DOI: 10.22146/jnteti.v14i2.17157

# *Ensemble Voting Classifier* Berbasis Multi-Algoritma dan Metode SMOTE untuk Klasifikasi Penyakit Jantung

**Dede Kurniadi[1], Asri Indah Pertiwi[1], Asri Mulyani[1]**

[1] Program Studi Teknik Informatika, Jurusan Ilmu Komputer, Institut Teknologi Garut, Garut, Jawa Barat 44151, Indonesia

[Diserahkan: 19 November 2024, Direvisi: 23 Januari 2025, Diterima: 17 April 2025]
Penulis Korespondensi: Dede Kurniadi (dede.kurniadi@itg.ac.id)

**INTISARI** — Jantung adalah organ penting tubuh yang berfungsi untuk memompa darah. Gangguan pada jantung dapat mengganggu sirkulasi darah dalam tubuh dan menjadi salah satu penyebab utama kematian global. Menurut laporan World Health Organization (WHO) tahun 2021, jumlah kematian akibat penyakit jantung mencapai angka yang signifikan. Sementara itu, prevalensi penyakit jantung di Indonesia mencapai 1,5%. Maka, diperlukan upaya pencegahan dan deteksi dini penyakit jantung dengan memanfaatkan teknologi pemelajaran mesin. Penelitian ini bertujuan untuk mengembangkan model klasifikasi penyakit jantung menggunakan algoritma *naïve Bayes* dan *random forest* melalui pendekatan *ensemble voting classifier*. Data yang digunakan berasal dari Kaggle, yang terdiri atas 1.000 *record* dengan 14 variabel, satu di antaranya sebagai target klasifikasi. Ketidakseimbangan data diatasi dengan teknik *synthetic minority oversampling technique* (SMOTE), sedangkan seleksi fitur dikonsultasikan dengan dokter spesialis jantung untuk memastikan relevansi klinis. Model dilatih menggunakan algoritma *naïve Bayes*, *random forest*, serta kombinasi keduanya melalui metode *ensemble voting classifier*, berbeda dengan penelitian sebelumnya yang hanya membandingkan beberapa algoritma untuk menentukan akurasi tertinggi. Hasil pengujian menunjukkan bahwa model yang dilatih dengan *ensemble voting classifier* memiliki kinerja terbaik, dengan akurasi 98,28%, presisi 98,41%, *recall* 98,41%, dan *F*1-*score* 98,41%. Penelitian ini membuktikan bahwa penerapan metode *ensemble voting classifier* mampu mencapai akurasi yang lebih baik dibandingkan penggunaan algoritma secara terpisah. Model ini termasuk kategori *excellent classification* dan diharapkan dapat berkontribusi dalam bidang kedokteran serta mendukung praktisi medis dalam pengembangan sistem pendukung keputusan untuk diagnosis penyakit jantung.

**KATA KUNCI** — *Ensemble Voting Classifier*, Klasifikasi Penyakit Jantung, *Naïve Bayes*, *Random Forest*.

## I. PENDAHULUAN

Di Indonesia, tantangan kesehatan makin beragam, dengan berbagai penyakit yang menyerang masyarakat, mulai dari yang ringan hingga yang berpotensi membahayakan nyawa [1]. Fokus penting saat ini terarah pada penyakit jantung, sebuah kondisi yang menjadi sorotan besar. Sebagai organ yang memompa darah, jantung memiliki peran penting dalam menyalurkan oksigen dan nutrisi ke seluruh tubuh. Gangguan pada fungsi jantung dapat menghambat aliran darah di dalam tubuh [2]. Penyakit kardiovaskular merupakan masalah kesehatan yang makin memburuk dari tahun ke tahun, menjadi penyebab utama kematian dan menunjukkan peningkatan keparahan secara global [3].

Menurut data World Health Organization (WHO) tahun 2019, sebanyak 17,9 juta kematian di seluruh dunia setiap tahunnya disebabkan oleh penyakit kardiovaskular, termasuk penyakit jantung [4]. Berdasarkan Laporan Nasional Riskesdas 2018, prevalensi penyakit jantung di Indonesia yang didiagnosis oleh dokter mencapai 1,5% pada semua kelompok usia [5]. Oleh karena itu, perlu adanya tindakan pencegahan, deteksi dini, dan penanganan penyakit jantung yang lebih optimal. Deteksi dini berperan penting dalam mengelola risiko penyakit jantung, sehingga memperbesar kemungkinan seseorang hidup lebih sehat dan lebih lama [6]. Deteksi dini penyakit jantung sangat dianjurkan, terutama bagi individu dengan usia di atas 40 tahun serta orang-orang yang memiliki faktor risiko tinggi, seperti pengidap hipertensi atau diabetes [7].

Kemajuan teknologi memungkinkan penanganan masalah ini dengan bantuan sistem yang mendukung tenaga medis dalam pencegahan dan deteksi dini penyakit jantung, melalui penerapan pemelajaran mesin untuk mengembangkan model yang dapat mengidentifikasi penyakit jantung. Kemampuan pemelajaran mesin untuk menganalisis data dengan cepat dan tepat juga berpotensi menghemat waktu dan biaya dalam proses diagnosis serta menurunkan risiko kesalahan manusia [8]. Dengan demikian, penggunaan teknologi pemelajaran mesin ini dapat meningkatkan efektivitas dalam deteksi dini dan penanganan penyakit jantung, mengurangi biaya terkait, serta memperbaiki kualitas hidup masyarakat.

## II. PENELITIAN TERKAIT

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan teknologi pemelajaran mesin dalam mendeteksi dan menganalisis penyakit jantung. Studi-studi ini membandingkan berbagai algoritma, seperti *naïve Bayes* dan *random forest*, untuk meningkatkan akurasi dan keandalan prediksi. Sebuah penelitian mencari algoritma yang sesuai dengan data yang digunakan dalam klasifikasi penyakit jantung dengan algorima *decision tree*, *naïve Bayes*, dan *random forest classifier* [9]. Penelitian ini menggunakan *dataset* Heart Attack Analysis & Prediction yang diambil dari Kaggle, terdiri atas 918 baris data dan 12 atribut, dengan 11 atribut di antaranya sebagai masukan (*age*, *sex*, *chest pain type*, *resting BP*, *cholesterol*, *fasting BS*, *resting ECG*, *max HR*, *exercise angina*, *oldpeak*, *ST slope*) dan 1 atribut sebagai keluaran (*heart*

*disease*). Pada tahap prapemrosesan data dalam penelitian ini dilakukan seleksi fitur, penanganan *outlier*, dan penanganan label. Untuk pengujian, *dataset* dibagi menjadi dua bagian, dengan rasio 80:20, yaitu 80% data digunakan sebagai data latih dan 20% data digunakan sebagai data uji. Dalam hasil pengujian seluruh model, *random forest classifier* tampil sebagai yang paling unggul, dengan skor 0,868 dari *grid search hyperparameter tuning* dan 0,852 dari *random search*.

Referensi [10] menentukan model terbaik dalam menganalisis penyakit jantung dengan menerapkan berbagai algoritma. Dengan menggunakan beberapa metode, penelitian ini berfokus pada peningkatan kinerja model untuk menghasilkan prediksi yang lebih akurat dan mendukung proses diagnosis yang lebih tepat. Algorima yang digunakan adalah *random forest*, Algoritma C45, *logistic regression*, dan *support vector machine* (SVM). Data yang digunakan sebanyak 300.000 data dengan variabel yang digunakan sejumlah 18 dan 1 target. Selain itu, kinerja klasifikasi dibandingkan menggunakan metode *synthetic minority oversampling technique* (SMOTE) dan *adaptive synthetic* (ADASYN). Metode ekstraksi fitur juga digunakan dalam mengidentifikasi variabel yang paling berpengaruh. Hasil menunjukkan bahwa algoritma *random forest* adalah algoritma terbaik, dengan akurasi awal 90,71% dan meningkat menjadi 94,54% setelah penerapan teknik *oversampling*.

Selanjutnya, penelitian lain menganalisis perbandingan model algoritma *decision tree*, *naïve Bayes*, dan *random forest* dalam klasifikasi penyakit jantung [11]. Sebanyak 319.795 data yang digunakan mengalami ketidakseimbangan data, sehingga dilakukan pengolahan dengan teknik *random undersampling* dan menghasilkan kumpulan data sebanyak 54.746. *Dataset* yang telah diproses kemudian dibagi menjadi dua bagian untuk digunakan dalam pelatihan dan pengujian model, dengan perbandingan 80% (43.796 data latih) dan 20% (10.950 data uji). Hasil penelitian menunjukkan akurasi terbaik pada algoritma *random forest*, dengan nilai akurasi 75%, presisi 77%, *recall* 74%, dan *F*1-*score* 76%.

