KLASIFIKASI PENYAKIT JANTUNG MENGGUNAKAN ALGORITMA *RANDOM FOREST* DENGAN PENDEKATAN *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE)

Elya Juni Arta Sinaga, Arnita, Said Iskandar Al Idrus, Debi Yandra Niska
Ilmu Komputer, Universitas Negeri Medan
Jalan Willem Iskandar Pasar V Medan Estate
elyasinaga15@gmail.com

ABSTRAK

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia sehingga deteksi dini sangat diperlukan. Penelitian ini bertujuan untuk mengembangkan sistem klasifikasi penyakit jantung menggunakan algoritma Random Forest yang diimplementasikan dalam bentuk website. Data penelitian berasal dari RSUD Parapat dengan total 304 data, terdiri dari 210 pasien penyakit jantung dan 94 pasien bukan penyakit jantung. Karena data tidak seimbang, metode *Synthetic Minority Over-sampling Technique* (SMOTE) digunakan untuk menyeimbangkan jumlah data antar kelas. Penelitian dilakukan dengan dua skenario, yaitu klasifikasi tanpa SMOTE dan dengan SMOTE. Pada skenario tanpa SMOTE, model Random Forest menghasilkan akurasi 91,80%, presisi 92%, recall 90%, dan f1-score 92%. Hasil ini menunjukkan akurasi yang baik, tetapi model cenderung lebih fokus pada kelas penyakit jantung sehingga prediksi menjadi kurang seimbang. Pada skenario dengan SMOTE, akurasi tetap baik dan performa antar kelas menjadi lebih seimbang, yang terlihat dari peningkatan nilai akurasi 96,72%, presisi 97,61%, recall 97,61%, dan f1-score 95,80% pada kelas bukan penyakit jantung. Dengan demikian, penerapan SMOTE terbukti mampu meningkatkan kemampuan Random Forest dalam melakukan klasifikasi yang lebih adil antara dua kelas. Selain itu, website yang dibangun berhasil menampilkan hasil prediksi dengan baik sehingga dapat digunakan sebagai alat bantu dalam deteksi dini risiko penyakit jantung.

Kata kunci : *Penyakit Jantung, Klasifikasi, Random Forest, SMOTE*

1. PENDAHULUAN

Algoritma Random Forest merupakan salah satu metode pembelajaran mesin berbasis ansambel yang terdiri dari kombinasi banyak pohon keputusan. Algoritma ini bekerja dengan prinsip pemisahan biner rekursif pada struktur pohon untuk menghasilkan keputusan klasifikasi maupun regresi [1]. Keunggulan Random Forest antara lain mampu meningkatkan akurasi meskipun terdapat data yang hilang, relatif tahan terhadap outlier, serta efisien dalam pengolahan data berskala besar dengan parameter kompleks [2].

Beberapa penelitian terdahulu menunjukkan performa algoritma ini lebih stabil dibandingkan pohon tunggal. Misalnya, penelitian Suwardika & Suniantara menemukan bahwa akurasi klasifikasi menggunakan Random Forest meningkat signifikan dibandingkan CART, yaitu dari 65,41% menjadi 93,23%. Hal ini menegaskan bahwa Random Forest merupakan algoritma yang andal untuk permasalahan klasifikasi dengan kompleksitas data tinggi.

Penerapan pendekatan ini menjadi relevan dalam bidang kesehatan, khususnya penyakit jantung. Menurut data World Health Organization (2021), penyakit jantung menjadi penyebab utama kematian global dengan jumlah 17,9 juta kasus setiap tahun, termasuk di Indonesia. Faktor risiko meliputi usia, jenis kelamin, riwayat merokok, diabetes, hipertensi, kadar kolesterol, hingga gaya hidup tidak sehat [6][7].

Berdasarkan observasi di RSUD Parapat, belum didukung memiliki sistem klasifikasi berbasis website. Oleh karena itu, diperlukan suatu sistem klasifikasi berbasis Random Forest dengan pendekatan SMOTE yang mampu membantu proses deteksi dini risiko penyakit jantung. Sistem ini diharapkan dapat mempercepat diagnosis awal, mendukung keputusan dokter, serta meningkatkan efisiensi pelayanan kesehatan [8].

Namun demikian, salah satu tantangan utama dalam klasifikasi adalah ketidakseimbangan distribusi kelas pada dataset. Ketidakseimbangan data menyebabkan model lebih cenderung mengenali kelas mayoritas sehingga performanya menurun pada kelas minoritas [3]. Salah satu solusi yang banyak digunakan adalah Synthetic Minority Over-Sampling Technique (SMOTE). Teknik ini bekerja dengan membuat data sintetis melalui interpolasi sampel minoritas dengan tetangga terdekatnya, sehingga distribusi data menjadi lebih seimbang [4]. Dibandingkan metode oversampling sederhana, SMOTE dinilai lebih efektif karena mampu menambah variasi data tanpa menyebabkan overfitting.

Penelitian Lauw menunjukkan kombinasi SMOTE dengan Random Forest meningkatkan akurasi klasifikasi kanker paru dari 89,1% menjadi 94,1%. Hal tersebut membuktikan bahwa integrasi SMOTE dan Random Forest dapat mengatasi masalah ketidakseimbangan sekaligus meningkatkan kinerja model [5].

Keakuratan hasil klasifikasi penyakit sangat penting untuk memastikan diagnosis yang tepat [9].

Salah satu metode untuk mengetahui potensi seseorang terkena penyakit jantung adalah melalui teknik klasifikasi. Klasifikasi dalam *machine learning* digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang telah dipelajari. Dalam konteks ini, klasifikasi bertujuan menentukan apakah seseorang berisiko tinggi atau rendah terhadap penyakit jantung.

