Peningkatan Performa *Classification and Regression Tree* Menggunakan *Bagging* pada Diagnosis Penyakit Jantung

http://dx.doi.org/10.28932/jutisi.v12i1.12439

Riwayat Artikel
Received: 15 Juli 2025 | *Final Revision*: 17 Februari 2026 | *Accepted*: 24 Februari 2026

Kokom Hera Fitriyana[✉ #1], Fitri Ayuning Tyas[*2], Abdul Jamil[#3]

[#]*Teknik Informatika, Universitas Muhammadiyah Brebes*
Jl. Pangeran Diponegoro, Grengseng No.184, Taraban, Kecamatan Paguyangan, Kabupaten Brebes 52276, Indonesia
[1]kokomherafitriyana14@gmail.com

[*]*Sistem Informasi, Universitas Muhammadiyah Brebes*
Jl. Pangeran Diponegoro, Grengseng No.184, Taraban, Kecamatan Paguyangan, Kabupaten Brebes 52276, Indonesia
[2]tyas_fa@umbs.ac.id

[#]*Manajemen, Universitas Muhammadiyah Brebes*
Jl. Pangeran Diponegoro, Grengseng No.184, Taraban, Kecamatan Paguyangan, Kabupaten Brebes 52276, Indonesia
[3]abdul.jamil@umbs.ac.id

[✉]Corresponding author: kokomherafitriyana14@gmail.com

Abstrak — **Penyakit jantung merupakan salah satu penyebab utama kematian di dunia, sehingga dibutuhkan metode diagnosis yang cepat dan akurat untuk menanggulanginya. Salah satu pendekatan yang dapat dimanfaatkan adalah data mining, khususnya metode klasifikasi untuk menganalisis data kesehatan. Algoritma *Classification and Regression Tree* (CART) memiliki keunggulan dalam interpretabilitas, namun masih kurang stabil terhadap perubahan data. Untuk mengatasi keterbatasan tersebut, penelitian ini menerapkan teknik *Bootstrap Aggregating* (*Bagging*) guna meningkatkan kestabilan dan akurasi model. Data penelitian diperoleh dari tiga *dataset* yang tersedia di *platform Kaggle*, yaitu *Heart Disease*, *Heart Disease Cleveland*, dan *Heart Disease Prediction*. Eksperimen dilakukan pada setiap *dataset* secara terpisah atau individual serta pada data gabungan karena ketiganya memiliki karakteristik atribut yang serupa. Tahapan penelitian meliputi *preprocessing data* berupa penyesuaian tipe atribut, penanganan *missing value*, serta *identifikasi outlier,* kemudian dilanjutkan dengan pelatihan dua model klasifikasi, yaitu CART tunggal dan CART berbasis *Bagging* menggunakan parameter *default*. Evaluasi performa dilakukan menggunakan metrik akurasi melalui metode *k-fold cross-validation*. Hasil penelitian menunjukkan bahwa *Bagging* secara konsisten meningkatkan akurasi CART pada seluruh *dataset*. Pada *dataset Heart Disease*, akurasi meningkat dari 72,89% menjadi 78,00%. Pada *Heart Disease Cleveland*, akurasi naik dari 81,89% menjadi 85,78%. Pada *Heart Disease Prediction*, akurasi mengalami peningkatan dari 77,44% menjadi 82,44%. Sementara itu, pada data gabungan, akurasi CART naik dari 76,00% menjadi 79,11% setelah penerapan *Bagging*. Dengan demikian, teknik *Bagging* terbukti efektif dalam meningkatkan akurasi dan kestabilan model CART untuk diagnosis penyakit jantung.**

Kata kunci— ***Bagging*; CART; Klasifikasi; Penyakit Jantung.**

Improving Classification and Regression Tree Performance Using Bagging in Heart Disease Diagnosis

Abstract — ***Heart disease is one of the leading causes of death worldwide, thus requiring fast and accurate diagnostic methods to address it. One approach that can be utilized is data mining, particularly classification methods for analyzing health data. The Classification and Regression Tree (CART) algorithm offers advantages in interpretability but remains less stable against variations in data. To overcome this limitation, this study applies the Bootstrap Aggregating (Bagging) technique to enhance model stability and accuracy. The research uses three datasets available on the Kaggle platform, namely Heart Disease, Heart Disease Cleveland, and Heart Disease Prediction. Experiments were conducted on each dataset individually as well as on a combined dataset, as they share similar attribute characteristics. The research stages include data preprocessing—such as adjusting attribute types, handling missing values, and identifying outliers—followed by training two classification models: a single CART model and a Bagging-based CART model using default parameters. Model performance was evaluated using accuracy metrics through k-fold cross-validation. The results show that Bagging consistently improves CART accuracy across all datasets. On the Heart Disease dataset, accuracy increased from 72.89% to 78.00%. On the Heart Disease Cleveland dataset, accuracy rose from 81.89% to 85.78%. On the Heart Disease Prediction dataset, accuracy improved from 77.44% to 82.44%. Meanwhile, for the combined dataset, CART accuracy increased from 76.00% to 79.11% after applying Bagging. Thus, the Bagging technique is proven effective in enhancing the accuracy and stability of the CART model for heart disease diagnosis.***

Keywords— ***Bagging***; ***CART***; ***Classification***; ***Heart Disease.***

I. PENDAHULUAN

Penyakit jantung atau dikenal sebagai penyakit kardiovaskular adalah penyebab utama kematian secara global. Menurut data *World Health Organization* (WHO) diperkirakan 17,9 juta orang meninggal karena penyakit kardiovaskular pada tahun 2019, mewakili 32% dari seluruh kematian global. Dari kematian tersebut, 85% disebabkan oleh serangan jantung dan stroke [1]. Berdasarkan data Riskesdas, prevalensi penyakit jantung yang didiagnosis oleh dokter di Indonesia mencapai 1,5% [2]. Penyakit jantung merupakan penyakit yang terjadi akibat adanya gangguan yang memengaruhi fungsi jantung [3]. Gejala seperti tekanan darah tinggi, kolesterol tinggi, stres, obesitas dan diabetes merupakan faktor utama yang dapat memengaruhi kesehatan jantung [4]. Sehingga diperlukan suatu pendekatan yang lebih efektif dalam menganalisis faktor risiko dan pola penyakit jantung, salah satunya dengan memanfaatkan *data mining* untuk menggali informasi dari data kesehatan yang tersedia.

Data mining merupakan proses eksplorasi dan analisis data dalam jumlah besar untuk mengidentifikasi pola serta aturan yang memiliki makna [5],[6],[7]. Klasifikasi merupakan metode analisis data yang bertujuan untuk membangun model yang dapat menggambarkan kategori data yang signifikan [8]. Metode klasifikasi terdiri dari beberapa algoritma yang umum digunakan, salah satunya yaitu algoritma CART.