Referensi [12] melakukan perbandingan dengan tujuh metode klasifikasi pemelajaran mesin, yaitu *naïve Bayes*, *k-nearest neighbor* (KNN), *random forest*, *logistic regression*, SVM, *decision tree*, dan *adaptive boosting* (AdaBoost). Data yang digunakan sebanyak 297 data bersih dari keseluruhan data yang terkumpul, yaitu 303 *record* (6 data memiliki variabel yang tidak lengkap) dengan 14 variabel. *Dataset* ini merupakan data Cleveland Clinic Foundation yang diperoleh dari UCI Machine Learning Repository. Hasil ekperimen menunjukkan bahwa algoritma *naïve Bayes* memberikan kinerja terbaik, dengan tingkat akurasi sebesar 84,67%.

Penelitian lainnya melibatkan pendekatan perbandingan antara algoritma *naïve Bayes*, *random forest*, dan KNN dalam klasifikasi penyakit jantung [13]. *Dataset* yang digunakan terdiri atas 304 entri data yang dipisahkan menjadi 294 data untuk proses pelatihan dan 10 data lainnya digunakan untuk pengujian. Sumber data ini dari Cleveland Clinic Foundation dan telah diadopsi oleh Hungarian Institute of Cardiology di Budapest. Hasilnya menunjukkan bahwa algoritma *naïve Bayes* adalah algoritma yang memberikan akurasi tertinggi, yaitu *area under the curve* (AUC) 0,91, *calibration* (CA) 0,84, *F*1-*score* 0,84, presisi 0,839, dan *recall* 0,84.

Berdasarkan penelitian-penelitian tersebut, diketahui bahwa algoritma *naïve Bayes* dan *random forest* menjadi algoritma terbaik dalam kasus klasifikasi penyakit jantung. Kedua algoritma tersebut telah menunjukkan kinerja yang konsisten dan akurasi yang tinggi dalam mendeteksi penyakit jantung. Selain itu, penelitian-penelitian sebelumnya hanya membandingkan beberapa algoritma untuk mendapatkan algoritma dengan akurasi tertinggi. Maka, pada penelitian ini dilakukan pemodelan untuk klasifikasi penyakit jantung dengan menggabungkan kedua algoritma tersebut (*naïve Bayes* dan *random forest*) dan memanfaatkan kekuatan dari masing-masing algoritma menggunakan pendekatan *ensemble voting classifier*.

### A. PENYAKIT JANTUNG

Penyakit jantung adalah suatu kondisi medis yang melibatkan berbagai jenis gangguan pada jantung. Gangguan tersebut dapat terjadi pada pembuluh darah jantung, katup jantung, atau bahkan otot jantung itu sendiri. Selain itu, penyakit jantung dapat juga disebabkan oleh berbagai faktor lain, termasuk infeksi dan kelainan bawaan [14]. Dengan beragam bentuk gangguan yang mungkin terjadi, penyakit jantung tergolong sebagai salah satu masalah kesehatan yang kompleks dan seringkali memerlukan penanganan medis yang tepat.

Terdapat sejumlah faktor yang berperan dalam menyebabkan pasien dengan penyakit jantung mengalami kondisi yang parah dan memerlukan perawatan medis. Salah satu faktor yang signifikan adalah keberadaan faktor risiko kardiovaskular serta kondisi penyakit lain yang dialami pasien, yang sangat berperan dalam menentukan risiko penyakit jantung. Faktor risiko kardiovaskular dapat diklasifikasikan ke dalam dua kelompok utama, yaitu faktor yang tidak dapat diubah dan faktor yang dapat diubah. Faktor risiko yang tidak dapat diubah meliputi usia, jenis kelamin, serta riwayat genetik, yang semuanya berpengaruh terhadap kemungkinan seseorang mengembangkan penyakit jantung. Sementara itu, faktor risiko yang dapat diubah adalah variabel yang dapat diubah atau dikelola melalui tindakan tertentu. Contohnya adalah tekanan darah tinggi (hipertensi), kadar kolesterol tinggi dalam darah, kebiasaan merokok, diabetes, serta kelebihan berat badan atau obesitas [2].

### B. PEMELAJARAN MESIN

Pemelajaran mesin merupakan subdisiplin dalam bidang kecerdasan buatan yang mengizinkan sistem untuk mempelajari pola dari data secara alami dan otomatis serta meningkatkan kemampuannya melalui pemelajaran dari pengalaman, tanpa memerlukan pemrograman langsung [15]. Fokus pemelajaran mesin adalah mengembangkan program komputer yang dapat mengakses data dan memperoleh pemelajaran dari data tersebut.

### C. KLASIFIKASI

Klasifikasi merupakan proses penting dalam analisis data, yang bertujuan mengidentifikasi pola-pola yang ada dalam *dataset* dan memisahkan data ke dalam beberapa kelas yang berbeda. Proses ini melibatkan pembentukan model atau fungsi yang mampu mempelajari pola-pola tersebut dari data yang telah diberi label kelas. Dengan menggunakan berbagai algoritma dan teknik, klasifikasi memungkinkan untuk memprediksi kelas dari data baru yang belum memiliki label kelas. Model klasifikasi dapat digunakan untuk berbagai bidang, seperti pengenalan pola, analisis citra, dan biomedis. Dengan memahami pola-pola yang terdapat dalam data serta mengklasifikasikannya dengan akurat, dapat dibuat keputusan yang unggul dan diperoleh wawasan yang berharga dari data tersebut [16].

### D. NAÏVE BAYES

*Naïve Bayes* adalah pendekatan dalam klasifikasi yang berbasis pada Teorema Bayes, suatu prinsip yang memanfaatkan konsep probabilitas dan statistik yang ditemukan oleh ilmuwan Inggris, Thomas Bayes. Teorema Bayes memungkinkan dilakukannya perkiraan probabilitas kejadian di masa yang akan datang berdasarkan pengalaman masa lalu [17].

Algoritma *naïve Bayes*, sebagai salah satu algoritma dari pemelajaran mesin, menjadi pilihan umum dalam menyelesaikan masalah klasifikasi, terutama dalam konteks klasifikasi teks yang melibatkan data pelatihan dengan dimensi yang tinggi. *Naïve Bayes* memperkirakan probabilitas kelas target berdasarkan fitur-fitur yang diamati, dengan asumsi bahwa setiap fitur independen satu sama lain, tetapi dapat memberikan klasifikasi yang cukup akurat [18].

Berdasarkan [19], rumus yang dapat digunakan dalam implementasi *naïve Bayes* dituliskan dalam (1).

$$P\,(Ci|D) = \frac{P\,(D|Ci) \times P\,(Ci)}{P\,(D)} \quad (1)$$

dengan $P\,(Ci|D)$ merupakan *conditional probability* kategori terhadap dokumen, $P\,(D|Ci)$ menunjukkan *conditional probability* dokumen terhadap kategori, $P(Cj)$ merepresentasikan probabilitas kategori atau teks yang akan diklasifikasi, dan $P(D)$ menunjukkan probabilitas dokumen atau data.

### E. RANDOM FOREST

*Random forest* adalah suatu metode pemelajaran mesin yang berakar pada pemelajaran mesin terbimbing dengan menerapkan konsep *decision tree* secara berulang, membentuk sekumpulan *decision tree* yang disebut sebagai *forest* atau hutan [20]. *Random forest* adalah evolusi dari algoritma *classification and regression trees* (CART), yang juga terkait dengan teknik *decision tree*. Yang membedakan *random forest* dari metode CART adalah penggunaan metode *bootstrap aggregating* (*bagging*) dan seleksi fitur yang dilakukan secara acak, yang sering disebut sebagai *random feature selection* [21]. Dalam implementasinya, *random forest* menggunakan teknik *bagging* untuk membuat sejumlah besar *decision tree* secara paralel. Masing-masing *decision tree* dibangun dengan memanfaatkan sampel data yang diambil secara acak dengan penggantian dari *dataset* yang tersedia. Lebih lanjut lagi, hanya sebagian kecil dari fitur yang dipertimbangkan dalam pembuatan keputusan di setiap *decision tree* dan fitur tersebut dipilih secara acak. Persamaan (2) digunakan dalam implementasi *random forest*.

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

dengan *pi* adalah frekuensi relatif dan *c* menunjukkan jumlah kelas. Persamaan (2) merupakan persamaan Gini yang digunakan untuk menentukan keputusan tentang pemisahan *node* dalam cabang *decision tree*.