Penelitian ini menggunakan algoritma Random Forest dengan pendekatan SMOTE untuk menganalisis faktor-faktor seperti usia, jenis kelamin, riwayat merokok, diabetes, tekanan darah, denyut jantung, kolesterol total, dan nyeri dada untuk mengklasifikasikan apakah seseorang penyakit jantung atau bukan penyakit jantung. Dengan hasil akhir pembangunan sebuah sistem berbasis website untuk klasifikasi penyakit jantung, diharapkan mampu menjadi alat untuk prediksi awal apakah seseorang beresiko terkena penyakit jantung atau tidak.

2. TINJAUAN PUSTAKA

2.1. Penyakit Jantung

Jantung yang sehat sangat penting bagi kualitas hidup seseorang karena memungkinkan aktivitas sehari-hari berjalan normal. Sebaliknya, penyakit jantung menyebabkan gangguan fungsi tubuh bahkan kematian. Selama 20 tahun terakhir, penyakit jantung menjadi penyebab utama kematian di dunia dengan jumlah kematian global mencapai 18,6 juta per tahun dan diperkirakan meningkat hingga 24,2 juta pada 2030 [10]. Deteksi dini melalui pemeriksaan tekanan darah, kadar kolesterol, dan gula darah secara rutin sangat diperlukan. Selain itu, perubahan pola hidup sehat, olahraga teratur, berhenti merokok, serta pemeriksaan rekam jantung dapat membantu mencegah sekaligus memperlambat perkembangan penyakit jantung [11].

2.2. Klasifikasi

Klasifikasi merupakan salah satu algoritma dalam machine learning yang digunakan untuk menentukan kelas atau kategori suatu data berdasarkan informasi yang tersedia. Algoritma ini bekerja dengan cara mempelajari pola dari data berlabel, kemudian mengklasifikasikan data baru ke dalam kelas yang sesuai Dalam konteks pengolahan teks, klasifikasi membantu mengelompokkan dokumen sesuai topik sehingga mempermudah pengorganisasian informasi. Pemilihan fitur yang tepat sangat penting karena dapat meningkatkan kecepatan sekaligus efektivitas proses pembelajaran [12]. Proses klasifikasi umumnya melalui dua tahap, yaitu pelatihan model menggunakan data contoh dan penerapan model pada data baru.

2.3. *Machine Learning*

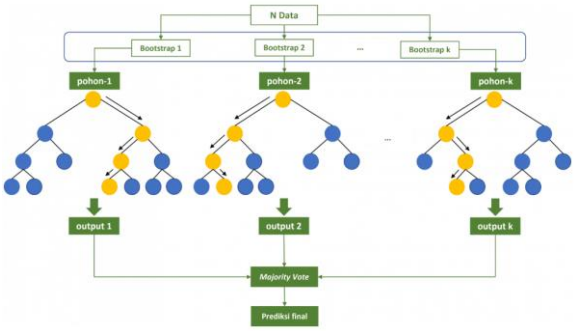
Teknologi pembelajaran mesin atau machine learning saat ini semakin populer di berbagai bidang. Istilah ini pertama kali diperkenalkan oleh Arthur Samuel sebagai kemampuan komputer untuk belajar tanpa pemrograman eksplisit. Tom Mitchell mendefinisikan machine learning sebagai kemampuan komputer belajar dari pengalaman untuk meningkatkan kinerja, sementara Murphy menekankan analisis data, identifikasi pola, dan prediksi. Secara umum, pembelajaran mesin adalah bidang ilmu komputer yang memungkinkan komputer belajar dari data masa lalu secara otomatis dan meningkatkan kinerjanya [13]. Terdapat tiga pendekatan utama:

a. *Supervised Learning*, model belajar dari data berlabel untuk prediksi atau klasifikasi.
b. *Unsupervised Learning*, model menemukan pola tersembunyi tanpa label, misalnya clustering
c. *Reinforcement Learning*, model belajar melalui interaksi dengan lingkungan berdasarkan reward atau punishment [14].

Penerapan machine learning mencakup bisnis, kesehatan, keamanan, serta pengambilan keputusan berbasis data.

2.4. *Random Forest*

Random Forest merupakan salah satu algoritma pembelajaran mesin yang memanfaatkan metode pemisahan biner berulang untuk mencapai node akhir dalam struktur pohon, dengan dasar pohon klasifikasi [15].



Gambar 1. Visualisasi random forest

Berikut adalah langkah-langkah pembentukan random forest (Breiman, 2014) :

a. Ambil dataset acak sebanyak n (jumlah total data) dari dataset latih dengan pergantian, sehingga dataset yang diambil memiliki kemungkinan untuk terambil kembali di pengambilan dataset berikutnya.
b. Dari dataset acak tersebut, buatlah sebuah pohon keputusan dengan aturan *splitting* Gini Index atau Entropy. Pohon keputusan ini akan menjadi *decision tree.*

$$Gini\ (D) = 1 - \sum_{i=1}^{c} p_i^2 \quad (1)$$

$$Entropy\ (D) = -\sum_{i=1}^{c} p_i\ log_2\ (p_i) \quad (2)$$

Dimana :

c : jumlah kelas target

p_i : proporsi jumlah sampel kelas i terhadap jumlah total sampel.

c. Ulangi langkah 1 dan 2 sebanyak n kali untuk membuat n *decision tree.* Setiap *decision tree* dibuat dengan dataset acak yang berbeda-beda.
d. Tentukan *threshold* yaitu nilai ambang yang dipakai untuk membagi data menjadi dua subset (kiri dan kanan) , dengan menggunakan nilai mean dicari dengan persamaan :