CART (*Classification and Regression Tree*) merupakan salah satu algoritma *machine learning* non-parametrik yang umum digunakan dalam analisis klasifikasi dengan pohon keputusan, yang dapat diterapkan baik pada variabel respon kategorik maupun kontinu [9]. Jika target atau variabel dependennya kontinu, maka pohon yang dihasilkan adalah pohon regresi. Sementara itu, jika targetnya kategoris, maka pohon yang dihasilkan adalah pohon klasifikasi [10]. CART menggunakan *Gini Index* sebagai dasar pemilihan atribut optimal dalam proses pemisahan data [11]. CART memiliki beberapa keunggulan dibandingkan algoritma klasifikasi lainnya, antara lain memiliki pohon keputusan yang mudah diinterpretasikan, memiliki akurasi yang cukup baik, dan memiliki perhitungan yang lebih cepat [12],[13],[14]. Beberapa penelitian telah memanfaatkan algoritma CART ini dalam berbagai bidang, khususnya dalam bidang kesehatan. Seperti penelitian yang dilakukan oleh [15] dan [16]. Penelitian-penelitian tersebut membandingkan algoritma CART dengan beberapa algoritma klasifikasi lain untuk mengolah data medis dan hasilnya menunjukkan bahwa algoritma CART mampu memberikan tingkat akurasi yang lebih tinggi dibandingkan algoritma-algoritma lainnya. Namun, dibalik kelebihan yang dimiliki oleh algoritma CART terdapat kekurangan yaitu sangat bergantung pada jumlah sampel dan modelnya kurang stabil, yang apabila terjadi sedikit saja perubahan pada data training akan berpengaruh besar pada model pembelajaran yang dihasilkan [14]. Untuk menutupi kekurangan tersebut, *bagging* yang merupakan bagian dari teknik *ensemble learning* dapat menjadi solusi karena dapat membuat model klasifikasi menjadi lebih stabil dan meningkatkan akurasi klasifikasi.

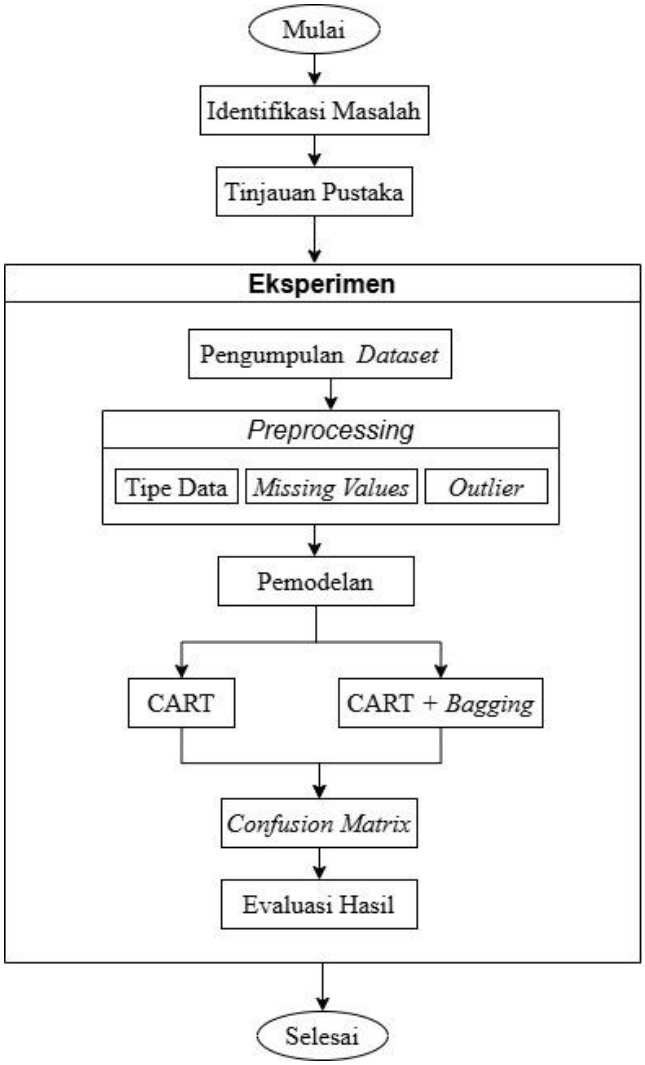
Bagging (*Bootstrap Aggregating*) adalah salah satu teknik *ensemble* dalam *machine learning* yang berfungsi untuk meningkatkan performa klasifikasi dengan mengombinasikan berbagai model klasifikasi secara acak pada *dataset training*

yang dapat membantu mengurangi variasi serta mencegah terjadinya *overfitting* [17],[18]. *Bagging* bersifat paralel yang membuat setiap model bekerja secara independen, sehingga dapat meningkatkan akurasi [19]. Teknik *Bagging* dapat diterapkan untuk mempertahankan keragaman model, meningkatkan kecepatan dalam proses klasifikasi, serta mengurangi kebutuhan memori [18]. Beberapa penelitian telah membuktikan bahwa penggunaan teknik *Bagging* dapat secara efektif meningkatkan akurasi suatu algoritma klasifikasi diantaranya yaitu penelitian oleh [20] dan [21].

Berdasarkan pemaparan tersebut, penelitian ini berfokus pada peningkatan performa algoritma CART menggunakan teknik *Bagging* dalam mendiagnosis penyakit jantung melalui pendekatan penggabungan beberapa *dataset*. Penggabungan *dataset* dilakukan untuk memperkaya variasi data serta meningkatkan kemampuan generalisasi model yang dibangun. Adapun *dataset* yang digunakan dalam penelitian ini terdiri dari tiga *dataset* yang bersumber dari *Kaggle*, yang memiliki karakteristik atribut yang relatif serupa sehingga memungkinkan untuk dilakukan integrasi data secara efektif.

II. METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini adalah metode eksperimen. Metode eksperimen dilakukan dengan mengamati berbagai variabel yang menjadi objek penelitian [22]. Penelitian ini menerapkan metode eksperimen untuk menerapkan teknik *Bagging* pada algoritma CART guna meningkatkan performa klasifikasi dalam diagnosis penyakit jantung dibandingkan model algoritma CART tanpa *Bagging*. Penelitian ini merancang alur yang tersusun secara terstruktur untuk menggambarkan tahapan penelitian yang akan dilakukan. Alur tersebut disajikan dalam bentuk gambar yang menampilkan tahapan penelitian secara runtut. Rincian tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1 Tahapan Penelitian

A. Pengumpulan Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari tiga *dataset* terkait penyakit jantung yang diperoleh dari *platform Kaggle*. Ketiga *dataset* tersebut adalah: *Heart Disease*, *Heart Disease Cleveland*, dan *Heart Disease Prediction*.

B. Preprocessing

Preprocessing merupakan langkah awal dalam mempersiapkan data, agar data lebih mudah diolah dan dianalisis [23]. Data yang akan diolah dan dianalisis tidak selalu dalam kondisi baik untuk diproses, karena dapat mengandung format yang tidak konsisten, *missing value*, dan *outlier* [24]. Oleh karena itu, tahap *preprocessing* sangat penting untuk meningkatkan kualitas data dan memastikan hasil analisis yang lebih akurat. Beberapa tahapan *preprocessing* yang dilakukan dalam penelitian ini meliputi:

1) Format Tipe Data

Format tipe data yang tidak sesuai dapat menyebabkan kesalahan perhitungan, kendala dalam penerapan algoritma, serta ketidaksesuaian dalam integrasi data dari berbagai sumber [25]. Kemudian dilakukan *one hot encoding* untuk mengonversi variabel kategorikal kedalam bentuk numerik agar dapat digunakan dalam algoritma *machine learning* [26].

