data dilakukan untuk membagi dataset menjadi dua subset utama: data training dan data testing. Dalam pembagian ini, 80% dari data akan digunakan untuk melatih model (data training), sementara 20% sisanya akan dialokasikan untuk menguji performa model (data testing). Alasan di balik pembagian ini adalah untuk memastikan bahwa model memiliki cukup data untuk belajar dan mengenali pola dari data training, serta memiliki data yang memadai untuk mengevaluasi kinerjanya secara objektif. Sehingga dapat memberikan estimasi yang andal mengenai kemampuan generalisasi model terhadap data baru.

II.5. *Imbalance Class*

Ketidakseimbangan kelas (*imbalance class*) terjadi ketika jumlah sampel atau observasi dalam satu kelas jauh lebih besar atau lebih kecil dibandingkan kelas lainnya dalam konteks klasifikasi. Hal ini dapat menyulitkan proses pembelajaran classifier dan menyebabkan kinerja klasifikasi menjadi tidak optimal. Ketika kelas tidak seimbang, kualitas data dalam hal kinerja klasifikasi menurun, membuat model kurang efektif dalam memprediksi hasil dengan akurasi tinggi.

II.6. Resampling SMOTE

Penanganan *class imbalance* menggunakan metode SMOTE *(Synthetic Minority Over-Sampling Technique)* untuk mengatasi ketidakseimbangan antara kelas target, sehingga model dapat mempelajari pola dengan lebih baik. Dengan melakukan tahap ini, data siap digunakan untuk tahap selanjutnya dalam analisis dan pemodelan.

II.7. Pemodelan

1. *Random Forest*

Algoritma Random Forest merupakan suatu teknik dalam ensemble learning yang mengintegrasikan beberapa pohon keputusan melalui proses sampling terpadu dengan penggunaan prediktor acak. Metode ini memprediksi dengan menggabungkan hasil dari setiap pohon, menggunakan suara mayoritas untuk klasifikasi dan nilai rata-rata untuk regresi [5]. Random Forest, meskipun menggunakan ansambel pohon keputusan, tidak dapat menetapkan signifikansi individu dari setiap variabel tetapi dapat menunjukkan tingkat kepentingan variabel [6]. Tingkat kepentingan variabel diukur menggunakan Mean Decrease in Gini (MDG). Jika terdapat q variabel penjelas dengan $h = 1, 2, \ldots, q$, maka *Mean Decrease in Gini* (MDG) akan memastikan pentingnya variabel X_h sebagai berikut:

$$MDG_h = \frac{1}{k}\Sigma_t[d(h,t)I(h,t)] \quad (1)$$

dengan

k = jumlah pohon dalam struktur random forest
$d(h,t)$ = besar penurunan indeks Gini untuk variabel penjelas X_h pada sebuah simpul

$I(h,t)] = \begin{cases} 1; & X_h \text{ memilih simpul } t, \\ 0; & \text{selainnya} \end{cases}$

2. *Logistic Regression*

Logistic Regression adalah sebuah teknik analisis yang dipakai untuk menilai hubungan antara variabel independen dan variabel dependen yang terdiri dari nilai biner atau nominal [7]. Tujuannya adalah untuk menjelaskan hubungan antara satu atau lebih variabel independen dengan variabel dependen yang memiliki dua atau lebih kategori, menggunakan estimasi rasio peluang yang disesuaikan [8]. Rumus dasar dari Logistic Regression adalah

$$g(X) = sigmoid(\alpha + \beta X) \quad (2)$$

$$sigmoid(x) = \frac{1}{1 + exp - x}$$

keterangan :
α = konstanta
β = koefisien regresi (kemiringan)
X = variabel faktor penyebab (independent)

Dimana fungsi sigmoid menghasilkan probabilitas yang berada dalam rentang 0 hingga 1, sesuai dengan karakteristik data biner atau kategorikal yang umumnya digunakan dalam analisis prediktif dan interpretatif.