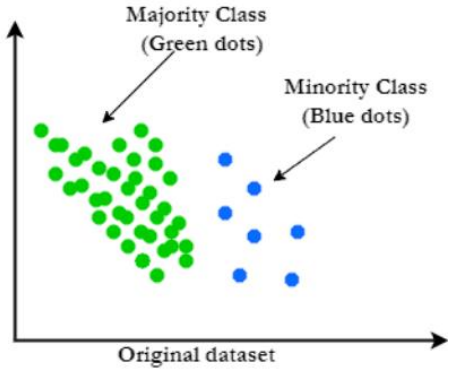
$$\bar{x} = \sum_{j=1}^{n} x_j \quad (3)$$

Dimana:
$\bar{x}$ = Rata- rata (mean)
x_j = Nilai data ke-j
n = Jumlah observasi
j = 1,2,…,n = indeks penjumlahan sampai data ke-n

e. Setiap observasi diklasifikasikan menggunakan seluruh pohon keputusan yang telah dibangun, sehingga diperoleh n hasil prediksi untuk masing-masing observasi.
f. Gabungkah hasil prediksi dari n *decision tree* untuk menghasilkan suatu prediksi akhir. Gunakan metode *voting* atau *weighted voting* untuk menentukan hasil akhir.
g. Evaluasi akurasi model pada dataset uji.

2.5. *Synthetic Minority Oversampling Technique* (SMOTE)

Ketidakseimbangan data terjadi ketika rasio objek suatu kelas data lebih banyak dibandingkan dengan kelas lain [16]. Ketidakseimbangan data dapat menimbulkan masalah pada model klasifikasi karena model cenderung lebih akurat pada kelas mayoritas sehingga cenderung mengabaikan kelas minoritas yang mungkin saja merupakan kelas yang lebih penting.



Gambar 2. Visualisasi *imbalanced* data

Imbalanced data menunjukkan distribusi kelas dalam dataset yang tidak merata, yaitu terdapat perbedaan yang signifikan antara jumlah data pada setiap kelas. Misal pada kasus data penyakit, terdapat banyak data pasien yang sehat (kelas mayoritas) dibanding dengan data pasien yang sakit (kelas minoritas) [17].

2.6. Website

Website adalah sekumpulan halaman yang saling terhubung dan dapat diakses melalui internet dengan menggunakan peramban web. Website dapat berfungsi sebagai media informasi, komunikasi, transaksi bisnis, teknologi, website telah menjadi bagian integral dalam kehidupan modern, digunakan oleh individu, perusahaan, institusi pendidikan, hingga pemerintahan untuk berbagai keperluan. Berdasarkan fungsinya, website dapat dikategorikan menjadi beberapa jenis, seperti website statis yang hanya menampilkan informasi tanpa perubahan dinamis, serta website dinamis yang memungkinkan interaksi pengguna, seperti media sosial, *e-commerce*, dan sistem berbasis web [18].

2.7. Penelitian Terdahulu

Suparyati, Emma Utami, dan Alva Hendi Muhammad melakukan penelitian terdahulu dengan judul *Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease* [19]. Pada penelitian ini ditunjukkan bagaimana perbedaan hasil yang di dapat saat random forest digunakan tanpa SMOTE dan dengan SMOTE. Penelitian ini menggunakan 2 kelas data yaitu *lumpy* dan *no lumpy*. Hasil dari penelitian ini menunjukkan pemodelan random forest tanpa SMOTE memiliki recall 96%, precision 94%, f1-score 94,9%. Sedangkan, untuk pemodelan *Random Forest* dengan SMOTE menunjukkan hasil *recall* 97.7% , *precision* 96.2% , dan *f1-score* 96.9% [19].

Penelitian lain juga dilakukan oleh Jelita et al. dengan judul "*Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest*". Penelitian ini menunjukkan hasil yang baik untuk pemodelan dengan menggunakan *Random Forest* dengan akurasi sebesar 96.42% , precision 94.44%, recall 96.66% , f1-score 95.56% [20] .

3. METODE PENELITIAN

3.1. Lokasi dan Waktu penelitian

Penelitian ini dilaksanakan di Rumah Sakit Umum Daerah (RSUD) Parapat yang beralamat di Jl.Kolonel TPR Sinaga Parapat,Simalungun, Sumatera Utara. Waktu penelitian berlangsung selama 30 hari, Penelitian ini dilaksanakan di Rumah Sakit Umum Daerah (RSUD) Parapat yang beralamat di Jl.Kolonel TPR Sinaga Parapat, Simalungun, Sumatera Utara. Waktu penelitian berlangsung selama 30 hari.

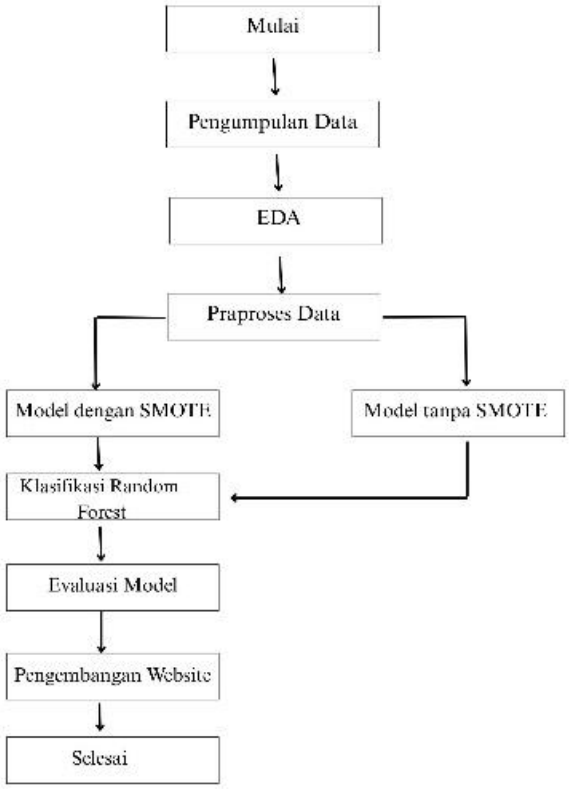
3.2. Kriteria Data Penelitian

Pengumpulan data dilakukan secara manual dengan menuliskan data yang di dapat ke dalam *Microsoft excel*. Berikut ditampilkan tabel rancangan kriteria data penelitian yang digunakan dalam penelitian ini:

Tabel 1. Kriteria data penelitian

No	Atribut	Keterangan
1	Usia	Usia dalam tahun
2	Jenis Kelamin	Laki-laki / Perempuan
3	Riwayat Merokok	ya / tidak
4	Diabetes Melitus	ya / tidak
5	Tekanan Darah	numerik
6	Kolestrol Total	numerik
7	Nyeri Dada	ada/ tidak ada
8	Label	Penyakit Jantung / Bukan Penyakit Jantung

3.3. Prosedur Penelitian



Gambar 3. Diagram alur penelitian

Pengumpulan data dalam penelitian ini dilakukan secara manual dengan mengakses dua sumber utama, yaitu buku catatan status rekam medis pasien dan website berisi hasil pemeriksaan laboratorium. Setelah data terkumpul, dilakukan Exploratory Data Analysis (EDA) untuk memahami karakteristik data secara menyeluruh. Analisis ini dilengkapi dengan visualisasi data guna memberikan gambaran yang lebih jelas dan mudah dipahami, seperti pola distribusi, hubungan antar variabel, dan potensi adanya ketidakseimbangan kelas. Tahap EDA berperan penting sebagai dasar dalam menentukan strategi praproses dan pemodelan yang akan dilakukan.

Selanjutnya, tahap praproses data mencakup pemeriksaan missing values, encoding variabel kategorik ke bentuk numerik, serta normalisasi atau standarisasi fitur numerik agar berada pada skala yang seragam. Data kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20. Dua skenario diterapkan, yaitu menggunakan teknik Synthetic Minority Oversampling Technique (SMOTE) dan tanpa SMOTE, untuk membandingkan performa model pada data seimbang dan tidak seimbang. Algoritma Random Forest digunakan sebagai metode klasifikasi, dan hasil evaluasi model diukur berdasarkan akurasi. Jika akurasi <70%, proses kembali ke tahap EDA; jika ≥70%, model dianggap layak digunakan.

3.4. Pengembangan Website

Implementasi model ke dalam website dilakukan dengan mengintegrasikan model *machine learning* yang telah dilatih ke dalam *backend* berbasis *Flask*. Model yang telah disimpan digunakan untuk menerima input dari pengguna, melakukan prediksi, dan menampilkan hasil dalam antarmuka web. *Backend* menangani permintaan dari *frontend*, sedangkan *frontend* menyediakan tampilan interaktif bagi pengguna untuk memasukkan data kesehatan.Dengan implementasi ini,

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

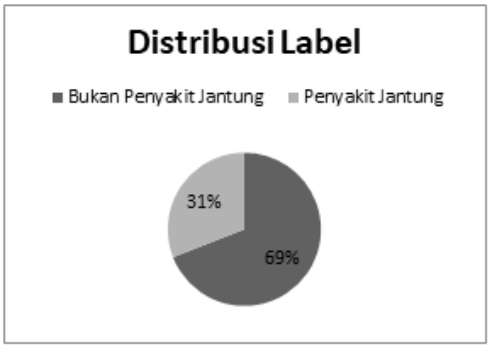
Data yang digunakan dalam penelitian ini berasal dari Rumah Sakit Umum Daerah Parapat yang beralamat di Jl Kol. TPR Sinaga, Tiga Raja, Kec. Girsang Sipangan Bolon, Kabupaten Simalungun, Sumatera Utara. Dalam waktu 30 hari, proses pengumpulan data menghasilkan 304 data catatan medis pasien.

Tabel 2. Sampel data

Usia	81	50	43	47	65
Jenis Kelamin	Lk	Lk	Lk	Pr	Pr
Riwayat merokok	ya	tidak	ya	tidak	tidak
Diabetes melitus	ya	ya	tidak	tidak	tidak
Tekanan Darah	156	167	112	132	159
Kolestrol Total	221	217	146	257	153
Nyeri Dada	ya	tidak	tidak	ya	tidak
Label	0	0	1	0	1

Keterangan :
0 = Penyakit Jantung
1 = Bukan Penyakit Jantung

2.2. Distribusi Kelas



Gambar 4. Grafik distribusi label

Dari gambar 4 dapat dilihat bahwa:

1. Bukan penyakit jantung (label 1) memiliki persentase 31% atau dengan jumlah 94 sampel.
2. Penyakit jantung (label 0) memiliki persentase 69% atau dengan jumlah 211 sampel.

Perbedaan jumlah data antara label bukan penyakit jantung dan penyakit jantung menunjukkan ketidakseimbangan data yang cukup signifikan sehingga akan digunakan teknik SMOTE untuk menyeimbangkan data.

2.3. *Exploratory Data Analisys* (EDA)

Exploratory Data Analisys (EDA) adalah proses penting untuk memeriksa dataset dengan tujuan menemukan pola, mencari hubungan atau mengidentifikasi kecenderungan yang mungkin tersembunyi di dalam data [21]. Pada tahap EDA peneliti melakukan pendekatan untuk menganalisis kumpulan data yang telah dikumpulkan, dengan tujuan untuk menemukan karakteristik utama data. Penelitian ini menggunakan teknik analisis *univariat* untuk menemukan informasi pada data.