### F. ENSEMBLE VOTING CLASSIFIER

*Ensemble voting classifier* merupakan suatu pendekatan pemelajaran mesin yang mengintegrasikan beberapa model pemelajaran untuk meningkatkan kinerja prediksi. Konsep ini melibatkan penggunaan beberapa algoritma pemelajaran untuk membangun sejumlah model independen, yang kemudian digabungkan bersama untuk menghasilkan prediksi akhir [18]. Dalam proses *ensemble voting classifier*, setiap model memberikan suara atau kontribusi berdasarkan prediksi yang dibuatnya. Suara dari setiap model kemudian dijumlahkan atau diambil rata-rata dan kelas yang mendapat suara terbanyak dipilih sebagai prediksi akhir.

## III. METODOLOGI

Metode yang digunakan dalam penelitian ini adalah *machine learning lifecycle* (MLLC). MLLC merujuk pada serangkaian langkah atau proses yang dimulai dari pengumpulan data hingga model yang dihasilkan siap untuk digunakan [22]. MLLC ini berlangsung secara progresif, maju ke depan, dan bisa berulang atau iteratif. Setiap iterasi bertujuan untuk meningkatkan akurasi dan kinerja model yang dikembangkan. Penelitian ini membangun model pemelajaran mesin dengan menggunakan bahasa pemrograman Python pada *tools* Jupyter Notebook versi 6.4.12 yang ada dalam aplikasi Anaconda. Tahapan yang dilakukan meliputi akuisisi data, prapemrosesan data, serta pelatihan model dan evaluasi.

### A. AKUISISI DATA

Akuisisi data dilakukan melalui dua aktivitas. Yang pertama adalah pengumpulan data. Tahap ini melakukan pencarian dan pengumpulan data mentah terkait penyakit jantung. Tahap kedua adalah analisis data yang dilakukan untuk memahami karakteristik data yang akan diproses dan digunakan dalam pemodelan klasifikasi pemelajaran mesin.

### B. PRAPEMROSESAN DATA

Tahap ini merupakan tahap yang sangat penting dalam memastikan kualitas dan keandalan data yang digunakan untuk melatih model. Terdapat tiga aktivitas yang dilakukan pada tahap ini. Aktivitas pertama adalah penyeimbangan data, yang dilakukan untuk menangani data yang tidak seimbang. Teknik yang digunakan untuk penyeimbangan data ini adalah SMOTE. Teknik ini memastikan distribusi data yang lebih merata antara kelas minoritas dan mayoritas [23]. Aktivitas kedua adalah rekayasa fitur (*feature engineering*). Pada tahap ini dilakukan wawancara terhadap dokter spesialis jantung untuk memastikan atribut-atribut yang dapat digunakan pada pemodelan. Terakhir, pemisahan data, yang dilakukan untuk membagi *dataset* menjadi subset berbeda untuk pelatihan dan pengujian model.

### C. PELATIHAN MODEL DAN EVALUASI

Pada tahap ini dilakukan aktivitas pembuatan model klasifikasi penyakit jantung. Penelitian ini menerapkan kombinasi algoritma *naïve Bayes* dan *random forest* melalui metode *ensemble voting classifier*. Oleh karena itu, terdapat beberapa aktivitas yang dilakukan, diawali dengan membangun model *naïve Bayes*. Dalam mengimplementasikan model tersebut, digunakan data yang sebelumnya telah disiapkan dan dipilih. Persamaan (1) digunakan dalam implementasi *naïve Bayes* [19]. Selanjutnya, dibangun model *random forest*. Rumus indeks Gini diterapkan untuk menentukan keputusan tentang pemisahan *node* dalam cabang *decision tree*. Persamaan (2) merupakan rumus indeks Gini yang digunakan dalam implementasi *random forest*. Kemudian, dilakukan implementasi *ensemble voting classifier*. Setiap model klasifikasi berupaya secara optimal sambil menyeimbangkan dampak kelemahan individu di berbagai bagian dengan *dataset*. Teknik dalam *ensemble voting classifier* memanfaatkan kekuatan dari masing-masing model, sehingga dapat menghasilkan prediksi yang lebih akurat dan stabil [24]. Berbagai metode *ensemble* diperkenalkan karena terbukti memberikan kinerja unggul pada *dataset* penyakit jantung yang berbeda [25]. Terakhir, semua model yang telah dibuat dan

dibangun kemudian dievaluasi. Pada tahap ini digunakan model evaluasi berupa akurasi, presisi, *recall*, dan *F*1-*score*.

## IV. HASIL

### A. AKUISISI DATA

Pada tahap akuisisi data dilakukan dua aktivitas, yaitu pengumpulan data dan analisis data.

#### 1) PENGUMPULAN DATA

Dalam tahap ini dilakukan pengumpulan data mengenai *dataset* penyakit jantung yang diperoleh secara *online* melalui salah satu *website* kumpulan *dataset*, yaitu Kaggle. Kumpulan data tersebut dapat dilihat pada kaggle.com/datasets/jocelyndumlao/cardiovaskular-disease-dataset [26], *Dataset* tersebut merupakan data yang digunakan dalam penelitian ini. *Dataset* tersebut memiliki 1.000 data dengan 14 variabel. Sebanyak 13 variabel merupakan variabel bebas dan 1 variabel merupakan variabel terikat atau target. Variabel target tersebut merupakan target biner dan 13 variabel lainnya bertipe data kategorial dan numerik, seperti tersaji dalam Tabel I.

#### 2) ANALISIS DATA

Pada tahap ini analisis data dilakukan untuk mengetahui karakteristik data yang sudah dikumpulkan sebelumnya. Terdapat beberapa aktivitas yang dilakukan dalam menganalisis data. Yang pertama, untuk mengetahui jumlah nilai yang kosong atau hilang pada setiap variabel dalam *dataset*, dilakukan analisis *missing value*. *Dataset* tersebut disimpan dalam variabel data dan dianalisis menggunakan fungsi dari *library* Pandas. Dalam menganalisis *missing value* di dalam data, digunakan fungsi df.isnull().sum(). Hasil analisis *missing value* adalah *dataset* yang telah dikumpulkan tidak memiliki nilai yang hilang, sehingga *dataset* tersebut dapat digunakan untuk analisis selanjutnya.

Selanjutnya, dilakukan analisis duplikasi data. Dalam menganalisis duplikasi data pada *dataset*, digunakan fungsi df.duplicated().sum(). Hasil dari analisis duplikasi data diperlihatkan pada Gambar 1. Tampak pada gambar bahwa *dataset* tidak memiliki duplikasi data pada semua variabelnya. Maka, *dataset* dapat dianalisis lebih lanjut melalui tahapan *data outlier*. Analisis terhadap *outlier* dilakukan untuk menemukan dan menangani data yang secara signifikan berbeda dari mayoritas. Analisis *outlier* dilakukan dengan memanfaatkan *library* Seaborn dengan fungsi df[columns_list].boxplot(figsize=(12, 6)). Visualisasi informasi *outlier* dilakukan menggunakan *boxplot* hasil pengolahan program, seperti ditunjukkan pada Gambar 2.

Berdasarkan Gambar 2, terlihat bahwa dari sebaran data pada setiap atribut tidak terdapat *outlier*. Maka, dapat dilakukan analisis selanjutnya, yaitu distribusi kelas. Analisis distribusi kelas pada label *dataset* penyakit jantung dilakukan untuk mengetahui proporsi masing-masing kelas. *Dataset* ini memiliki dua kelas biner, yaitu kelas 0, yang menunjukkan tidak adanya penyakit jantung; dan kelas 1, yang menunjukkan adanya penyakit jantung. Jumlah kelas pada *dataset* ini didistribusikan menggunakan fungsi df['target'].value_counts().tolist(). Distribusi kelas data pada *dataset* ditunjukkan pada Tabel II.

Pada Tabel II terlihat bahwa jumlah data pada kedua kelas berbeda. Jumlah data pada kelas 0 (tidak ada penyakit jantung) sebanyak 420 data, sedangkan jumlah data kelas 1 (ada penyakit jantung) adalah 580 data. Hal tersebut menunjukkan adanya ketidakseimbangan data.