3. *Support Vector Machine* (SVM)

Algoritma *Support Vector Machine* (SVM) bertujuan untuk menemukan hyperplane optimal dalam ruang dimensi tinggi yang dapat memisahkan dua kelas data. Proses ini melibatkan pencarian hyperplane yang bertujuan untuk memaksimalkan margin atau jarak antara pola data latih dan batas keputusan [9]. SVM menggunakan teknik kernel untuk mentransformasikan data ke ruang kernel, di mana kelas data dapat dipisahkan secara linear. Persamaan matematis yang mendasari SVM, seperti yang dijelaskan oleh Octaviani et al. [10], menggambarkan hyperplane dalam bentuk

$$f(x) = w^T x + b \quad (3)$$

dimana
w = vektor bobot
x = adalah vektor fitur
b = konstanta bias

Jarak atau margin antara dua kelompok objek dari kelas yang berbeda bisa dihitung sebagai $\frac{2}{||w||}$ di mana $||w||$ adalah norma Euclidean dari vektor bobot w yang digunakan untuk menentukan lebar margin maksimal.

4. *K-Nearest Neighbors* (KNN)

Metode K-Nearest Neighbor (KNN) merupakan salah satu pendekatan sederhana dalam data mining yang menggunakan contoh dan bersifat non-parametrik. Pemilihan nilai K, jumlah tetangga terdekat yang dipertimbangkan dalam proses klasifikasi, memiliki dampak signifikan terhadap akurasi algoritma ini. Nilai K yang rendah dapat

menyebabkan KNN peka terhadap gangguan pada data, sedangkan nilai K yang tinggi bisa mengakibatkan model menjadi bias [11]. Langkah-langkah utama algoritma meliputi menentukan nilai K, menghitung jarak Euclidean untuk setiap objek data, dan memilih kelas mayoritas dari tetangga terdekat sebagai hasil klasifikasi. KNN, meskipun sederhana, memerlukan pertimbangan yang hati-hati terhadap nilai K dan perhitungan jarak Euclidean untuk memastikan akurasi dan sensitivitas yang optimal dalam mengklasifikasikan objek baru dalam dataset kompleks [12].

II.8. Evaluasi Hasil

Sesudah prosedur pemodelan diselesaikan, selanjutnya adalah melakukan penilaian kinerja model menggunakan data pengujian dan validasi, dengan tujuan untuk menentukan kemampuan prediktif terhadap penyakit jantung. Model dievaluasi dengan metrik seperti akurasi, *recall*, dan F1-score, dan hasilnya divisualisasikan menggunakan confusion matrix. Pada evaluasi data uji, model dilatih dan diuji dengan data terpisah, sedangkan pada evaluasi data validasi, data latih dibagi menjadi data latih dan data validasi untuk mendeteksi overfitting dan memastikan generalisasi yang baik. Evaluasi ini bertujuan memilih model terbaik yang dapat digunakan dalam praktik klinis untuk prediksi penyakit jantung.

III. HASIL DAN PEMBAHASAN

III.1. Pengumpulan Data

Dataset yang digunakan memuat informasi karakteristik pasien, termasuk usia, jenis kelamin (M: Pria, F: Wanita), jenis nyeri dada (TA: Angina Tipikal, ATA: Angina Atipikal, NAP: Nyeri Non-Angina, ASY: Tanpa Gejala), tekanan darah istirahat, kolesterol serum, gula darah puasa (1 jika >120 mg/dl, 0 jika ≤120 mg/dl), hasil elektrokardiogram (Normal, ST, LVH), denyut jantung maksimum, angina saat berolahraga (Y: Ya, N: Tidak), nilai oldpeak (depresi ST), kemiringan segmen ST (Naik, Datar, Turun), dan status penyakit jantung (1: penyakit jantung, 0: normal). Data ini dianalisis untuk memahami hubungan antara faktor-faktor tersebut dengan penyakit jantung.

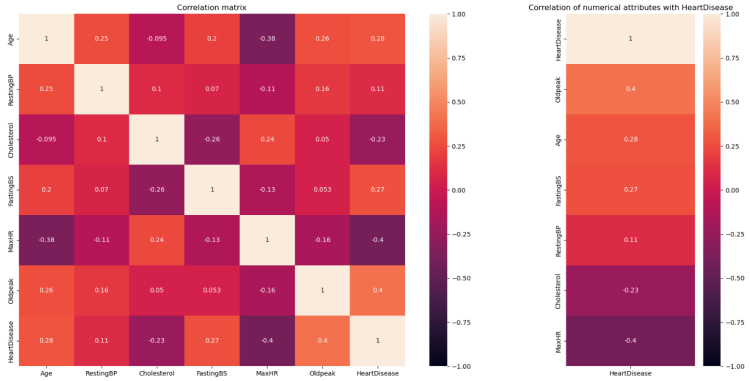
Tabel 1. Dataset

Age	Sex	ChestPain Type	RestingBP	Cholesterol	Fasting BS	RestingECG	Max HR	ExersizeAngina	OldPeak	ST_Slope	HearDisease
40	M	ATA	140	200	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
…	…	…	…	…	…	…	…	…	…	…	…
45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