2) Missing Values

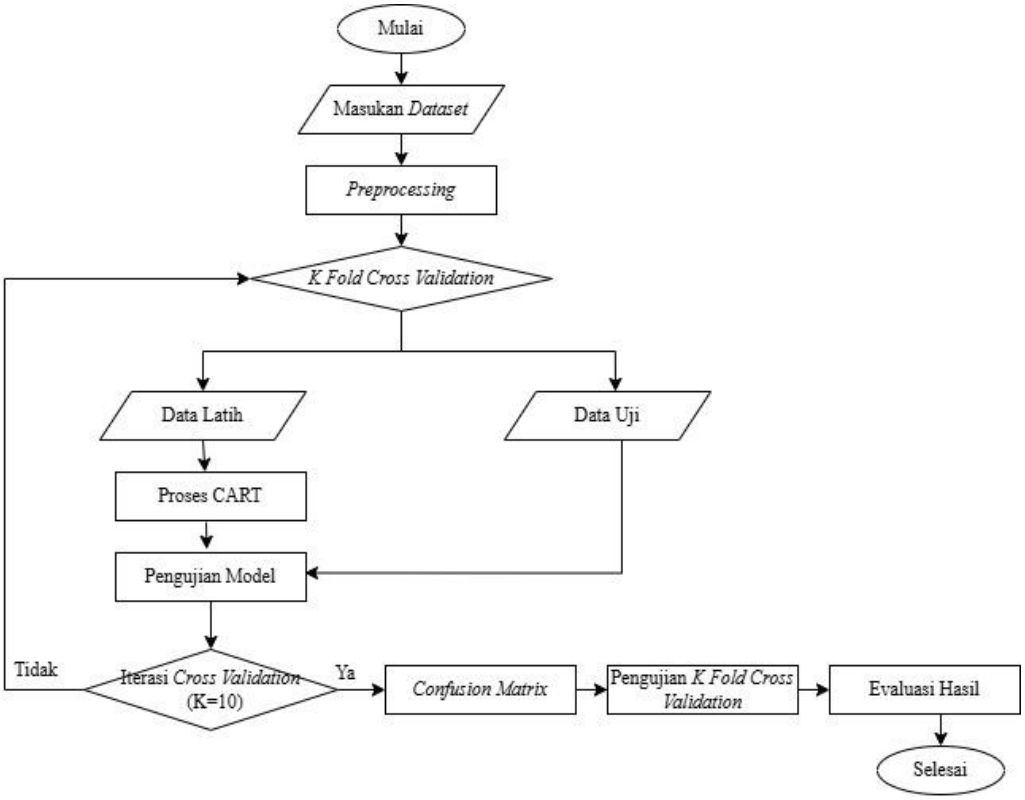
Missing values merupakan kondisi dimana suatu *dataset* mengandung data yang tidak lengkap atau terdapat nilai yang hilang dalam beberapa bagian [27]. *Missing values* sering kali menjadi kendala karena kehilangan data yang penting dapat menghambat analisis, mengurangi efisiensi, dan menurunkan akurasi [28].

3) Outlier

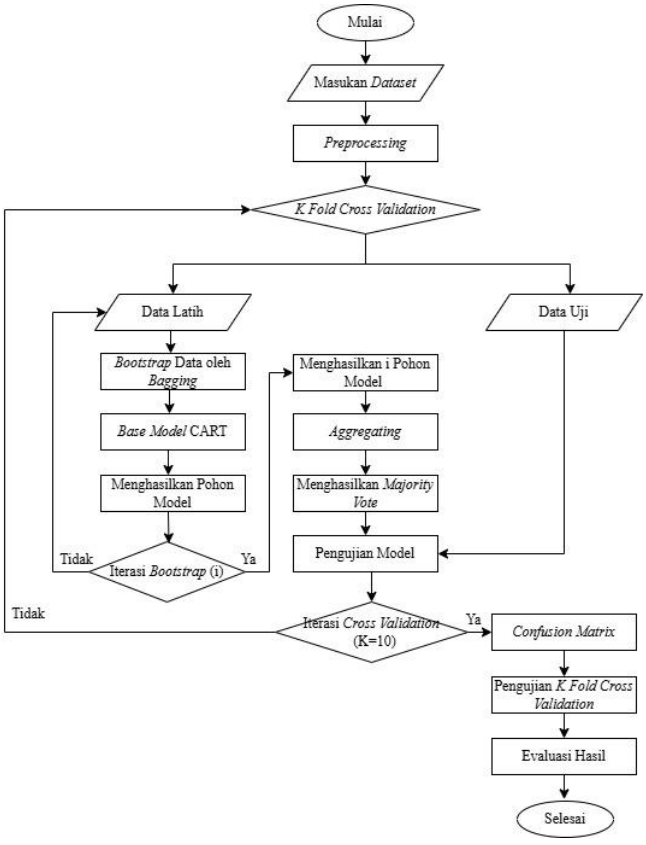
Pendeteksian dan penanganan *outlier* merupakan langkah penting dalam *preprocessing* data, karena keberadaan *outlier* dalam proses *data mining* dapat menyebabkan hasil yang kurang akurat [29]. *Outlier* adalah data yang secara signifikan menyimpang dari pola keseluruhan dalam suatu *dataset*, yang dapat mempengaruhi hasil analisis statistik serta mengurangi keakuratan model [30].

C. Pemodelan

Pada tahap penerapan model, penelitian ini menggunakan dua pendekatan klasifikasi, yaitu algoritma CART tunggal dan CART yang dikombinasikan dengan teknik *Bagging*. Sebelum proses pelatihan model dilakukan, data tidak dibagi secara manual ke dalam data training dan data testing, melainkan divalidasi menggunakan skenario *K-Fold Cross Validation* dengan nilai K = 10. Pendekatan ini dipilih untuk memastikan bahwa seluruh data memperoleh kesempatan yang sama untuk menjadi data latih dan data uji, sehingga hasil evaluasi lebih stabil dan tidak bergantung pada satu kali pembagian data. Model CART diterapkan sebagai model dasar dengan parameter *default*, sedangkan teknik *Bagging* digunakan untuk meningkatkan kestabilan dan akurasi model melalui penggabungan beberapa pohon keputusan yang dilatih pada subset data berbeda. Gambar 2 dan Gambar 3 merupakan penerapan model dalam penelitian ini.



Gambar 2 Model CART



Gambar 3 Model CART *Bagging*

D. Evaluasi Hasil

Evaluasi model dilakukan dengan menggunakan *confusion matrix* untuk membandingkan tingkat akurasi antara model CART tunggal dan model CART yang dikombinasikan dengan teknik *Bagging*. Akurasi digunakan sebagai metrik utama karena menunjukkan persentase prediksi yang sesuai dengan label aktual. Melalui analisis hasil evaluasi, dapat dinilai sejauh mana penerapan teknik *Bagging* mampu meningkatkan kinerja model algoritma CART dan mengurangi tingkat kesalahan dalam mengklasifikasikan kondisi pasien. *Confusion Matrix* memiliki empat komponen utama yang menjadi dasar dalam perhitungan berbagai metrik evaluasi klasifikasi [19], yaitu seperti pada Tabel 1.

TABEL 1
CONFUSION MATRIX

Correct Classification	***Classified as***	
	Positive	***Negative***
Positive	TP	FN
Negative	FP	TN

Keterangan :
TP (*True Positive*) : Data positif yang terklasifikasi di kelas yang benar.
TN (*True Negative*) : Data negatif yang terklasifikasi di kelas yang benar.
FP (*False Positive*) : Data negatif yang salah diklasifikasikan sebagai positif.
FN (*False Negative*) : Data positif yang salah diklasifikasikan sebagai negatif.