TABEL I
VARIABEL DALAM *DATASET*

| No. | Variabel | Tipe Data |
|---|---|---|
| 1. | *patientid* | Numerik |
| 2. | *age* | Numerik |
| 3. | *gender* | Kategorial |
| 4. | *restingbp* | Numerik |
| 5. | *serumcholestrol* | Numerik |
| 6. | *fastingbloodsugar* | Kategorial |
| 7. | *chestpain* | Kategorial |
| 8. | *restingelectro* | Kategorial |
| 9. | *maxheartrate* | Numerik |
| 10. | *exerciseangina* | Kategorial |
| 11. | *oldpeak* | Numerik |
| 12. | *slope* | Kategorial |
| 13. | *noofmajorvessels* | Kategorial |
| 14. | *classification (target)* | Kategorial |

```
Jumlah baris duplikat: 0

Baris duplikat:
 Empty DataFrame
Columns: [patientid, age, gender, chestpain, restingBP, serumchole
strol, fastingbloodsugar, restingrelectro, maxheartrate, exercisea
ngia, oldpeak, slope, noofmajorvessels, target]
Index: []
```

**Gambar 1.** Hasil analisis duplikasi data.

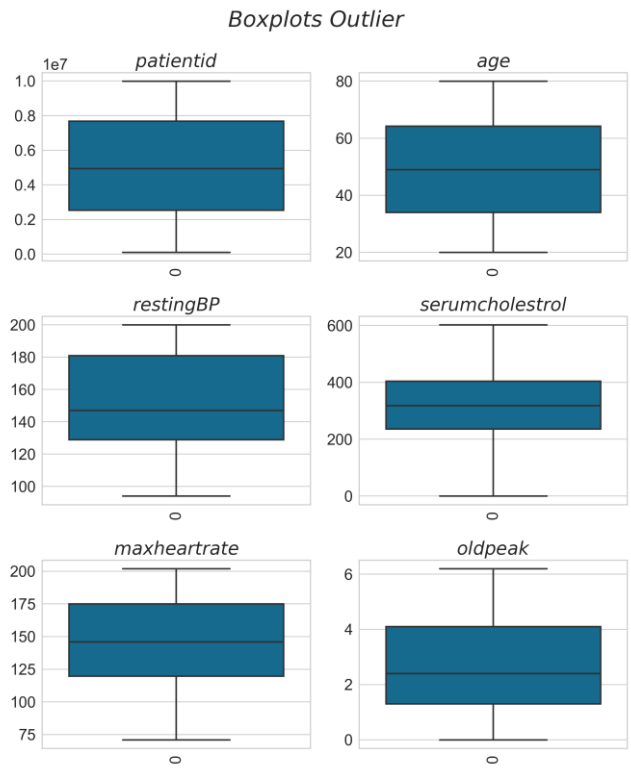
Dari beberapa analisis yang telah dilakukan pada *dataset* penyakit jantung, ditemukan satu masalah, yaitu adanya ketidakseimbangan data. Ketidakseimbangan dalam data dapat memiliki pengaruh besar terhadap hasil klasifikasi [27]. Maka, masalah ini perlu ditangani agar tidak memengaruhi kinerja model yang dibuat. Penanganan masalah ketidakseimbangan data ini dilakukan pada tahap selanjutnya, yaitu tahap prapemrosesan data.

### B. PRAPEMROSESAN DATA

Pada tahap prapemrosesan data terdapat tiga aktivitas yang dilakukan. Berikut merupakan detail dari setiap aktivitas tersebut.

#### 1) PENYEIMBANGAN DATA *(SMOTE)*

Berdasarkan Tabel II, diketahui bahwa terdapat data yang tidak seimbang, sehingga perlu dilakukan proses penyeimbangan data dengan menerapkan teknik SMOTE. Teknik SMOTE ini melakukan *over-sampling* dengan membuat sampel sintetis untuk kelas minoritas, berdasarkan analisis dan interpolasi dari data yang sudah ada guna memperbaiki distribusi data [23]. Namun, jika data tidak terlalu kompleks, kekurangan dari teknik SMOTE ini dapat memperkenalkan bias sintetik yang tidak ada dalam data asli, sehingga dapat memengaruhi kinerja model terhadap data nyata, terutama jika model lebih banyak belajar dari pola sintetik daripada pola asli. Langkah pertama dalam penerapan SMOTE adalah memvisualisasikan distribusi kelas target dengan menggunakan fungsi countplot dari *library* Seaborn. Visualisasi ini akan membantu dalam memahami proporsi data pada kelas minoritas dan mayoritas. Selanjutnya, atribut masukan atau fitur ($x$) dan keluaran atau target ($y$) dipisahkan dari *dataset* agar memudahkan proses penyeimbangan data. Atribut fitur berisi informasi rekam medis, seperti *restingbp* dan *serumcholestrol*, sedangkan target menunjukkan adanya penyakit jantung atau tidak pada pasien. Kemudian, objek SMOTE diinisialisasi dengan *seed random* untuk reproduksibilitas dengan menggunakan parameter *default*




**Gambar 2.** Informasi *outlier*.

TABEL II
DISTRIBUSI KELAS DATA

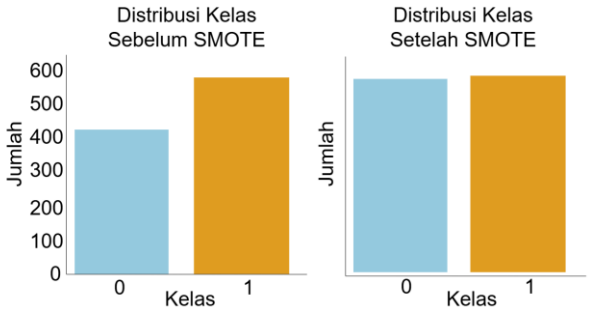
| No. | Kelas | Jumlah Data |
|---|---|---|
| 1. | 0 | 420 |
| 2. | 1 | 580 |

*k_neighbors* = 5 dan fungsi SMOTE(random_state = 99) diterapkan pada data untuk interpolasi linier dalam menciptakan sampel sintetik dari kelas minoritas. Dalam memvalidasi data sintetik yang telah diciptakan, model dilatih ulang dengan *dataset* yang seimbang dan diuji untuk memastikan bahwa data sintetik tidak memengaruhi kinerja model secara negatif. Setelah SMOTE diterapkan, dihasilkan distribusi kelas yang seimbang dengan fungsi X_resampled, y_resampled = smote.fit_resample(X, y), dengan *X* merupakan atribut fitur atau masukan dan *Y* adalah target atau keluaran.

Hasil pengolahan program sebelum dan sesudah penerapan SMOTE tersaji dalam Gambar 3. Pada gambar tampak bahwa ketidakseimbangan data telah diatasi dengan teknik SMOTE. Setelah dilakukannya penyeimbangan menggunakan SMOTE, data pada kelas 0 (tidak ada penyakit jantung), yang semula sebanyak 420 data, bertambah 160 data, sehingga menjadi 580, sedangkan data kelas 1 (ada penyakit jantung), tetap berjumlah 580 data. Maka, jumlah keseluruhan data setelah penerapan SMOTE adalah 1.160 data.

### 2) REKAYASA FITUR

Pada proses rekayasa fitur, dilakukan wawancara kepada salah satu dokter spesialis jantung, yaitu dr. Gusti Made Odi Sidharta, Sp.JP. Menurut dr. Gusti Made Odi Sidharta, Sp. JP., variabel-variabel yang ada pada *dataset* dapat diterapkan dan digunakan dalam mengklasifikasi penyakit jantung. Namun, dari 14 variabel yang ada dalam *dataset* terdapat satu variabel yang tidak berpengaruh dalam mengklasifikasi penyakit jantung, yaitu variabel "*patiented*", sehingga variabel tersebut tidak akan digunakan pada proses selanjutnya. Sementara itu,




**Gambar 3.** Hasil penerapan teknik SMOTE.

TABEL III
HASIL PROSES REKAYASA FITUR

| No. | Variabel | Tipe Data |
|---|---|---|
| 1. | *age* | *Int64* |
| 2. | *gender* | *Int64* |
| 3. | *restingbp* | *Int64* |
| 4. | *serumcholestrol* | *Int64* |
| 5. | *fastingbloodsugar* | *Int64* |
| 6. | *chestpain* | *Int64* |
| 7. | *restingelectro* | *Int64* |
| 8. | *maxheartrate* | *Int64* |
| 9. | *exerciseangina* | *Int64* |
| 10. | *oldpeak* | *float64* |
| 11. | *slope* | *Int64* |
| 12. | *noofmajorvessels* | *Int64* |
| 13. | *target* | *Int64* |

13 variabel lainnya tetap digunakan karena menurut dr. Gusti Made Odi Sidharta, Sp. JP, penyakit jantung ini merupakan penyakit yang sangat berbahaya, sehingga variabel-variabel tersebut sangat penting dalam menentukan seseorang memiliki penyakit jantung atau tidak. Tabel III menyajikan variabel hasil seleksi yang akan digunakan pada proses selanjutnya.