III.2. Analisis Data Eksploratif

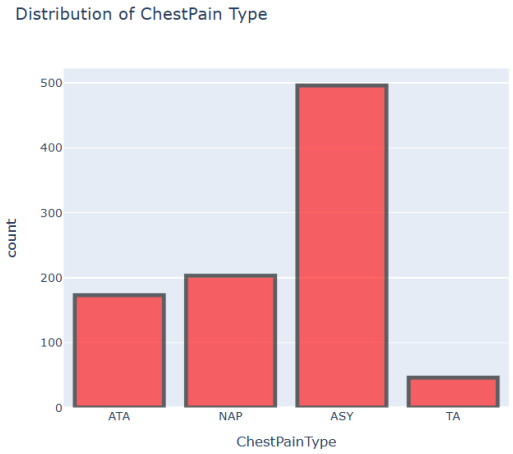
Eksplorasi data adalah tahap penting dalam proses analisis data dan digunakan untuk berbagai tujuan yang esensial dalam memahami dataset yang akan digunakan.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 2. Heatmap Korelasi

Korelasi antar variabel di atas mengindikasi bahwa variabel "Age", "Oldpeak", dan "RestingBP" memiliki korelasi positif yang signifikan dengan "HeartDisease", artinya pasien yang lebih tua, dengan nilai "Oldpeak" dan "RestingBP" yang lebih tinggi, memiliki risiko penyakit jantung yang lebih tinggi. Sebaliknya, "MaxHR" memiliki korelasi negatif yang signifikan, menunjukkan bahwa pasien dengan nilai "MaxHR" yang lebih rendah memiliki risiko penyakit jantung yang lebih tinggi. Variabel "Cholesterol" dan "FastingBS" memiliki korelasi yang lemah dengan "HeartDisease", sehingga mungkin kurang penting dalam memprediksi risiko penyakit jantung.



Gambar 3. Plot Distribusi Jenis Nyeri Dada

Grafik tersebut menunjukkan bahwa sebagian besar pasien dalam dataset mengalami nyeri dada asimptomatik (ASY) dengan sekitar ± 500 kasus, diikuti oleh nyeri non-angina (NAP) dengan sekitar ± 200 kasus, angina tipikal (ATA) dengan sekitar ± 180 kasus, dan angina atipikal (TA) dengan sekitar ± 50 kasus. Hal ini membuktikan bahwa terdapat sebagian besar pasien tidak mengalami nyeri dada yang berhubungan langsung dengan penyakit jantung, tetapi beberapa mengalami nyeri dada yang mungkin mengindikasikan penyakit jantung. Informasi ini berguna untuk memprediksi kemungkinan penyakit jantung pada pasien berdasarkan jenis nyeri dada.

III.3. *Preprocessing Data*

Pada tahap ini, dilakukan beberapa langkah untuk memurnikan data, sehingga menghasilkan dataset yang rapi dan siap digunakan untuk proses selanjutnya. Tidak terdapat

missing value pada data, sehingga langsung dilanjutkan ke pemisahan variabel x dan y. Variabel dependen, 'HeartDisease', dipisahkan dan disimpan dalam variabel y, sementara variabel independen disimpan dalam variabel X dengan menghapus kolom 'HeartDisease'. Selanjutnya, variabel kategorikal seperti Sex, ChestPainType, RestingECG, ExerciseAngina, dan ST_Slope diubah menjadi representasi numerik menggunakan LabelEncoder.