Metode untuk mengevaluasi kinerja atau performa suatu algoritma meliputi *accuracy*, *precision*, *recall*, dan *F1-score* dengan rumus yang digunakan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Penelitian ini melakukan peningkatan performa pada algoritma CART menggunakan teknik *Bagging* dengan tujuan untuk mengetahui sejauh mana teknik ini dapat meningkatkan kinerja model dalam mendiagnosis penyakit jantung. Proses klasifikasi, peningkatan performa, dan perhitungan akurasi dilakukan menggunakan bahasa pemrograman *Python* pada *platform google colaboratory*. *Hardware* yang digunakan adalah HP Laptop 14s-dq5xxx dengan spesipikasi:

1) Operating System : Windows 11 Home Single Language 64-bit (10.0, Build 26100)

2) Processor : 12th Gen Intel® Core™ i3-1215U (8 CPUs), ~1.2GHz

3) RAM : 8GB

A. Pengumpulan Dataset

Eksperimen dalam penelitian ini menggunakan tiga *dataset* berbeda yang berkaitan dengan penyakit jantung, yaitu *Heart Disease* (HD), *Heart Disease Cleveland* (HDC), dan *Heart Disease Prediction* (HDP) yang diperoleh dari *platform Kaggle*. Ketiga *dataset* tersebut terlebih dahulu digunakan secara terpisah untuk menganalisis performa model pada karakteristik data yang berbeda. Selain itu, ketiganya juga digabungkan menjadi satu *dataset* komprehensif mengingat secara umum memiliki kesamaan atribut maupun kesesuaian nilai pada masing-masing fitur. Proses penggabungan ini dilakukan untuk memperoleh ukuran data yang lebih besar dan variatif, sehingga dapat memberikan gambaran performa model yang lebih stabil serta meningkatkan kemampuan generalisasi dalam diagnosis penyakit jantung. Deskripsi lengkap mengenai atribut pada masing-masing *dataset* disajikan dalam Tabel 2, Tabel 3, dan Tabel 4.

TABEL 2
ATRIBUT HEART DISEASE

No	**Atribut**	**Deskripsi**	**Keterangan**
1	Age	Usia Pasien	Usia dalam tahun
2	Sex	Jenis Kelamin	0 = Perempuan,1 = Laki-laki

No	Atribut	Deskripsi	Keterangan
3	Cp	Jenis nyeri dada	0 = *Typical Angina*,1 = *Atypical Angina*, 2 = *Non Angina Pain*, 3 = *Asymptomatic*
4	Trestbps	Tekanan darah saat istirahat	mm/hg
5	Chol	Kadar kolestrol	mg/dl
6	Fbs	Kadar gula darah	0 = *No*, 1 = *Yes*
7	restecg	Hasil test elektrokardiografi	0 = *Normal*, 1 = *Abnormal*, 2 = *Probable*
8	Thalach	Detak jantung maksimal	-
9	Exang	Nyeri saat olahraga	0 = *No*, 1 = *Yes*
10	Oldpeak	Depresi ST disebabkan oleh olahraga dibandingkan dengan istirahat	-
11	Slope	Slope puncak ST setelah berolahraga	0 = *Upsloping*, 1 = *Flat*, 2 = *Downsloping*
12	Ca	Jumlah pembuluh darah	0 - 4
13	Thal	Thallium	1 = *Normal*, 2 = *Fixed Defect*, 3 = *Reversable Defect*
14	Target	Terkena penyakit jantung	0 = *No*, 1 = *Yes*

TABEL 3
ATRIBUT HEART DISEASE CLEVELAND

No	Atribut	Deskripsi	Keterangan
1	Age	Usia Pasien	Usia dalam tahun
2	Sex	Jenis Kelamin	0 = Perempuan,1 = Laki-laki
3	Cp	Jenis nyeri dada	0 = *Typical Angina*,1 = *Atypical Angina*, 2 = *Non Angina Pain*, 3 = *Asymptomatic*
4	Trestbps	Tekanan darah saat istirahat	mm/hg
5	Chol	Kadar kolestrol	mg/dl
6	Fbs	Kadar gula darah	0 = *No*, 1 = *Yes*
7	restecg	Hasil test elektrokardiografi	0 = *Normal*, 1 = *Abnormal*, 2 = *Probable*
8	Thalach	Detak jantung maksimal	-
9	Exang	Nyeri saat olahraga	0 = *No*, 1 = *Yes*
10	Oldpeak	Depresi ST disebabkan oleh olahraga dibandingkan dengan istirahat	-
11	Slope	Slope puncak ST setelah berolahraga	0 = *Upsloping*, 1 = *Flat*, 2 = *Downsloping*
12	Ca	Jumlah pembuluh darah	0 - 4
13	Thal	Thallium	1 = *Normal*, 2 = *Fixed Defect*, 3 = *Reversable Defect*
14	Condition	Terkena penyakit jantung	0 = *No*, 1 = *Yes*

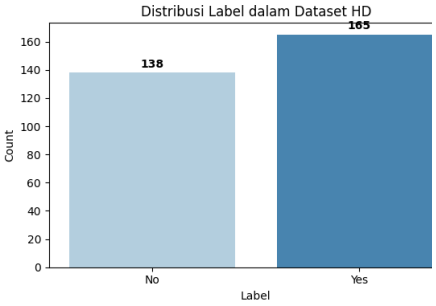
TABEL 4
ATRIBUT HEART DISEASE PREDICTION

No	Atribut	Deskripsi	Keterangan
1	Age	Usia Pasien	Tahun
2	Sex	Jenis Kelamin	0 = Perempuan, 1 = Laki-laki
3	Cp	Jenis nyeri dada	1 = Typical Angina, 2 = Atypical Angina, 3 = Non Angina Pain, 4 = Asymptomatic
4	Trestbps	Tekanan darah saat istirahat	-
5	Chol	Kadar kolestrol	mg/dl
6	Fbs	Kadar gula darah > 120 mg/dl	0 = *No*, 1 = *Yes*
7	Restecg	Hasil test elektrokardiografi	0 = Normal, 1 = *Abnormal*, 2 = *Probable*
8	Thalach	Detak jantung maksimal	-
9	Exang	Nyeri saat olahraga	0 = *No*, 1 = *Yes*
10	Oldpeak	Depresi ST disebabkan oleh olahraga dibandingkan dengan istirahat	-
11	Slope	Slope puncak ST setelah berolahraga	1 = *Upsloping*, 2 = *Flat*, 3 = *Downsloping*
12	Ca	Jumlah pembuluh darah	0 - 3
13	Thal	Thallium	3 = normal, 6 = *fixed defect*, 7 = *reversable defect*
14	Heart Disease	Terkena penyakit jantung	1 = *No*, 2 = *Yes*

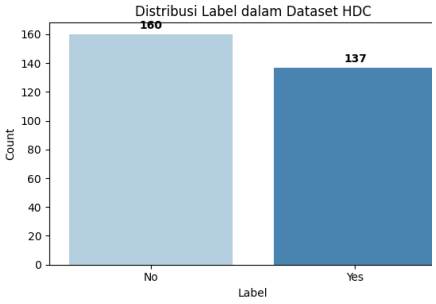
Pada Tabel 4 *dataset Heart Disease Prediction* beberapa atribut dilakukan transformasi nilai agar sesuai dengan skema pengkodean pada *dataset Heart Disease* dan *Heart Disease Cleveland*. Transformasi ini bertujuan untuk menyeragamkan

representasi data kategorikal sehingga proses penggabungan *dataset* tidak menimbulkan perbedaan interpretasi nilai serta dapat mendukung proses pemodelan secara optimal.