2.4. *Cek Missing value*

Tabel 3. Cek *Missing Value*

Usia	0
Jenis Kelamin	0
Riwayat Merokok (ya/tidak)	0
Diabetes Melitus (ya/tidak)	0
Tekanan Darah Sistolik (mmHg)	0
Kolestrol total (mg/dL)	0
Nyeri Dada	0
Target	0

Pada tahap pemeriksaan data ditemukan hasil bahwa seluruh fitur dalam dataset tidak memiliki nilai yang hilang (*missing value*). Hal ini menunjukkan bahwa data yang digunakan sudah lengkap dan dapat langsung digunakan dalam proses analisis.

2.5. Data Duplikat

Langkah selanjutnya dalam proses pemeriksaan data adalah melakukan identifikasi terhadap keberadaan data duplikat (*duplicate value*).

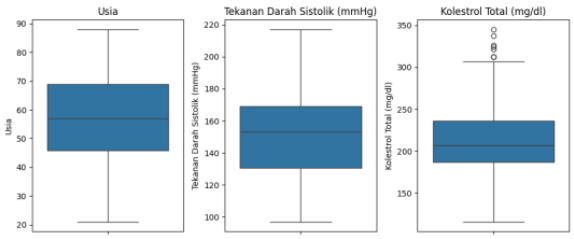
Tabel 4. Jumlah Data Duplikat

Jumlah Data Duplikat	0

Berdasarkan hasil analisis, tidak ditemukan adanya baris data yang bersifat duplikat dalam dataset yang digunakan. Hal ini menunjukkan bahwa setiap entri data bersifat untuk dan merepresentasikan informasi yang berbeda. Ketiadaan data duplikat ini menjadi indikasi bahwa proses pengumpulan data telah dilakukan dengan baik dan akurat, sehingga tidak diperlukan langkah tambahan untuk membersihkan data dari duplikasi.

2.6. Cek Outlier

Tahap akhir proses EDA dalam penelitian ini yaitu analisis *outlier*, proses identifikasi *outlier* dilakukan dengan menggunakan metode *Interquartile Range* (IQR) dimana nilai-nilai *outlier* ditentukan berdasarkan batas bawah (IQ-1.5 x IQR) dan batas atas (Q3 + 1.5 x IQR).Setiap nilai yang berada di luar rentang ini dianggap sebagai *outlier*. Pada penelitian ini terdapat 3 kolom data numerik yang akan ditunjukkan nilai *outlier* yaitu kolom usia, tekanan darah sistolik (mmHg) dan kolestrol total.



Gambar 5. Visualisasi *outlier*

Pada boxplot terlihat bahwa tidak ada outlier ditemui pada variabel usia dan tekanan darah. Namun ada beberapa titik outlier pada kriteria kolestrol total. Berdasarkan perhitungan, jumlah outlier pada variabel ini adalah 8 data (2,63%) dari total keseluruhan data.

2.7. *Cleaning Data*

Cleaning data dilakukan untuk mengatasi *outlier* yang ditemukan pada data. Penanganan outlier dilakukan dengan mengombinasikan imputasi median untuk mengganti nilai ekstrem dengan nilai tengah distribusi, serta *capping* untuk membatasi nilai agar tetap berada dalam rentang yang wajar.

Gambar 6. Visualisasi data setelah outlier diatasi

2.8. *Label Encoding*

Tabel 5. Data setelah label *encoding*

Usia	Jenis Kelamin	Riwayat Merokok	DM	Tekanan Darah	Kolestrol Total	Nyeri Dada	Label
55	1	0	1	169	265	1	0
50	1	1	1	167	217	1	0
55	1	0	1	212	199	1	0
53	1	1	0	108	152	0	1
47	0	0	0	143	188	0	1

Dalam penelitian ini, variabel kategorikal seperti jenis kelamin,riwayat merokok,diabetes melitus dan nyeri dada diubah kedalam bentuk numerik menggunakan metode *label encoding*. Sebagai contoh, kategori "Laki-laki" dan "Perempuan" pada variabel jenis kelamin diubah menjadi nilai 1 dan 0. Kategori "ya" dan "tidak" pada variabel riwayat merokok diubah menjadi nilai 1 dan 0.Kategori "ada" dan "tidak ada" dalam variavel nyeri dada diubah menjadi 1 dan 0.Dan kategori "penyakit jantung" dan "bukan penyakit jantung" dalam variabel target diubah menjadi 0 dan 1.

Tabel 6. Data setelah normalisasi data

Usia	Jenis Kelamin	Riwayat Merokok	DM	Tekanan Darah	Kolestrol Total	Nyeri Dada
0.043829	-0.78446	1.19049	1.197824	0.074581	0.074581	-1.290994
1.809118	1.274755	-0.83999	1.197824	-0.042971	2.249753	-1.290994
-1.857251	-0.784465	1.19049	-0.834847	-0.323988	-0.444495	0.774597
-2.128834	1.274755	-0.83999	-0.834847	-1.553437	-0.988288	0.774597
-2.196730	1.274755	-0.83999	-0.834847	-1.623692	-1.111877	0.774597

Standard scalling , yang juga dikenal sebagai *Z-Score Normalization* atau *standardization*, merupakan salah satu metode normalisasi data yang digunakan untuk mengubah nilai-nilai data sebelum proses pemodelan. Teknik ini diterapkan sebagai bagian dari tahap praproses dengan memanfaatkan fungsi yang tersedia pada *library Scikit-learn* [22].

2.9. *Split Data*

Proses *split* data membagi data menjadi dua bagian yaitu data *training* (uji) dan data *testing* (latih). Pembagian data training dan testing dalam penelitian ini yaitu dengan rasio 80:20 yaitu 80% (243 data) untuk data *training* dan 20% (61 data) untuk data *testing*.