Tabel 2. Dataset Setelah Transformasi

Age	Sex	ChestPain Type	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	Exersize Angina	OldPeak	ST_Slope
40	1	1	140	200	0	1	172	0	0.0	2
49	0	2	160	180	0	1	156	0	1.0	1
37	1	1	130	283	0	2	98	0	0.0	2
48	0	0	138	214	0	1	108	1	1.5	1
54	1	2	150	195	0	1	122	0	0.0	2
…	…	…	…	…	…	…	…	…	…	…
45	1	3	110	264	0	1	132	0	1.2	1
68	1	0	144	193	1	1	141	0	3.4	1
57	1	0	130	131	0	1	115	1	1.2	1
57	0	1	130	236	0	0	174	0	0.0	1
38	1	2	138	175	0	1	173	0	0.0	2

Setelah itu, data dinormalisasi menggunakan MinMaxScaler untuk memastikan semua fitur berada dalam rentang yang sama.

```
array([[0.24489796, 1.        , 0.33333333, ..., 0.        , 0.29545455,
        1.        ],
       [0.42857143, 0.        , 0.66666667, ..., 0.        , 0.40909091,
        0.5       ],
       [0.18367347, 1.        , 0.33333333, ..., 0.        , 0.29545455,
        1.        ],
       ...,
       [0.59183673, 1.        , 0.        , ..., 1.        , 0.43181818,
        0.5       ],
       [0.59183673, 0.        , 0.33333333, ..., 0.        , 0.29545455,
        0.5       ],
       [0.20408163, 1.        , 0.66666667, ..., 0.        , 0.29545455,
        1.        ]])
```

Gambar 4. Dataset Setelah Normalisasi

Langkah berikutnya adalah memisahkan data menjadi dua kategori, yakni data untuk pelatihan (80%) dan data untuk pengujian (20%). Data pelatihan berfungsi sebagai bahan untuk mengembangkan model, sementara data pengujian digunakan untuk mengevaluasi kemampuan model yang telah dikembangkan.

III.4. *Imbalance Class*

Ketidakseimbangan kelas (*class imbalance*) terjadi ketika jumlah sampel atau observasi dalam satu kelas jauh lebih besar atau lebih kecil dibandingkan kelas lainnya dalam konteks klasifikasi. Hal ini dapat menyulitkan proses pembelajaran classifier dan menyebabkan kinerja klasifikasi menjadi tidak optimal.

```
1    406
0    328
Name: HeartDisease, dtype: int64
```

Gambar 5. Hasil *Imbalance Class*

Didapatkan hasil yang menunjukkan bahwa kelas dengan label 1 memiliki 406 sampel, sedangkan kelas dengan label 0 memiliki 328 sampel. Distribusi kelas ini menunjukkan ketidakseimbangan yang signifikan antara kelas 1 dan kelas 0 dalam dataset.

III.5. SMOTE (*Synthetic Minority Over-sampling Technique*)

Dalam upaya mengatasi ketidakseimbangan kelas dalam dataset, peneliti menerapkan teknik oversampling SMOTE (*Synthetic Minority Over-sampling Technique*) untuk meningkatkan banyaknya sampel pada kelas minoritas dan mencapai keseimbangan kelas yang lebih baik. Setelah penerapan SMOTE menunjukkan bahwa distribusi kelas dalam dataset pelatihan sekarang sudah seimbang.

```
1    406
0    406
Name: HeartDisease, dtype: int64
```

Gambar 6. Hasil SMOTE

Kedua kelas tersebut telah mencapai keseimbangan, di mana kelas 1 dan kelas 0 masing-masing memiliki 406 sampel, sehingga jumlah sampel antara kedua kelas menjadi sama. Dengan demikian, ketidakseimbangan kelas dalam dataset telah berhasil ditangani dengan menggunakan teknik oversampling menggunakan SMOTE.

III.6. Evaluasi Model pada Data Uji

Untuk membandingkan efektivitasnya, penelitian ini bertujuan mengevaluasi algoritma Random Forest dalam memprediksi penyakit jantung dengan algoritma *machine learning* lainnya seperti *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Logistic Regression* pada data uji. Hasil evaluasi model pada data uji memberikan wawasan tentang kemampuan masing-masing model dalam mengklasifikasikan data baru yang belum pernah dilihat sebel

III.6.1. Algoritma *Random Forest*

Tabel 3. Evaluasi Model Data Uji Algoritma *Random Forest*

Accuracy	***Recall***	***F1-Score***
0.87745	0.87378	0.87805

Algoritma *Random Forest* menunjukkan kinerja yang unggul, dengan akurasi tinggi dan keseimbangan yang baik antara *recall* dan *F1-Score*. Hal ini menunjukkan kemampuan model dalam mengidentifikasi kasus penyakit jantung dengan benar.