### 3) PEMISAHAN DATA

Aktivitas terakhir pada tahap prapemrosesan data adalah pemisahan data. Aktivitas ini dilakukan untuk membagi data menjadi dua subset, yaitu pelatihan dan pengujian. Pembagian subset tersebut menggunakan rasio 90:10, dengan 90% data (sebanyak 1044 data) sebagai data latih dan 10% data (sebanyak 160 data) sebagai data uji. Hasil penyeimbangan data selanjutnya digunakan untuk melatih dan menguji model *naïve Bayes*, *random forest*, dan *ensemble voting classifier*.

## C. PELATIHAN MODEL DAN EVALUASI

Pada tahap ini dilakukan beberapa aktivitas dalam membangun dan mengevaluasi model untuk klasifikasi penyakit jantung.

### 1) MODEL *NAÏVE BAYES*

Implementasi atau pembangunan model *naïve Bayes* dimulai dari inisialisasi model *naïve Bayes* dengan GaussianNB dari *library* scikit-learn karena model ini tidak memerlukan banyak konfigurasi *hyperparameter* dan hanya menggunakan pengaturan *default* yang sederhana. Kemudian, model dilatih menggunakan data latih sebanyak 90% dari *dataset*. Fungsi yang digunakan dalam melatih model ini yaitu nb_model.fit(X_train, y_train).

Pada tahap ini, model dilatih agar dapat mempelajari pola data. Setelah model dilatih, dihitung probabilitas prediksi dengan nb_model.predict_proba(X_test)[:, 1] dan dibuat prediksi nilai target pada data uji dengan fungsi nb_model.predict(X_test). Perhitungan probabilitas prediksi dilakukan menggunakan (1). Hasil perhitungan probabilitas

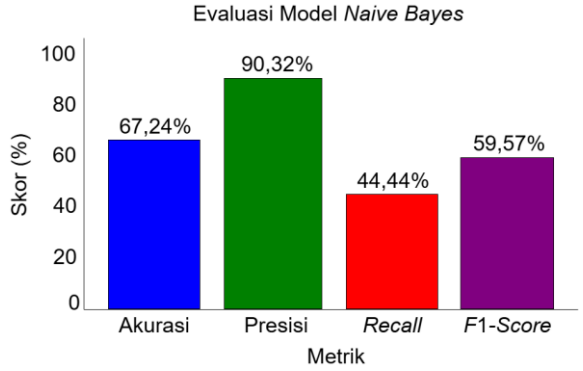
prediksi selanjutnya digunakan dalam evaluasi model untuk menghasilkan *confusion matrix* (akurasi, presisi, *recall*, dan *F*1-*score*). Hasil pengolahan program implementasi model klasifikasi *naïve Bayes* ditunjukkan dalam Gambar 4.

Dari Gambar 4 dapat dilihat bahwa hasil evaluasi terhadap pemodelan klasifikasi penyakit jantung dengan menggunakan algoritma *naïve Bayes* menunjukkan kinerja yang bervariasi di beberapa metrik evaluasi. Algoritma ini berhasil mencapai tingkat akurasi sebesar 67,24%, yang menunjukkan bahwa lebih dari setengah prediksi yang dihasilkan oleh model sesuai dengan label sebenarnya. Selain itu, algoritma *naïve Bayes* juga menghasilkan nilai presisi yang cukup tinggi, yaitu 90,32%, yang mengindikasikan bahwa dari semua prediksi positif yang dihasilkan, sebagian besar benar-benar merupakan kasus penyakit jantung. Namun, dari sisi *recall*, algoritma ini hanya mencapai nilai sebesar 44,44%, yang berarti masih banyak kasus positif yang tidak terdeteksi oleh model. Hal ini menyebabkan nilai *F*1-*score*, yang merupakan rata-rata antara presisi dan *recall*, berada pada angka 59,57%. Evaluasi lebih lanjut juga disajikan melalui *confusion matrix* yang dapat dilihat pada Gambar 5(a) dan Gambar 5(b), yang memberikan gambaran lebih rinci tentang jumlah prediksi benar dan salah yang dihasilkan oleh model untuk setiap kategori.

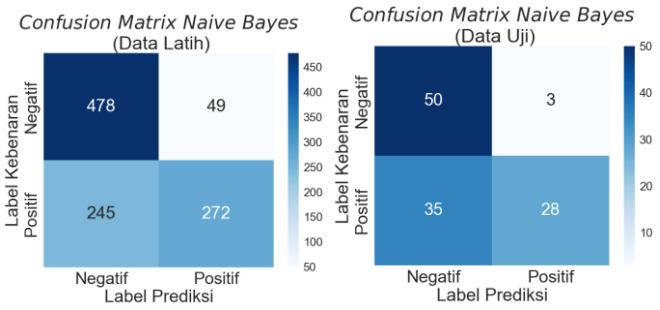
Gambar 5 dihasilkan melalui pengolahan program. Gambar 5(a) menjelaskan analisis kinerja model dengan memberikan informasi tentang jumlah prediksi benar dan salah yang dihasilkan oleh model dibandingkan dengan nilai aktual dari data. Tampak bahwa pada data latih, terdapat 478 data yang diprediksi TN (kelas 0) dan 272 diprediksi TP (kelas 1). Namun, terdapat 49 data yang terdeteksi FP dan 245 data FN. Sementara itu, pada Gambar 5(b) ditampilkan *confusion matrix* dari hasil prediksi model menggunakan data uji. Dapat dilihat bahwa algoritma telah berhasil mengidentifikasi sejumlah besar data dengan tepat. Dari seluruh data yang diuji, terdapat 50 data yang berhasil diprediksi sebagai TN, yaitu data yang memang berasal dari kelas 0 dan diprediksi dengan benar oleh model. Selain itu, terdapat 28 data yang dikategorikan sebagai TP, yang berarti data tersebut berasal dari kelas 1 dan juga diprediksi dengan akurat oleh model. Namun, tidak semua prediksi yang dihasilkan oleh model sesuai dengan kategori sebenarnya. Terdapat tiga data yang diprediksi sebagai FP, yaitu data yang sebenarnya berasal dari kelas 0, tetapi diprediksi oleh model sebagai kelas 1. Selain itu, juga terdapat 35 data yang diprediksi sebagai FN, yaitu data yang seharusnya diklasifikasikan sebagai kelas 1, tetapi diprediksi oleh model sebagai kelas 0. *Confusion matrix* ini memberikan gambaran yang lebih detail mengenai kinerja model, baik dalam mengidentifikasi kelas yang benar maupun dalam membuat kesalahan klasifikasi, yang dapat menjadi acuan untuk peningkatan model lebih lanjut.

### 2) MODEL *RANDOM FOREST*

Tahap selanjutnya merupakan implementasi atau membangun model *random forest*. Pemodelan klasifikasi penyakit jantung dengan algoritma *random forest* dimulai dari inisialisasi *random forest* dengan menggunakan data latih, dengan jumlah 90% dari *dataset*, melalui fungsi RandomForestClassifier(criterion='gini', random_state=77). Model ini dibuat dengan kriteria pemisahan Gini yang digunakan untuk menentukan pemisahan terbaik di setiap *node* serta menggunakan parameter *random state* untuk memastikan hasil yang konsisten. Kemudian, model mulai dilatih dengan fungsi *fit(X_train, y_train)* agar model mempelajari pola dari




**Gambar 4**. Hasil evaluasi model *naïve Bayes*.



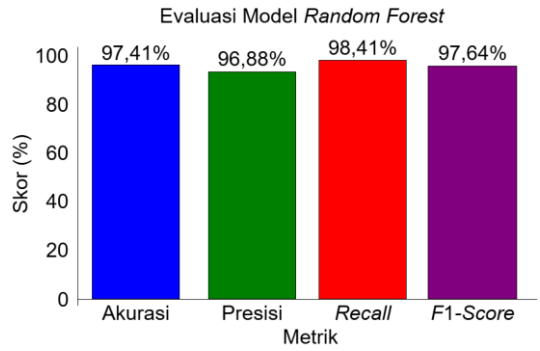

**Gambar 5**. *Confusion matrix* model *naïve Bayes*, (a) data latih, (b) data uji.

data. Lalu, probabilitas prediksi dihitung. Dalam pemodelan digunakan rumus indeks Gini pada (2) serta dibuat prediksi menggunakan predict(X_test) dan data uji sebanyak 10% dari *dataset*. Kemudian, dilanjutkan dengan tahap evaluasi model.