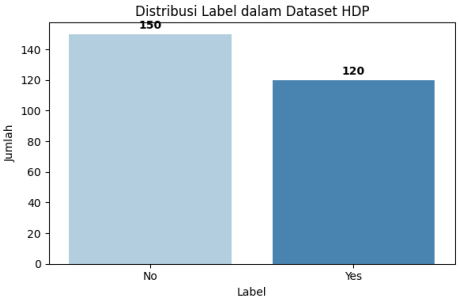
Setelah data dikumpulkan, data tersebut dilakukan proses *load dataset* atau pemuatan data ke dalam lingkungan pemrograman *Python* menggunakan *library Pandas.* Proses ini bertujuan untuk mengimpor *dataset* dalam format *.csv* agar dapat diolah dan dianalisis lebih lanjut. Kemudian, dilakukan pemeriksaan distribusi kelas pada label dari ketiga *dataset*. Pemeriksaan ini penting untuk mengidentifikasi adanya ketidakseimbangan kelas (*class imbalance*), yang dapat memengaruhi kinerja model selama proses pelatihan dan evaluasi.



Gambar 4 Distribusi Kelas HD

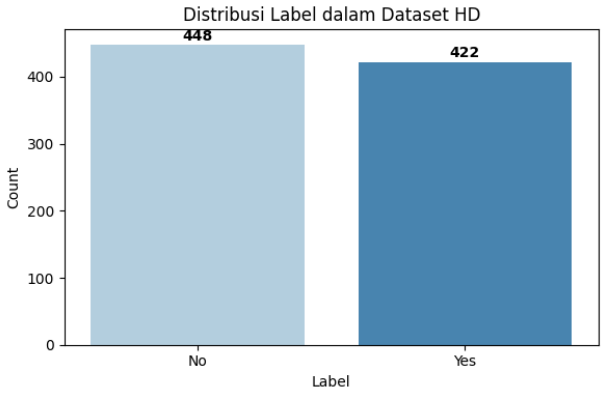


Gambar 5 Distribusi Kelas HDC



Gambar 6 Distribusi Kelas HDP

Gambar 4, Gambar 5, dan Gambar 6, menunjukkan variasi jumlah pasien penyakit jantung antara label *Yes* dan *No* pada *dataset*, namun selisihnya tidak terlalu signifikan sehingga masih dapat dikategorikan sebagai relatif seimbang. Hal ini penting karena ketidakseimbangan kelas yang ekstrem dapat menyebabkan model lebih cenderung memprediksi kelas mayoritas, sehingga mengurangi kemampuan generalisasi terutama pada kelas minoritas. Gambar 7 memperlihatkan distribusi label atau kelas pada *dataset* gabungan, yang merupakan hasil penggabungan tiga *dataset* HD, HDC, dan HDP. Pada *dataset* gabungan ini, jumlah pasien dengan label *Yes* dan *No* tetap menunjukkan perbedaan yang tidak terlalu besar. Meskipun total datanya lebih banyak karena merupakan akumulasi dari tiga sumber, komposisi kedua kelas masih berada dalam rentang yang dapat dikategorikan relatif seimbang.



Gambar 7 Distribusi Kelas Gabungan

B. Preprocessing

Tahap selanjutnya adalah melakukan *preprocessing* data untuk memastikan data dalam kondisi yang sesuai untuk digunakan dalam proses pelatihan model. Berikut tahapan *preprocessing* data:

1) Penyesuaian Tipe Data

Penyesuaian tipe data dilakukan untuk memastikan setiap atribut dalam *dataset* memiliki format yang sesuai dengan kebutuhan proses analisis. Penyesuaian ini meliputi konversi tipe data numerik dan kategorikal agar dapat dikenali dengan benar oleh algoritma pemodelan yang digunakan.

TABEL 5
PERBANDINGAN DATASET SEBELUM KONVERSI TIPE DATA

Dataset	Jumlah Baris	Jumlah Kolom	Kolom Target	Tipe Data	
				Jumlah int64	Jumlah float64
HD (*Heart Disease*)	303	14	*target*	13	1
HDC (*Heart Disease Cleveland*)	297	14	*condition*	13	1
HDP (*Heart Disease Prediction*)	270	14	*heart_disease*	13	1
Dataset Gabungan	870	14	*target*	13	1

Berdasarkan hasil eksplorasi data sebelum konversi tipe data pada Tabel 5, seluruh *dataset* memiliki 14 atribut yang didominasi oleh tipe data int64 sebanyak 13 atribut dan satu atribut bertipe float64, yaitu *oldpeak*. Meskipun secara umum seluruh atribut telah terbaca sebagai numerik, beberapa atribut yang bersifat kategorikal (seperti sex, cp, fbs, restecg, exang, slope, ca, dan thal) masih bertipe numerik sehingga memerlukan penyesuaian agar dapat direpresentasikan sesuai karakteristik datanya dalam proses pemodelan.

TABEL 6
KONDISI DATASET SESUDAH KONVERSI TIPE DATA

Dataset	Tipe Data		Jumlah object	Kolom Target
	Jumlah int64	Jumlah float64		
HD	4	1	9	target (object)
HDC	4	1	9	condition (object)
HDP	4	1	9	heart_disease (object)
Gabungan	4	1	9	target (object)

Setelah dilakukan penyesuaian tipe data dengan hasil pada Tabel 6, atribut yang bersifat kategorikal seperti sex, cp, fbs, restecg, exang, slope, ca, thal, serta atribut target dikonversi menjadi tipe object. Perubahan ini menghasilkan 4 atribut bertipe int64, 1 atribut bertipe float64, dan 9 atribut bertipe object pada masing-masing *dataset*. Penyesuaian ini dilakukan agar setiap atribut sesuai dengan karakteristik datanya sehingga proses pemodelan menggunakan algoritma CART dan teknik *Bagging* dapat berjalan secara optimal.

Kemudian, memastikan bahwa algoritma pemodelan dapat memproses atribut kategorikal dengan benar. Karena sebagian besar algoritma pembelajaran mesin tidak dapat menangani data kategorikal secara langsung, maka dilakukan proses *encoding*. Salah satu teknik yang digunakan adalah *one hot encoding*, yang mengubah atribut kategorikal menjadi representasi numerik biner agar dapat dibaca dan diproses oleh model. *One hot encoding* dilakukan terhadap atribut-atribut dengan tipe data *object* yang dianggap nominal seperti *cp*, *restecg*, dan *thal*. Sedangkan atribut-atribut lain pada kategori tipe data *object* yang dianggap ordinal seperti *sex*, *fbs*, *exang*, *slope*, *ca*, target (label HD), *condition* (*label* HDC), dan *heart disease* (label HDP) dikonversi ke tipe data *integer*. Hal ini dilakukan karena atribut-atribut tersebut bersifat biner atau ordinal, sehingga dapat diwakili dengan angka tanpa perlu *one hot encoding*. Berikut ditampilkan hasil dari proses *one hot encoding* yang telah diterapkan pada atribut-atribut nominal ketiga *dataset*:

TABEL 7
HASIL ONE HOT ENCODING SELURUH DATASET

Dataset	Atribut nominal		
	cp	restecg	thal
HD	cp_1, cp_2, cp_3	restecg_1, restecg_2	thal_1, thal_2, thal_3
HDC	cp_1, cp_2, cp_3	restecg_1, restecg_2	thal_1, thal_2
HDP	cp_1, cp_2, cp_3	restecg_1, restecg_2	thal_2, thal_3
Gabungan	cp_1, cp_2, cp_3	restecg_1, restecg_2	thal_1, thal_2, thal_3

Berdasarkan hasil *one hot encoding* pada Tabel 7, atribut nominal yaitu cp, restecg, dan thal dikonversi menjadi beberapa kolom biner pada masing-masing *dataset*. Sebagai contoh, atribut cp (*chest pain type*) yang merepresentasikan jenis nyeri dada dan memiliki beberapa kategori, diubah menjadi kolom cp_1 (*Atypical Angina*), cp_2 (*Non-Anginal Pain*), dan cp_3 (*Asymptomatic*). Nilai pada setiap kolom hasil *one hot encoding* berupa *True* atau *False*, yang menunjukkan keberadaan suatu kategori pada setiap baris data. Nilai *True* menandakan bahwa data tersebut termasuk dalam kategori yang diwakili oleh kolom tersebut, sedangkan *False* menunjukkan bahwa data tersebut tidak termasuk

dalam kategori tersebut. Setiap kategori direpresentasikan dalam kolom terpisah untuk menghindari asumsi adanya hubungan ordinal antar kategori. Perbedaan jumlah kolom hasil *encoding* pada atribut thal menunjukkan adanya variasi kategori yang muncul pada masing-masing *dataset*. Melalui proses ini, seluruh atribut nominal telah dikonversi ke dalam bentuk numerik biner sehingga dapat digunakan secara optimal dalam tahap pemodelan menggunakan algoritma CART dan teknik *Bagging*.

2) Handling Missing Values

Proses pemeriksaan dilakukan terhadap keberadaan *missing values* pada masing-masing *dataset*. Pemeriksaan ini bertujuan untuk memastikan bahwa tidak terdapat data yang hilang yang dapat memengaruhi hasil analisis dan performa model. Hasil deteksi *missing values* pada *dataset* HD, HDC, HDP, dan *dataset* gabungan disajikan pada Tabel 8.

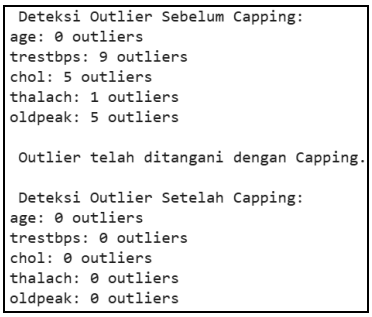
TABEL 8
HASIL DETEKSI MISSING VALUES

Dataset	**Jumlah Atribut**	**Total *Missing Values***	**Keterangan**
HD	14	0	Tidak terdapat *missing values*
HDC	14	0	Tidak terdapat *missing values*
HDP	14	0	Tidak terdapat *missing values*
Gabungan	14	0	Tidak terdapat *missing values*

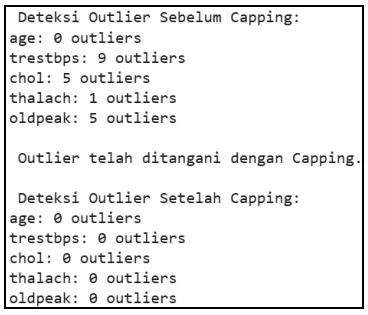
Berdasarkan hasil deteksi pada Tabel 8, seluruh atribut pada masing-masing *dataset* menunjukkan nilai 0, yang berarti tidak terdapat data yang hilang. Dengan demikian, tidak diperlukan proses penanganan lanjutan seperti imputasi maupun penghapusan data. *Dataset* yang digunakan dalam penelitian ini dinyatakan lengkap dan siap untuk diproses pada tahap pemodelan.

3) Handling Outlier

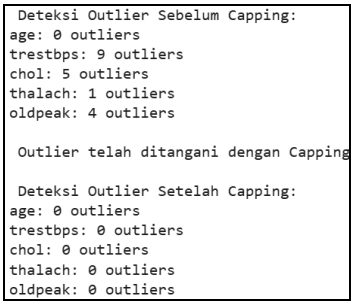
Proses identifikasi *outlier* pada ketiga *dataset* dilakukan dengan menggunakan metode *Interquartile Range* (IQR). Nilai-nilai yang berada di luar batas bawah (Q1 - 1.5*IQR) dan batas atas (Q3 + 1.5*IQR) dianggap sebagai *outlier.* Sebagai alternatif dari penghapusan data, pendekatan capping digunakan untuk mempertahankan integritas *dataset*. Nilai yang melebihi batas atas diganti dengan nilai maksimum yang masih dianggap wajar, sedangkan nilai di bawah batas bawah diganti dengan nilai minimum yang dapat diterima. Pendekatan ini memungkinkan data tetap utuh tanpa kehilangan informasi penting, serta menjaga keseimbangan distribusi data dalam proses pelatihan model.



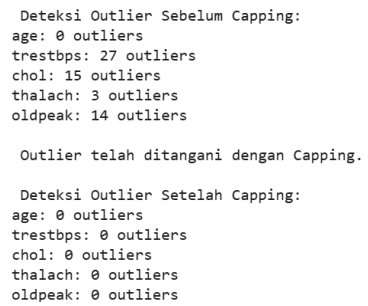
Gambar 8 Outlier HD



Gambar 9 *Outlier* HDC



Gambar 10 *Outlier* HDP



Gambar 11 *Outlier* Gabungan

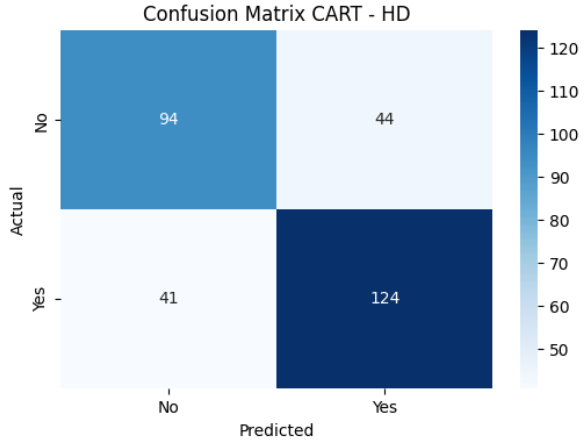
Berdasarkan Gambar 8, Gambar 9, Gambar 10, dan Gambar 11, terlihat bahwa sebelum dilakukan *capping* masih terdapat *outlier* pada beberapa atribut. Setelah proses *capping* diterapkan, seluruh atribut pada keempat *dataset* menunjukkan tidak adanya *outlier*, sehingga data menjadi lebih bersih dan siap untuk tahap analisis selanjutnya.