III.6.2. Algoritma *Support Vector Machine* (SVM)

Tabel 4. Evaluasi Model Data Uji Algoritma *Support Vector Machine*

Accuracy	***Recall***	***F1-Score***
0.86764	0.82608	0.87557

Algoritma SVM juga menunjukkan kinerja yang baik, meskipun sedikit di bawah Random Forest, namun akurasi dan tingginya angka *recall*, model ini terbukti efektif dalam mengidentifikasi kasus penyakit jantung.

III.6.3. Algoritma *K-Nearest Neighbors* (KNN)

Tabel 5. Evaluasi Model Data Uji Algoritma *K-Nearest Neighbors*

Accuracy	*Recall*	*F1-Score*
0.85784	0.85784	0.85784

Algoritma KNN memiliki performa yang konsisten dengan akurasi, recall, dan F1-Score yang sama. Ini menunjukkan bahwa model ini dapat diandalkan untuk prediksi penyakit jantung, meskipun tidak sebaik Random Forest dan SVM.

III.6.4. Algoritma *Logistic Regression*

Tabel 6. Evaluasi Model Data Uji Algoritma *Logistic Regression*

Accuracy	*Recall*	*F1-Score*
0.85784	0.84112	0.86124

Algoritma Logistic Regression menunjukkan kinerja yang sangat baik pada data uji. Model ini menunjukkan bahwa Logistic Regression adalah pilihan yang kuat untuk memprediksi penyakit jantung.

III.7. Evaluasi Model pada Data Validasi

III.7.1. Algoritma *Random Forest*

Tabel 7. Evaluasi Model Data Validasi Algoritma *Random Forest*

Accuracy	*Recall*	*F1-Score*
0.92638	0.90909	0.92105

Algoritma Random Forest menunjukkan kinerja yang sangat baik pada data validasi, dengan recall tertinggi di antara semua model. Hasil tersebut mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam mendeteksi kasus penyakit jantung yang aktual, sehingga dapat diandalkan untuk membantu diagnosis yang akurat.

III.7.2. Algoritma *Support Vector Machine* (SVM)

Tabel 8. Evaluasi Model Data Validasi Algoritma *Support Vector Machine*

Accuracy	*Recall*	*F1-Score*
0.90184	0.88311	0.89473

Algoritma SVM juga menunjukkan performa yang solid, meskipun sedikit di bawah Random Forest dan KNN. Akurasinya cukup tinggi, dan recall serta F1-Score yang baik menunjukkan kemampuan model dalam mengklasifikasikan kasus penyakit jantung.

III.7.3. Algoritma *K-Nearest Neighbors* (KNN)

Tabel 9. Evaluasi Model Data Validasi Algoritma *K-Nearest Neighbors*

Accuracy	*Recall*	*F1-Score*
0.91411	0.90909	0.90909

Algoritma KNN memiliki performa yang sangat baik dengan akurasi dan F1-Score yang tinggi, hampir setara dengan Random Forest. Ini menunjukkan bahwa model ini dapat diandalkan untuk prediksi penyakit jantung.

III.7.4. Algoritma *Logistic Regression*

Tabel 10. Evaluasi Model Data Validasi Algoritma *Logistic Regression*

Accuracy	***Recall***	***F1-Score***
0.90797	0.88311	0.90066

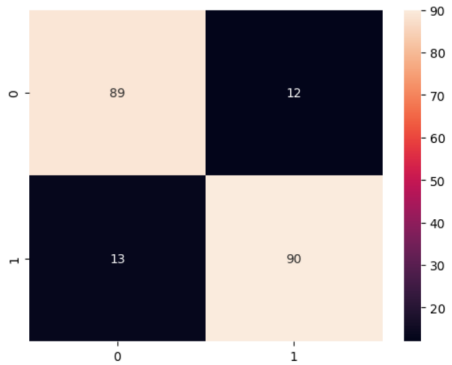
Algoritma Logistic Regression menunjukkan performa yang baik dengan akurasi yang cukup tinggi. Recall dan F1-Score yang seimbang menunjukkan kemampuan model dalam mendeteksi kasus positif dan mengurangi false positive.