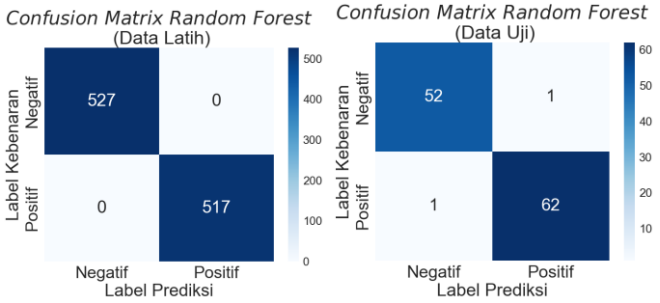
Hasil evaluasi model *random forest* ditunjukkan pada Gambar 6. Gambar tersebut menyajikan persentase hasil evaluasi model *random forest* yang telah dibangun untuk klasifikasi penyakit jantung. Model ini berhasil mencapai kinerja yang sangat baik, dengan akurasi sebesar 97,41%, presisi mencapai 96,88%, *recall* sebesar 98,41%, serta *F*1-*score* 97,64%. Hasil evaluasi ini secara jelas menunjukkan keunggulan model *random forest* dibandingkan model sebelumnya, dengan peningkatan signifikan pada setiap metrik evaluasi yang digunakan. Persentase yang lebih tinggi ini mengindikasikan bahwa *random forest* mampu mengklasifikasikan data dengan tingkat ketepatan dan kesesuaian yang lebih baik.

Selain itu, evaluasi lebih lanjut dilakukan dengan menggunakan *confusion matrix*, yang ditampilkan pada Gambar 7(a) dan Gambar 7(b). Kedua gambar memberikan visualisasi yang lebih rinci tentang distribusi prediksi model, termasuk tingkat keberhasilan model ini mengidentifikasi kategori yang benar dan menghindari kesalahan klasifikasi.

Gambar 7 dihasilkan melalui pengolahan program. Pada Gambar 7(a), hasil evaluasi model *random forest* yang divisualisasikan melalui *confusion matrix* memberikan gambaran yang sangat jelas mengenai kinerja model saat diterapkan pada data latih. Sebanyak 527 data berhasil diklasifikasikan dengan benar sebagai TN, yang berarti bahwa model mampu mengidentifikasi kasus negatif penyakit jantung dengan akurasi tinggi. Selain itu, sebanyak 517 data juga berhasil diklasifikasikan dengan tepat sebagai TP, yang menunjukkan kemampuan model dalam mengenali data yang memiliki penyakit jantung. Selain itu, hasil menunjukkan bahwa tidak ada prediksi yang salah pada data latih ini, baik dalam bentuk FP maupun FN. Hasil ini mengindikasikan




**Gambar 6**. Hasil evaluasi model *random forest*.




**Gambar 7**. *Confusion matrix* model *random forest*, (a) data latih, (b) data uji.

bahwa model *random forest* mampu melakukan pemetaan dengan presisi sempurna terhadap data latih. Sementara itu, pada Gambar 7(b), *confusion matrix* dari hasil evaluasi menggunakan data uji menunjukkan adanya beberapa kesalahan prediksi. Terdapat satu data yang salah diklasifikasikan sebagai kelas 1, padahal seharusnya berada di kelas 0 (FP). Selain itu, satu data yang sebenarnya termasuk kelas 1 justru terprediksi sebagai kelas 0 (FN). Meskipun demikian, model berhasil mengidentifikasi 52 data dengan benar sebagai TN dan 62 data lainnya sebagai TP. Walaupun ada beberapa kesalahan, model *random forest* tetap menunjukkan kinerja yang cukup baik pada data uji, meskipun tidak seakurat kinerjanya pada data latih.

### 3) *ENSEMBLE VOTING CLASSIFIER*

Tahap terakhir merupakan penerapan *ensemble voting classifier* untuk mengombinasikan algoritma *naïve Bayes* dan *random forest*. Penerapan *ensemble voting classifier* dimulai dengan inisialisasi model *naïve Bayes* dan *random forest*, yang selanjutnya diikuti dengan inisialisasi *ensemble voting classifier* pada kedua model tersebut dengan menggunakan funsgi VotingClassifier. Algoritma yang digabungkan hanya dua algoritma, sehingga *ensemble* yang digunakan adalah *soft voting* dengan fungsi voting = 'soft'. Model dengan metode *ensemble* ini akan mengambil keputusan akhir berdasarkan rata-rata probabilitas yang dihasilkan dari algoritma *naïve Bayes* dan *random forest*. Dalam hal ini, *naïve Bayes* menghasilkan probabilitas prediksi berdasarkan asumsi independen antarfitur. Pendekatan yang digunakan dalam *naïve Bayes* ini dapat menyederhanakan proses perhitungan probabilitas gabungan dari fitur-fitur untuk memprediksi kelas target. Sementara itu, *random forest* menghasilkan probabilitas prediksi dengan membangun beberapa *decision tree* dan mengombinasikan hasil prediksi dari setiap *tree*. Kombinasi kedua model dengan menggunakan *ensemble voting classifier* ini dapat mengurangi distribusi *error* yang mungkin terjadi pada model individu. Setelah inisialisasi *ensemble*, proses dilanjutkan dengan melatih model *ensemble* menggunakan data latih (90% dari *dataset*), dengan fungsi fit(), dan menghitung probabilitas prediksi serta membuat prediksi menggunakan data uji (10% dari *dataset*), lalu diakhiri dengan evaluasi model. Hasil pengolahan program evaluasi pada model ini ditunjukkan pada Gambar 8.

Sebagaimana terlihat pada Gambar 8, model klasifikasi yang menggunakan metode *ensemble voting classifier* memberikan hasil evaluasi yang sangat memuaskan. Model ini berhasil mencapai akurasi sebesar 98,28%, dengan nilai presisi yang tinggi, yaitu 98,41%. Ini menunjukkan bahwa sebagian besar prediksi positif adalah benar. Selain itu, model juga menunjukkan kinerja yang stabil dalam mengidentifikasi data dengan benar, terbukti dari nilai *recall* sebesar 98,41%. Hal ini memastikan bahwa mayoritas pasien dengan penyakit jantung teridentifikasi dengan benar. Secara keseluruhan, model ini menghasilkan nilai *F*1-*score* yang juga mencapai 98,41%, yang mengindikasikan keseimbangan antara presisi dan *recall*. Hasil evaluasi ini menyatakan bahwa penggabungan algoritma *naive Bayes* dan *random forest clasiffier* menggunakan *ensemble voting clasiffier* cocok dan dapat memperkuat keandalan model dalam prediksi penyakit jantung. Jika kedua algoritma tidak cocok, kemungkinan akurasi yang dihasilkan akan menurun dibandingkan dengan saat digunakan algoritma tunggal.

Selain itu, untuk memahami lebih dalam kinerja model, dilakukan analisis menggunakan *confusion matrix*. Gambar 9(a) dan Gambar 9(b), yang dihasilkan melalui pengolahan program, merupakan visualisasi dari *confusion matrix*.

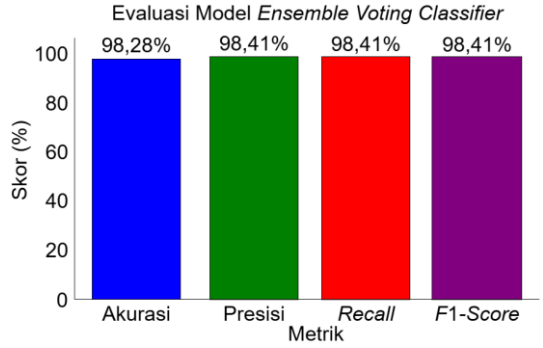
Berdasarkan Gambar 9(a), *confusion matrix* dengan data latih tidak menunjukkan kesalahan sama sekali, dengan 527 berada pada kelas 0 dan 517 berada pada kelas 1 yang terprediksi benar. Namun, berdasarkan Gambar 9(b), terlihat bahwa penerapan metode *ensemble voting classifier* menghasilkan sedikit kesalahan pada penggunaan data uji. Kesalahan hanya ada pada satu data yang terprediksi FP dan satu data lainnya yang terprediksi FN. Selain itu, tidak ada lagi kesalahan yang terlihat. Sementara itu, 52 data lainnya terprediksi benar berada pada kelas 0 dan 62 data terprediksi benar pada kelas 1.

Berdasarkan hasil uji coba keseluruhan model yang terdiri atas algoritma *naïve Bayes*, *random forest*, dan *ensemble voting classifier*, diperoleh hasil evaluasi yang menunjukkan bahwa kinerja terbaik dicapai oleh model yang menggabungkan beberapa algoritma, yaitu *ensemble voting classifier*. Setiap model telah diuji dengan cermat dan hasil kinerjanya disajikan secara rinci dalam Tabel IV, yang menggambarkan perbandingan berbagai metrik evaluasi, seperti akurasi, presisi, *recall*, dan *F*1-*score*.