C. *Pemodelan*

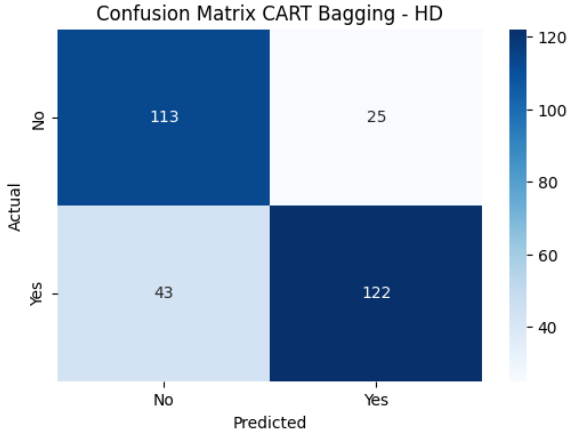
Pada tahap ini dilakukan proses pembangunan model klasifikasi untuk mendiagnosis penyakit jantung berdasarkan data yang telah melalui tahap *preprocessing*. Pada penelitian ini, pemodelan dilakukan dengan menggunakan dua pendekatan, yaitu algoritma CART sebagai model dasar, serta CART yang dikombinasikan dengan teknik *Bagging* sebagai model pengembangannya. Model CART digunakan untuk membentuk pohon keputusan yang mampu melakukan klasifikasi berdasarkan pemilihan atribut terbaik menggunakan ukuran tertentu, sehingga menghasilkan aturan-aturan keputusan yang mudah dipahami. Selanjutnya, untuk meningkatkan stabilitas dan performa model, diterapkan teknik *Bagging* dengan cara membangun beberapa model CART pada data latih yang diambil secara acak menggunakan metode *bootstrap sampling*. Hasil dari beberapa model tersebut kemudian digabungkan melalui mekanisme *majority voting* untuk menghasilkan keputusan akhir.

D. *Confusion Matrix*

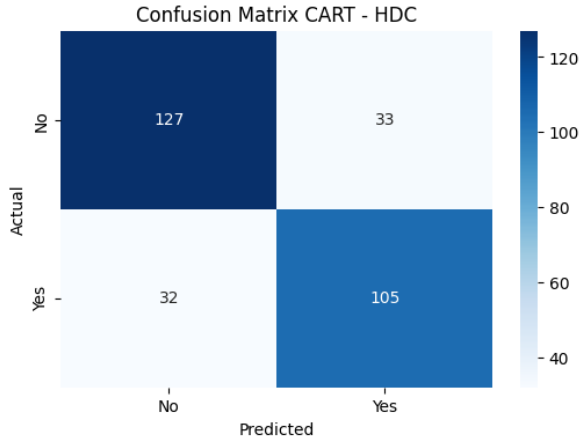
Confusion matrix menyajikan informasi mengenai jumlah prediksi yang benar dan salah untuk masing-masing kelas, sehingga dapat memberikan gambaran lebih rinci terhadap kinerja model.