III.8. Perbandingan Performa Model Klasifikasi

Tabel 11. Evaluasi Model Data Validasi Algoritma *Logistic Regression*

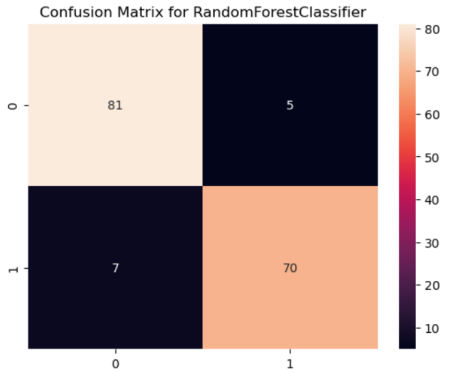
Algoritma	***Accuracy***	***Recall***	***F1-Score***	**Nilai data validasi Akurasi**
Random Forest	0.88	0.87	0.88	92.63%
Support Vector Machine (SVM)	0.87	0.83	0.87	90.18%
K-Nearest Neighbors (KNN)	0.86	0.83	0.86	91.41%
Logistic Regression	0.85	0.84	0.86	90.79%

Algoritma Random Forest terbukti sebagai model yang paling optimal dalam penelitian ini, dengan performa terbaik baik pada data uji maupun data validasi. *K-Nearest Neighbors* juga menunjukkan kinerja yang sangat baik dan dapat menjadi alternatif yang dapat diandalkan. Meskipun SVM dan *Logistic Regression* memberikan hasil yang baik, mereka tidak sebaik *Random Forest* dalam hal stabilitas dan performa keseluruhan. Oleh karena itu, *Random Forest* dapat dipertimbangkan sebagai model utama untuk prediksi penyakit jantung.



Gambar 7. *Confussion Matrix* Data Uji Algoritma *Random Forest*

Pada *Confusion Matrix* data uji, hasil klasifikasi mencakup 89 kasus *True Positive*, di mana individu dikenali secara akurat sebagai tidak terkena penyakit jantung, dan 12 kasus *False Positive*, di mana individu yang sebenarnya tidak terkena penyakit jantung salah diklasifikasikan sebagai positif. Selain itu, terdapat 90 kasus *True Negative*, di mana individu yang tidak terkena penyakit jantung diklasifikasikan dengan benar, dan 13 kasus *False Negative*, di mana individu yang sebenarnya terkena penyakit jantung diklasifikasikan sebagai tidak terkena.



Gambar 8. *Confussion Matrix* Data Validasi Algoritma *Random Forest*

Pada Confusion Matrix data validasi, kita memperoleh jumlah orang yang tidak terkena penyakit jantung atau data yang diklasifikasikan dengan benar sebagai negatif (*True Negative*) sebanyak 81 orang, jumlah data yang diklasifikasikan secara keliru sebagai positif (*False Positive*) sebanyak 5 orang, jumlah data yang diklasifikasikan dengan benar sebagai positif (*True Positive*) sebanyak 70 orang, dan jumlah data yang diklasifikasikan secara keliru sebagai negatif (*False Negative*) sebanyak 7 orang.

IV. KESIMPULAN

Penelitian ini mengevaluasi efektivitas algoritma machine learning dalam memprediksi penyakit jantung menggunakan *Random Forest, Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Logistic Regression*. Hasil menunjukkan bahwa Random Forest unggul dengan akurasi 87.7% pada data uji dan 92.6% pada data validasi, serta *recall* dan skor F1 yang tinggi. *Random Forest* terbukti efektif dalam mengklasifikasikan data baru dan mengidentifikasi kasus positif dengan tingkat kesalahan rendah, menjadikannya pilihan yang potensial untuk mendukung pengambilan keputusan klinis.