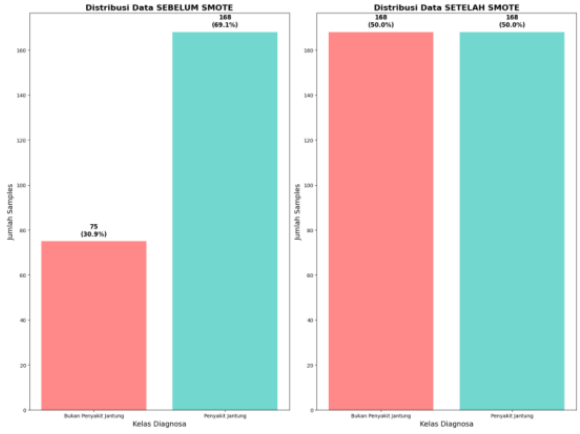
2.10. Model tanpa SMOTE

Selain menggunakan SMOTE, penelitian ini juga membangun model menggunakan data asli tanpa dilakukan oversampling. Tujuan utamanya adalah untuk membandingkan performa model antara kondisi data yang *imbalance* dengan kondisi data yang telah diseimbangkan menggunakan SMOTE. Pada data training terdapat total 243 data.

Distribusi ini menunjukkan adanya ketidakseimbangan kelas, di mana jumlah data pasien penyakit jantung lebih banyak dibandingkan pasien bukan penyakit jantung. Ketidakseimbangan ini berpotensi membuat model lebih condong (bias) dalam memprediksi kelas mayoritas (penyakit jantung) sehingga akurasi tampak tinggi, tetapi performa dalam mengenali kelas minoritas (bukan penyakit jantung) bisa rendah.

2.11. Model dengan Penerapan SMOTE

Untuk menangani ketidakseimbangan data tersebut peneliti menerapkan teknik SMOTE (*Syntetic Minority Oversampling Technique*) untuk menyeimbangkan jumlah data. SMOTE bekerja dengan membuat data sintetis baru unutk kelas minoritas, bukan hanya menduplikasi data yang sudah ada. Pada gambar 4.5. akan diperlihatkan visualisasi distribusi data sebelum dan sesudah penggunaan data terhadap data training.



Gambar 7. Distribusi data sebelum dan sesudah SMOTE

Jumlah data *training* dalam penelitian ini sebanyak 243 dengan 75 data untuk kelas "bukan penyakit jantung" dan 168 data untuk kelas "penyakit jantung". Setelah penerapan SMOTE maka data memiliki jumlah seimbang yaitu 168 data untuk masing-masing kelas.

2.12. Penerapan Random Forest

Pada penelitian ini, model *random forest* dibangun dengan parameter utama yang telah ditentukan untuk memperoleh hasil klasifikasi yang optimal. Jumlah pohon keputusan (*n_estimators* =100), sehingga model dapat menghasilkan prediksi yang stabil tanpa memerlukan waktu yang lama. Kedalaman maksimum pohon dibatasi hingga 10 (max_depth =10) agar model tidak mengalami *overfitting*.

Tabel 7. Parameter Random Forest

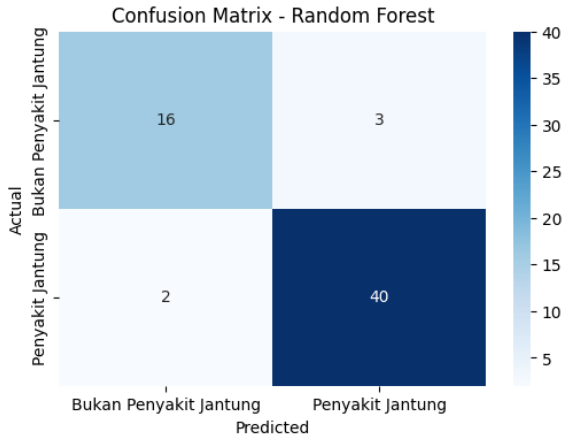
n-estimators	100
max_depth	10
min_samples_split	10
min_samples_leaf	5
max_features	sqrt

Setiap node hanya akan dilakukan pemisahan sampel apabila memiliki minimal 10 sampel (*min_samples_split* =10) dan jumlah sampel minimum pada setiap daun ditetapkan sebanyak 5

(*min_samples_leaf* = 5). Pada setiap pemisahan, fitur yang dipilih dibatasi dengan menggunakan akar kuadrat dari jumlah total fitur (*max_features = sqrt*).

2.13. Evaluasi Kinerja Model Tanpa SMOTE

Evaluasi kinerja model digunakan untuk mengetahui model sudah layak digunakan atau masih belum layak digunakan. Evaluasi kinerja model dilakukan berdasarkan *Confusion Matrix, Presisi, Recall, F1-Score.*



Gambar 8. *Confusion matrix* model tanpa SMOTE

- True Negative (TN) = 16
- False Positive (FP) = 3
- False Negative (FN) = 2
- True Positive (TP) = 40

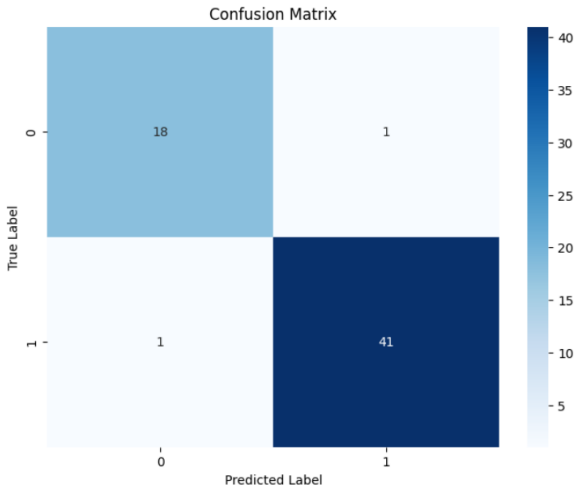
Tabel 8. Classification report model tanpa SMOTE

	Precision	***Recall***	***F1-score***
0	0.89	0.84	0.86
1	0.93	0.95	0.94
Akurasi			0.92
Macro Avg	0.91	0.90	0.90
Weighted Avg	0.92	0.92	0.92

Hasil evaluasi menunjukkan adanya indikasi bias terhadap kelas Penyakit Jantung. Hal ini terlihat dari nilai recall yang lebih tinggi pada kelas Penyakit Jantung (0.95) dibandingkan kelas Bukan Penyakit Jantung (0.84). Artinya, model lebih sering benar dalam mendeteksi pasien dengan penyakit jantung, tetapi relatif kurang optimal dalam mengenali pasien yang sebenarnya tidak sakit jantung. Kondisi ini terjadi karena data awal tidak seimbang, sehingga model cenderung "berpihak" pada kelas mayoritas.