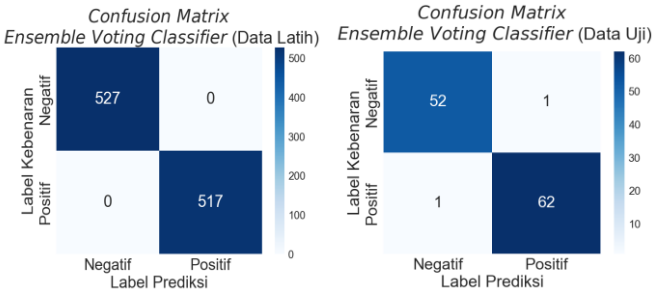
Dari Tabel IV, terlihat jelas bahwa model *ensemble voting classifier* menunjukkan kinerja yang lebih baik dibandingkan dengan pemodelan *naïve Bayes* dan *random forest* secara individual. Meskipun hasil *recall* memiliki persentase yang sama dengan algoritma lain, model *ensemble voting classifier* menunjukkan keunggulan pada metrik lainnya. Akurasi sebesar 98,28%, presisi 98,41%, dan *F*1-*score* 98,41% menegaskan bahwa penggabungan beberapa algoritma dalam satu model dapat meningkatkan akurasi dan keandalan prediksi.

## V. PEMBAHASAN

Penelitian ini berhasil membuat pemodelan klasifikasi untuk penyakit jantung dengan menggunakan penggabungan algoritma *naïve Bayes* dan algoritma *random forest* melalui *ensemble voting classifier*. *ensemble voting classifier*




**Gambar 8**. Hasil evaluasi model *ensemble voting classifier*.




(a) (b)

**Gambar 9**. *Confusion matrix* model *ensemble voting classifier*, (a) data latih, (b) data uji.

TABEL IV
HASIL EVALUASI SELURUH MODEL

| Model Klasifikasi | Matriks Evaluasi (%) | | | |
|---|---|---|---|---|
| | ***Akurasi*** | ***Presisi*** | ***Recall*** | ***F1-Score*** |
| *Naïve Bayes* | 67,24 | 90,32 | 44,44 | 59,57 |
| *Random forest* | 97,41 | 96,88 | 98,41 | 97,64 |
| *Ensemble voting classifier* | 98,28 | 98,41 | 98,41 | 98,41 |

digunakan untuk mengombinasikan kedua algoritma dengan cara kerjanya yang memanfaatkkan kekuatan dari setiap algoritma yang digunakan. Dengan *dataset* penyakit jantung yang telah diolah sejumlah 1.160 data, implementasi *ensemble voting classifier* berhasil diterapkan dan memberikan kinerja yang akurat, konsisten, serta berhasil meningkatkan akurasi dibandingkan dengan penggunaan algoritma *naïve Bayes* saja.

Hasil penelitian ini membuktikan bahwa penerapan *ensemble voting classifier* dapat digunakan dalam membuat model klasifikasi penyakit jantung yang lebih efektif dan akurat. Hal ini selaras dengan penelitian sebelumnya yang telah menguji kinerja model klasifikasi penyakit jantung dengan membandingkan algoritma *naïve Bayes* dan *random forest* [9]–[13]. Hasil dari masing-masing penelitian terkait diperlihatkan pada Tabel V.

Berdasarkan Tabel V, penelitian ini tidak hanya berfokus pada perbandingan antara dua algoritma, yaitu *naïve Bayes* dan *random forest*, tetapi juga memberikan kontribusi tambahan dengan menerapkan metode *ensemble voting classifier*. Metode ini diterapkan dengan tujuan untuk meningkatkan akurasi dan stabilitas model klasifikasi penyakit jantung. Hasil dari penelitian ini menunjukkan bahwa penggunaan *ensemble voting classifier* mampu menghasilkan akurasi yang lebih tinggi jika dibandingkan dengan hanya mengandalkan algoritma *naïve Bayes* atau *random forest* secara terpisah. Penerapan metode *ensemble* ini berhasil memadukan kekuatan dari kedua algoritma, sehingga dapat meningkatkan kinerja

TABEL V
HASIL PERBANDINGAN AKURASI PENELITIAN

| Penelitian | Metode | Akurasi (%) |
|---|---|---|
| Penelitian ini | *Ensemble voting classifier*, algoritma *naïve Bayes* dan *random forest* | *Naïve Bayes*: 67,24<br>*Random forest*: 97,41<br>*Ensemble voting classifier*: 98,28 |
| [9] | *Decision tree*, *naïve Bayes*, dan *random forest classifier* | *Decision tree*: 84,40<br>*Naïve Bayes*: 85,00<br>*Random forest classifier*: 85,20 |
| [10] | *Random forest classifier*, C45 *algorithm*, *logistic regression*, dan (SVM) | *Random forest classifier*: 94,54<br>C45 *algorithm*: 91,74<br>*Logistic regression*: 76,27<br>SVM: 76,24 |
| [11] | *Decision tree*, *naïve Bayes*, dan *random forest* | *Decision tree*: 72,00<br>*Naïve Bayes*: 71,00<br>*Random forest*: 75,00 |
| [12] | *Naïve Bayes*, KNN, *random forest*, *logistic regression*, SVM, *decision tree*, dan AdaBoost | *Naïve Bayes*: 84,67<br>KNN: 73,00<br>*Random forest*:81,70<br>*Logistic regression*: 84,30<br>SVM: 81,00<br>*Decision tree*: 74,00<br>AdaBoost: 71,30 |
| [13] | *Naïve Bayes*, *random forest classifier*, dan *k-nearest neighbor* (KNN) | *Naïve Bayes:* 91,00<br>*Random forest*: 89,90<br>KNN: 68,60 |

keseluruhan model dan memberikan hasil prediksi yang lebih akurat serta andal.

Pemodelan klasifikasi penyakit jantung ini menunjukkan bahwa penerapan metode *ensemble voting classifier* dapat berkontribusi signifikan dalam bidang medis, khususnya dalam diagnosis penyakit jantung. Dengan mengombinasikan *naïve Bayes* dan *random forest*, model ini menawarkan peningkatan akurasi dan kecepatan prediksi, yang mendukung pengambilan keputusan klinis yang lebih cepat dan tepat. Potensi integrasi hasil ini ke dalam sistem pendukung keputusan klinis di fasilitas kesehatan dapat membantu dalam identifikasi pasien berisiko tinggi secara lebih efisien, memungkinkan deteksi dini yang lebih efektif, serta berkontribusi pada peningkatan kualitas hidup pasien.

## VI. KESIMPULAN

Penelitian ini berhasil membangun model klasifikasi penyakit jantung dengan mengimplementasikan *ensemble voting classifier* dalam menggabungkan algoritma *naïve Bayes* dan *random forest*. Dalam prosesnya, terdapat masalah yang terindentifikasi, yaitu ketidakseimbangan kelas data. Untuk mengatasi masalah ini, diterapkan teknik SMOTE, yang terbukti efektif dalam menangani ketidakseimbangan kelas data. Proses pemisahan data dilakukan dengan membagi data menjadi data latih dan data uji, dengan rasio 90:10. Berdasarkan hasil evaluasi, terlihat adanya peningkatan kinerj setelah penerapan *ensemble voting classifier*. Penggunaan *ensemble voting classifier* untuk menggabungkan algoritma *naïve Bayes* dan *random forest* dalam model klasifikasi penyakit jantung menunjukkan kinerja yang sangat akurat dan

stabil. Model ini menghasilkan akurasi 98,28%, presisi 98,41%, *recall* 98,41%, dan *F*1-*score* 98,41%. Meskipun terdapat satu persamaan persentase pada metrik *recall*, metode *ensemble voting classifier* memberikan hasil yang lebih komprehensif dan unggul berkat kontribusi dari tiga *base classifier* lainnya (akurasi, presisi, dan *F*1-*score*). Hal ini menandakan bahwa penggabungan beberapa algoritma dalam satu model *ensemble* tidak hanya mampu memperbaiki hasil evaluasi, tetapi juga meningkatkan keandalan dalam model prediksi medis yang krusial, seperti klasifikasi penyakit jantung.