Algoritma lain seperti SVM, KNN, dan *Logistic Regression* juga menunjukkan kinerja baik, tetapi tidak sebaik *Random Forest.* Algoritma SVM mencapai akurasi 86.7% pada data uji dan 90.2% pada data validasi, sementara *Logistic Regression* mencapai akurasi 85.7% pada data uji dan 90.8% pada data validasi. Secara keseluruhan, *Random Forest* dapat dipertimbangkan sebagai model yang efektif untuk prediksi penyakit jantung, namun perlu evaluasi lanjutan dan validasi eksternal untuk memastikan keandalan dalam berbagai kondisi dan populasi pasien yang lebih luas. Penelitian ini memberikan dasar yang kuat untuk pengembangan lebih lanjut dan implementasi algoritma machine learning dalam prediksi penyakit jantung, yang dapat meningkatkan kualitas perawatan kesehatan dan mendukung pengambilan keputusan klinis.

REFERENSI

1. Karyatin. (2019). Faktor-Faktor Yang Berhubungan Dengan Kejadian Penyakit Jantung Koroner. *Jurnal Ilmiah Kesehatan*, *11*(1), 37–43.
2. Sumantiawan, D. I. (2024). Metode Analisis Menggunakan Algoritma Random Forest Untuk Prediksi Biaya Asuransi Kesehatan. In *Jurnal Informatika dan Teknologi (JICode)* (Vol. 1, Issue 1). https://doi.org/10.30599/jicode.v1i1.3294
3. Larassati, D., Zaidiah, A., & Afrizal, S. (2022). *Sistem Prediksi Penyakit Jantung Koroner Menggunakan Metode Naïve Bayes*. https://doi.org/https://doi.org/10.29100/jipi.v7i2.2842
4. Afiatuddin, N., Wicaksono, Mt., Rezky Akbar, V., & Wulandari, D. (2024). Komparasi Algoritma Machine Learning dalam Klasifikasi Kanker Payudara. *Jurnal Media Informatika Budidarma*, *8*(2), 889–899. https://doi.org/10.30865/mib.v8i2.7457

5. Afdhal, I., Kurniawan, R., Iskandar, I., Salambue, R., Budianita, E., & Syafria, F. (2022). Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *5*(1), 122–130. https://doi.org/https://doi.org/10.32672/jnkti.v5i1.4004
6. Britanthia, L., Tanujaya, C., Susanto, B., & Saragih, A. (2020). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify. *Indonesian Journal of Data and Science (IJODAS)*, *1*(3), 68–78.
7. Simanjuntak, W. O., Bijaksana, A., Negara, P., & Septriana, R. (2023). Perbandingan Algoritma Logistic Regression dan Random Foret (Studi Kasus : Klasifikasi Emosi Tweet) Comparison Of Logistic Regression And Random Forest Algorithms (Case Study: Tweet Emotion Classification). *Jurnal Aplikasi Dan Riset Informatika*, *02*(1), 160–164. https://doi.org/10.26418/juara.v2i1.69682
8. Rahmawati, I., & Fitriani, T. R. (2023). Analisis Sentimen Menggunakan Algoritma Logistic Regression Pada Penerbangan Lion Air berdasarkan Ulasan Pengguna Platform Online. In *Jejaring Penelitian dan Pengabdian Masyarakat (JPPM)* (Vol. 1, Issue 1). https://ejurnal.jejaringpppm.org/index.php/jriti
9. Abdusyukur, F. (2023). Penerapan Algoritma Support Vector Machine (Svm) Untuk Klasifikasi Pencemaran Nama Baik Di Media Sosial Twitter. *KOMPUTA : Jurnal Ilmiah Komputer Dan Informatika*, *12*(1).
10. Octaviani, P. A., Wilandari, Y., & Ispriyanti, D. (2014). Penerapan Metode Klasifikasi Support Vector Machine (Svm) Pada Data Akreditasi Sekolah Dasar (Sd) Di Kabupaten Magelang. *Jurnal Gaussian*, *3*(4), 811–820. http://ejournal-s1.undip.ac.id/index.php/gaussian
11. Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. In *IJCIT (Indonesian Journal on Computer and Information Technology)* (Vol. 6, Issue 2). http://ejournal.bsi.ac.id/ejurnal/index.php/ijcit
12. Dwi Fasnuari, H. A., Yuana, H., & Chulkamdi, M. T. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus. *Antivirus : Jurnal Ilmiah Teknik Informatika*, *16*(2), 133–142. https://doi.org/10.35457/antivirus.v16i2.2445