2.14. Evaluasi Kinerja Model dengan Penerapan SMOTE

Hasil evaluasi menunjukkan akurasi model sebesar 96%. Nilai precision dan recall pada kelas Penyakit Jantung relatif tinggi dan seimbang, sehingga model dapat mengenali pasien yang sakit dengan cukup baik. Pada kelas Bukan Penyakit Jantung, nilai *precision* 0.93 dan *recall* 0.96 menunjukkan kinerja yang masih stabil meskipun sedikit lebih rendah dibandingkan kelas penyakit jantung.



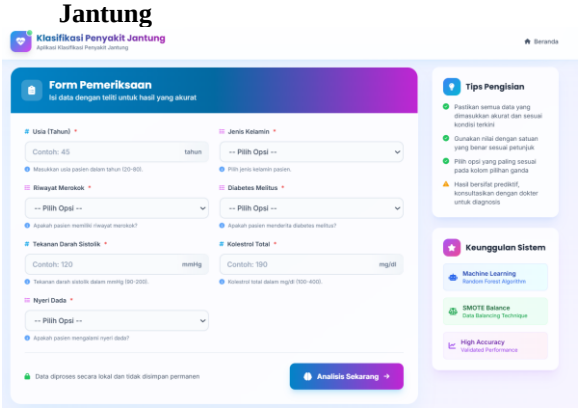
Gambar 9. *Confusion matrix* model dengan SMOTE

- True Negative (TN) = 17
- False Positive (FP) = 1
- False Negative (FN) =1
- True Positive (TP) = 39

Tabel 9. Classification report model dengan SMOTE

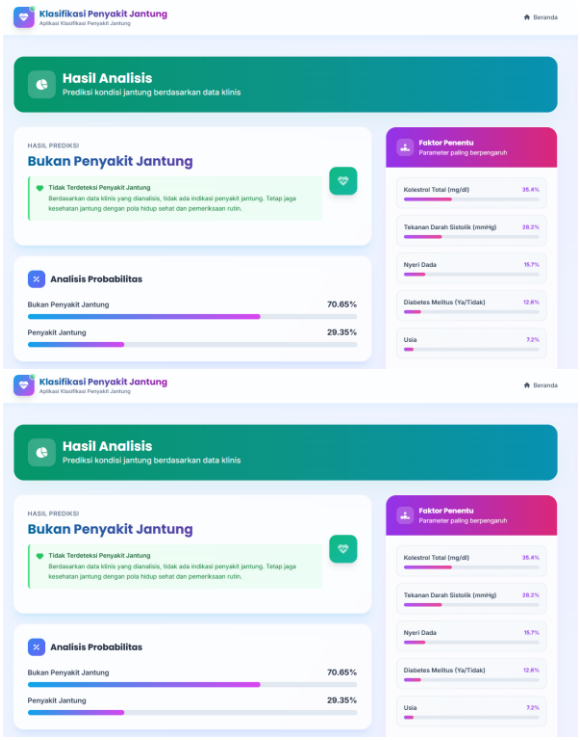
	Precision	***Recall***	***F1-score***
0	0.93	0.96	0.95
1	0.98	0.97	0.98
Akurasi			0.96
Macro Avg	0.96	0.97	0.92
Weighted Avg	0.97	0.97	0.97

2.15. Tampilan Website Klasifikasi Penyakit Jantung



Gambar 10. Tampilan awal website

Halaman utama website ini menyediakan informasi pengisian data kesehatan pasien yang sederhana dan mudah dipahami. Untuk memudahkan proses, disediakan tips pengisian setiap data yang dimasukkan agar sesuai dengan format yang benar. Terdapat pula penjelasan keunggulan sistem dimana dijelaskan bahwa keunggulan sistem ini terletak pada kemampuannya megolah data secara otomatis dengan algoritma *random forest* dan teknik SMOTE.



Gambar 11. Tampilan hasil analisis website

Halaman hasil analisis pada sistem ini menampilkan hasil klasifikasi kondisi jantung (penyakit jantung atau bukan penyakit jantung) berdasarkan data kesehatan yang telah di input pengguna. Selain memberikan hasil klasifikasi, halaman ini juga menyediakan informasi analisis probabilitas yaitu persentase kemungkinan seseorang terkena penyakit jantung atau tidak.

Gambar 12. Tampilan informasi penyakit jantung pada website.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penerapan algoritma Random Forest dengan teknik penyeimbangan data Synthetic Minority Oversampling Technique (SMOTE) mampu membangun model klasifikasi yang efektif dalam membedakan penyakit jantung dan bukan penyakit jantung. Hasil evaluasi menunjukkan bahwa skenario dengan SMOTE menghasilkan akurasi 96,72% dengan performa kelas yang seimbang, sedangkan tanpa SMOTE menghasilkan akurasi 91,80% namun cenderung lebih fokus pada kelas penyakit jantung. Website yang dibangun juga berhasil menampilkan hasil prediksi secara baik sebagai alat bantu deteksi dini. Untuk penelitian selanjutnya, disarankan menggunakan dataset yang lebih besar dan bervariasi, menambahkan fitur klinis lain seperti EKG, riwayat keluarga, serta mengeksplorasi algoritma klasifikasi dan teknik penyeimbangan data lainnya agar hasil prediksi semakin optimal**.**

DAFTAR PUSTAKA

[1] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[2] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 2, pp. 310–320, 2020, doi: 10.35957/jatisi.v7i2.289.