Penelitian ini menggunakan *ensemble voting classifier* dalam menggabungkan dua algoritma dan memanfaatkan kekuatan dari masing-masing algoritma (*naïve Bayes* dan *random forest*) dalam meningkatkan akurasi model. Namun, untuk penelitian di masa depan, disarankan untuk mengeksplorasi algoritma dan metode *ensemble* lainnya dalam meningkatkan kinerja model, guna mencapai hasil yang lebih akurat.

## KONFLIK KEPENTINGAN

Penulis menyatakan bahwa tidak ada konflik kepentingan.

## KONTRIBUSI PENULIS

Konseptualisasi, Dede Kurniadi dan Asri Indah Pertiwi; metodologi, Dede Kurniadi; perangkat lunak, Asri Indah Pertiwi; data dan validasi, Dede Kurniadi dan Asri Indah Pertiwi; penulisan—penyusunan draf asli, Dede Kurniadi dan Asri Indah Pertiwi; penulisan—peninjauan dan penyuntingan, Dede Kurniadi, Asri Indah Pertiwi, dan Asri Mulyani; visualisasi, Asri Indah Pertiwi; pengawasan, Dede Kurniadi dan Asri Mulyani.

## UCAPAN TERIMA KASIH

Ucapan terima kasih disampaikan kepada Institut Teknologi Garut yang telah mendanai penelitian ini dan kepada dr. Gusti Made Odi Sidharta, Sp.JP. yang telah meluangkan waktu untuk wawancara pada proses rekayasa fitur.

## REFERENSI

[1] L.P.C. Dewi. “Jenis, gejala, dan penyebab penyakit jantung.” Tanggal akses: 13-Mar-2024. [Online]. Tersedia: https://rs-soewandhi.surabaya.go.id/jenis-gejala-dan-penyebab-penyakit-jantung/

[2] S.D. Sawu, A.A. Prayitno, dan Y.I. Wibowo, “Analisis faktor risiko pada kejadian masuk rumah sakit penyakit jantung koroner di Rumah Sakit Husada Utama Surabaya,” *J. Sains Kesehat.*, vol. 4, no. 1, hal. 10–18, Jul. 2022, doi: 10.25026/jsk.v4i1.856.

[3] M. Ardiana, *Buku Ajar Prevensi dan Rehabilitasi Jantung*. Surabaya, Indonesia: Airlangga University Press, 2022.

[4] World Health Organization. “Cardiovascular diseases.” Tanggal akses: 7-Agu-2024. [Online]. Tersedia: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[5] Tim Riskesdas 2018, *Laporan Nasional Riskedas 2018*. Jakarta, Indonesia: Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan, 2019.

[6] Y.P. Santosa. “Memahami pentingnya cek kesehatan jantung.” Primaya Hospital. Tanggal akses: 8-Jul-2024. [Online]. Tersedia: https://primayahospital.com/jantung/pentingnya-cek-kesehatan-jantung/

[7] Direktorat Jenderal Pencegahan dan Pengendalian Penyakit. “Pemeriksaan, Gejala, dan Diet untuk Jantung.” Tanggal akses: 2-Jun-2024. [Online]. https://p2p.kemkes.go.id/pemeriksaan-gejala-dan-diet-untuk-jantung/

[8] A.M.A. Rahim, I.Y.R. Pratiwi, dan M.A. Fikri, “Klasifikasi penyakit jantung menggunakan metode synthetic minority over-sampling technique dan random forest clasifier,” *Indones. J. Comput. Sci.*, vol. 12, no. 5, hal. 2995–3011, Okt. 2023, doi: 10.33022/ijcs.v12i5.3413.

[9] J.D. Muthohhar dan A. Prihanto, “Analisis perbandingan algoritma klasifikasi untuk penyakit jantung,” *J. Inform. Comput. Sci. (JINACS)*, vol. 4, no. 3, hal. 298–304, Mar. 2023, doi: 10.26740/jinacs.v4n03.p298-304.

[10] A.F.N. Masruriyah dkk., “Evaluasi algoritma pembelajaran terbimbing terhadap dataset penyakit jantung yang telah dilakukan oversampling,” *MIND (Multimed. Artif. Intell. Netw. Database) J.*, vol. 8, no. 2, hal. 242–253, Des. 2023, doi: 10.26760/mindjournal.v8i2.242-253.

[11] D.H. Depari, Y. Widiastiwi, dan M.M. Santoni, “Perbandingan model decision tree, naive Bayes dan random forest untuk prediksi klasifikasi penyakit jantung,” *Inform., J. Ilmu Komput.*, vol. 18, no. 3, hal. 239–248, Des. 2022, doi: 10.52958/iftk.v18i3.4694.

[12] Ratnasari, A.J. Wahidin, A.E. Setiawan, dan P. Bintoro, “Machine learning untuk klasifikasi penyakit jantung,” *Aisyah J. Inform. Electr. Eng.*, vol. 6, no. 1, hal. 145–150, Feb. 2024, doi: 10.30604/jti.v6i1.272.

[13] A. Samosir, M.S. Hasibuan, W.E. Justino, dan T. Hariyono, “Komparasi algoritma random forest, naïve Bayes dan k-nearest neighbor dalam klasifikasi data penyakit jantung,” dalam *Pros. Semin. Nas. Darmajaya*, 2021, hal. 214–222.

[14] B. Asrun dan I. Irmayani, “Penerapan konsep non-deterministic finite automata dalam diagnosa penyakit jantung,” *Dewantara J. Technol.*, vol. 3, no. 1, hal. 122–125, Mei 2022, doi: 10.59563/djtech.v3i1.184.

[15] P.D. Kusuma, *Machine Learning Teori, Program, dan Studi Kasus*. Yogyakarta, Indonesia: Deepublish, 2020.

[16] P.B.N. Setio, D.R.S. Saputro, dan B. Winarno, “Klasifikasi dengan pohon keputusan berbasis algoritme C4.5,” dalam *PRISMA, Pros. Semin. Nas. Mat.*, 2020, hal. 64–71.

[17] Rayuwati, H. Gemasih, dan I. Nizar, “Implementasi algoritma naive Bayes untuk memprediksi tingkat penyebaran Covid,” *J. Ris. Rumpun Ilmu Tek.*., vol. 1, no. 1, hal. 38–46, Apr. 2022, doi: 10.55606/jurritek.v1i1.127.

[18] M. Al-Husaini, P.A. Saputra, M. Renaldi, dan R.A. Maulana, *Prediksi Tsunami dengan Metode Ensemble Machine Learning*. Jambi, Indonesia: PT. Sonpedia Publishing Indonesia, 2024.

[19] Sarwido, G.W.N. Wibowo, dan M.A. Manan, “Penerapan algoritma naive Bayes untuk prediksi heregistrasi calon mahasiswa baru,” *J. Tek. Inform.*, vol. 1, no. 1, hal. 1–10, Feb. 2022, doi: 10.02220/jtinfo.v1i1.126.

[20] S. Saadah dan H. Salsabila, “Prediksi harga Bitcoin menggunakan metode random forest,” *J. Komput. Terap.*, vol. 7, no. 1, hal. 24–32, Jun. 2021, doi: 10.35143/jkt.v7i1.4618.

[21] M.R. Adrian, M.P. Putra, M.H. Rafialdy, dan N.A. Rakhmawati, “Perbandingan metode klasifikasi random forest dan SVM pada analisis sentimen PSBB,” *J. Inform. UPGRIS*, vol. 7, no. 1, hal. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099.

[22] I. Daqiqil, *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python*, 1st ed. Riau, Indonesia: UR PRESS, 2021.

[23] A.J. Mohammed, M.M. Hassan, dan D.H. Kadir, “Improving classification performance for a novel imbalanced medical dataset using SMOTE method,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, hal. 3161–3172, Jun. 2020, doi: 10.30534/ijatcse/2020/104932020.

[24] M. Ardiansyah, “Model ensemble algoritma naive Bayes dan random forest dalam klasifikasi penyakit paru-paru untuk meningkatkan akurasi,” *SMARTLOCK, J. Sains dan Teknol.*., vol. 2, no. 2, hal. 32–38, Des. 2023, doi: 10.37476/smartlock.v2i2.4407.

[25] S. Bashir dkk., “A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction,” *IEEE Access*, vol. 9, hal. 130805–130822, Sep. 2021, doi: 10.1109/ACCESS.2021.3110604.

[26] J. Dumlao. “Cardiovascular Disease Dataset.” Kaggle. Tanggal akses: 28-Apr-2024. [Online]. Tersedia: https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset

[27] F. Rahman dan Mustikasari, “Optimization of student graduation predictions on time using binning and synthetic minority oversampling technique (SMOTE),” *Jagti*, vol. 4, no. 1, hal. 30–36, Feb. 2024, doi: 10.24252/jagti.v4i1.77.