[3] M. S. Shahab, A. Junaidi, and A. N. Sihananto, "PLANKTON ALGORITMA HIBRIDA CONVOLUTIONAL NEURAL," vol. 12, no. 3, 2024.

[4] H. N. Irmanda and Ria Astriratma, "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 915–922, 2020, doi: 10.29207/resti.v4i5.2313.

[5] C. Michael Lauw, H. Hairani, I. Saifuddin, J. Ximenes Guterres, M. Maariful Huda, and M. Mayadi, "Combination of Smote and Random Forest Methods for Lung Cancer Classification," *Int. J. Eng. Comput. Sci. Appl.*, vol. 2, no. 2, pp. 59–64, 2023, doi: 10.30812/ijecsa.v2i2.3333.

[6] K. K. Teo and T. Rafiq, "Cardiovascular Risk Factors and Prevention: A Perspective From Developing Countries," *Can. J. Cardiol.*, vol. 37, no. 5, pp. 733–743, 2021, doi: 10.1016/j.cjca.2021.02.009.

[7] F. Handayani, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network Dalam Prediksi Penyakit Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, p. 329, 2021, doi: 10.26418/jp.v7i3.48053.

[8] A. Pirzada *et al.*, "Risk Factors for Cardiovascular Disease: Knowledge Gained from the Hispanic Community Health Study/Study of Latinos," *Curr. Atheroscler. Rep.*, vol. 25, no. 11, pp. 785–793, 2023, doi: 10.1007/s11883-023-01152-9.

[9] W. S. Dharmawan, "Komparasi Algoritma Klasifikasi Svm-Pso Dan C4.5-Pso Dalam

Prediksi Penyakit Jantung," *I N F O R M a T I K a*, vol. 13, no. 2, p. 31, 2022, doi: 10.36723/juri.v13i2.301.

[10] D. W. Jones *et al.*, *2025 AHA/ACC/AANP/AAPA/ABC/ACCP/ACPM/AGS/AMA/ASPC/NMA/PCNA/SGIM Guideline for the Prevention, Detection, Evaluation and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Joint Committee* . 2025. doi: 10.1161/CIR.0000000000001356.

[11] W. Do Toka, F. Aulia Tamsil, and L. Armaijn, "Deteksi Dini Penyakit Jantung Koroner melalui Pemeriksaan Rekam Jantung pada Sivitas Akademika Universitas Khairun," *J. Kreat. Pengabdi. Kpd. Masy.*, vol. 7, no. 6, pp. 2447–2454, 2024, doi: 10.33024/jkpm.v7i6.14453.

[12] F. Istighfarizky, N. A. Sanjaya ER, I. M. Widiartha, L. G. Astuti, I. G. N. A. C. Putra, and I. K. G. Suhartana, "Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 11, no. 1, p. 167, 2022, doi: 10.24843/jlk.2022.v11.i01.p18.

[13] Verdy and Ery Hartati, "Klasifikasi Penyakit Mata Menggunakan Convolutional Neural Network Model Resnet-50," *J. Rekayasa Sist. Inf. dan Teknol.*, vol. 1, no. 3, pp. 199–206, 2024, doi: 10.59407/jrsit.v1i3.529.

[14] F. Marpaung, F. Aulia, and R. C. Nabila, *Computer Vision Dan Pengolahan Citra Digital*. 2022. [Online]. Available: www.pustakaaksara.co.id

[15] Riska Chairunisa, Adiwijaya, and Widi Astuti, "Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 805–812, 2020, doi: 10.29207/resti.v4i5.2083.

[16] F. Baharuddin and A. Tjahyanto, "Peningkatan Performa Klasifikasi Machine Learning Melalui Perbandingan Metode Machine Learning dan Peningkatan Dataset," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, pp. 25–31, 2022, doi: 10.32736/sisfokom.v11i1.1337.

[17] H. Akbar and W. K. Sanjaya, "Kajian Performa Metode Class Weight Random Forest pada Klasifikasi Imbalance Data Kelas Curah Hujan," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 3, no. 1, 2023, doi: 10.20885/snati.v3i1.30.

[18] B. A. Nugroho *et al.*, "Pengembangan Sistem Monitoring Admin Dan Pendaftaran User Pada UINSAFOOD Berbasis Web Aplikasi," vol. 12, no. 1, 2025.

[19] S. Suparyati, Emma Utami, and Alva Hendi Muhammad, "Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 555–562, 2022, doi: 10.29207/resti.v6i4.4147.

[20] J. Asian, M. D. Rosita, and T. Mantoro, "Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods," *J. Online Inform.*, vol. 7, no. 1, pp. 132–141, 2022, doi: 10.15575/join.v7i1.900.

[21] E. Febie, *Dasar Eksploratory Data Analisis (EDA)*. 2024. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=n7ofEQAAQBAJ&oi=fnd&pg=PR1&dq=teknik+Exploratory+Data+Analysis+(EDA)&ots=2R4NA3kEui&sig=n1sse6BtvC6JK3fgdDBVK8KmWNs&redir_esc=y#v=onepage&q=teknik Exploratory Data Analysis (EDA)&f=false

[22] A. Kurniatul Hidayah and R. Eka Putra, "Penerapan Metode Long Short Term Memory untuk Memprediksi Harga Beras di Indonesia," *J. Informatics Comput. Sci.*, vol. 6, no. 03, pp. 720–729, 2025, doi: 10.26740/jinacs.v6n03.p720-729.