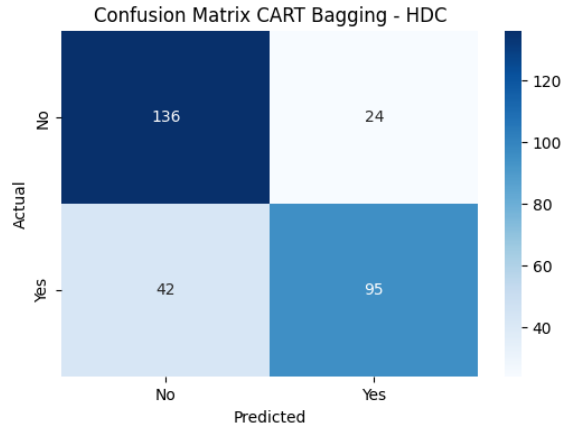
Gambar 12 Confusion Matrix CART HD



Gambar 13 Confusion Matrix CART Bagging HD

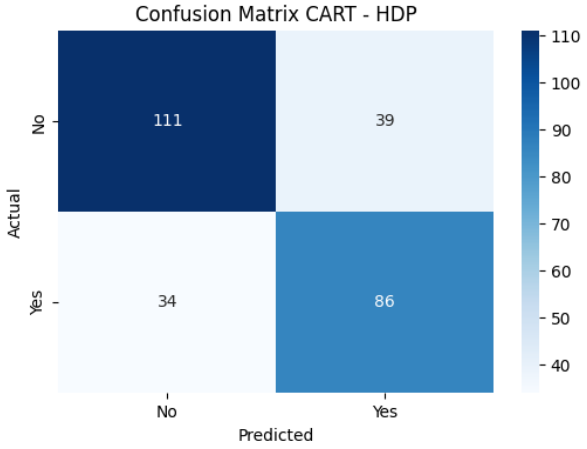


Gambar 14 *Confusion Matrix* CART HDC

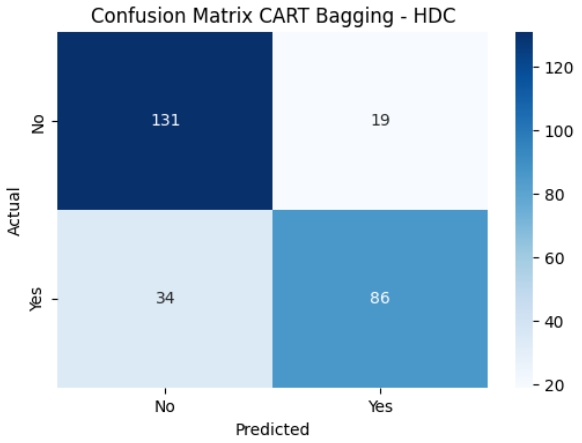


Gambar 15 *Confusion Matrix* CART *Bagging* HDC

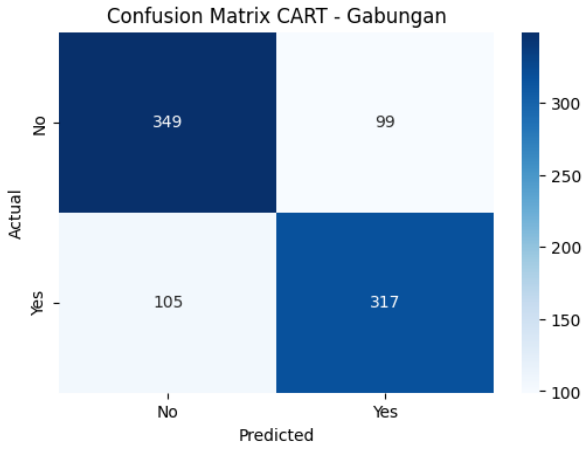
p-ISSN : 2443-2210
e-ISSN : 2443-2229



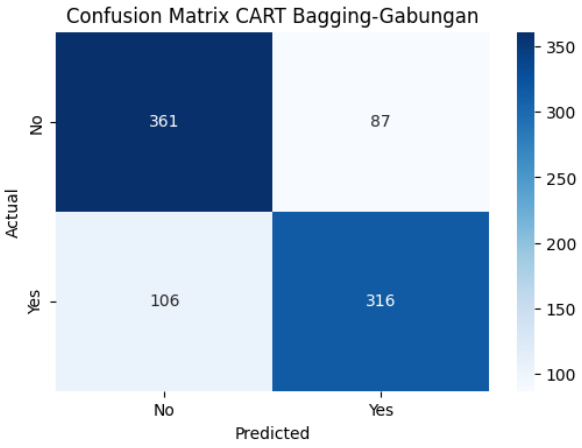
Gambar 16 *Confusion Matrix* CART HDP



Gambar 17 Confusion Matrix CART Bagging HD



Gambar 18 *Confusion Matrix* CART Gabungan



Gambar 19 *Confusion Matrix* CART *Bagging* Gabungan

Berdasarkan hasil *confusion matrix* pada ketiga *dataset* individual yang terdapat pada Gambar 12, Gambar 13, Gambar 14, Gambar 15, Gambar 16, dan Gambar 17, terlihat bahwa model CART secara konsisten menunjukkan kemampuan yang lebih baik dalam mendeteksi pasien yang benar-benar memiliki penyakit jantung. Hal ini ditunjukkan oleh nilai TP yang lebih tinggi dan FN yang lebih rendah dibandingkan model *Bagging*. Dengan demikian, CART memiliki sensitivitas yang lebih kuat pada ketiga *dataset* tersebut, yang berarti model lebih mampu menangkap pola-pola yang mengindikasikan adanya penyakit jantung. Namun, CART masih menghasilkan jumlah FP yang relatif lebih tinggi, sehingga beberapa pasien yang sebenarnya sehat terprediksi sebagai sakit. Ketika teknik *Bagging* diterapkan, terjadi peningkatan yang konsisten pada nilai TN dan penurunan FP pada seluruh *dataset* individual. Hal ini mengindikasikan bahwa *Bagging* mampu meningkatkan kemampuan model dalam mengidentifikasi pasien sehat dengan lebih akurat melalui peningkatan *specificity*. Meski demikian, peningkatan *specificity* tersebut disertai dengan meningkatnya FN dan menurunnya TP, yang menunjukkan bahwa Bagging menjadi lebih konservatif dalam memberikan prediksi positif. Kondisi ini mencerminkan karakteristik dasar *Bagging* yang mengurangi varians model, sehingga prediksi menjadi lebih stabil pada kelas negatif, namun sekaligus mengorbankan sensitivitas pada kelas positif.

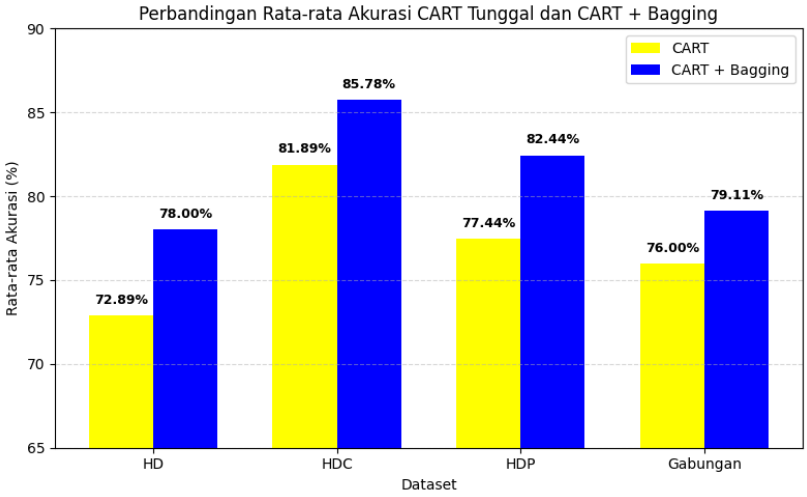
Pada *dataset* gabungan di Gambar 18 dan Gambar 19 pola performa model menunjukkan kecenderungan yang relatif serupa, tetapi dengan perbedaan yang lebih menonjol akibat ukuran data yang lebih besar dan heterogen. Model CART pada *dataset* gabungan menghasilkan nilai TP dan TN yang cukup tinggi, menandakan bahwa model masih mampu bekerja dengan baik dalam mendeteksi kedua kelas. Namun, jumlah FN yang masih cukup besar menunjukkan bahwa CART mengalami kesulitan dalam menangkap pola pasien sakit ketika data menjadi lebih kompleks dan bervariasi. Saat teknik *Bagging* diterapkan pada *dataset* gabungan, peningkatan TN dan penurunan FP kembali terlihat sangat konsisten,

menguatkan bukti bahwa Bagging memang efektif dalam meningkatkan kestabilan prediksi pada kelas negatif. Akan tetapi, Bagging kembali mengalami peningkatan pada FN dan sedikit penurunan TP, yang berarti sensitivitas model tidak membaik meskipun jumlah data lebih besar. Dengan demikian, Bagging tetap menunjukkan kecenderungan konservatif dalam memprediksi kelas positif, sehingga lebih sering gagal mengenali pasien yang benar-benar memiliki penyakit jantung.

Perbandingan antara *dataset* individual dan *dataset* gabungan menunjukkan bahwa CART lebih optimal ketika pola data lebih sederhana dan tidak terlalu heterogen, seperti pada *dataset* individual, sedangkan performanya semakin menurun dalam mendeteksi pasien sakit ketika *dataset* diperbesar dan digabungkan. Sebaliknya, *Bagging* menunjukkan kestabilan performa baik pada *dataset* mandiri maupun gabungan, terutama dalam peningkatan TN, tetapi tetap tidak mampu menekan jumlah FN. Kekurangan model pada seluruh pengujian ini disebabkan oleh beberapa faktor, antara lain tumpang tindih nilai fitur antar kelas (*class overlap*), tingkat heterogenitas yang tinggi pada *dataset* gabungan, serta karakteristik CART yang rentan mengalami *overfitting* dan karakteristik *Bagging* yang cenderung memperkuat pola kelas negatif. Untuk mengatasi hal ini, beberapa pendekatan dapat diterapkan, seperti menerapkan metode penyeimbangan kelas untuk menurunkan FN, menggunakan teknik *boosting* yang mampu memperbaiki kesalahan secara iteratif, mengoptimalkan *threshold* prediksi untuk meningkatkan sensitivitas, melakukan normalisasi dan seleksi fitur. Secara keseluruhan, hasil ini menunjukkan bahwa CART lebih unggul dari sisi sensitivitas, sedangkan *Bagging* lebih kuat dari sisi *specificity*, dan keduanya memiliki *trade-off* yang perlu dipertimbangkan dalam konteks diagnosis medis yang sangat sensitif terhadap kesalahan prediksi.

E. Evaluasi Hasil

Pada tahap ini dilakukan evaluasi terhadap performa akhir dari masing-masing model klasifikasi yang telah dibangun. Hasil evaluasi divisualisasikan pada grafik di Gambar 20 yang menunjukkan perbandingan akurasi kedua model.



Gambar 20 Hasil Perbandingan Akurasi

Secara konsisten, teknik *Bagging* memberikan peningkatan akurasi pada seluruh *dataset*. Pada *dataset* HD akurasi meningkat dari 72,89% menjadi 78%, pada HDC dari 81,89% menjadi 85,78%, pada HDP dari 77,44% menjadi 82,44%, dan pada *dataset* gabungan dari 76% menjadi 79,11%.

Peningkatan akurasi tersebut tidak hanya terlihat secara agregat, tetapi juga tergambar pada analisis *confusion matrix*. Pada seluruh *dataset*, *Bagging* menurunkan jumlah *False Positive* (FP), yaitu kasus ketika model memprediksi pasien memiliki penyakit padahal sebenarnya tidak. Penurunan FP ini terutama terjadi pada data dengan karakteristik variabel yang saling berkorelasi rendah dan mengandung *outlier*, seperti pada fitur kolesterol dan tekanan darah pada *dataset* HD dan HDC. CART tunggal cenderung mudah terpengaruh oleh nilai-nilai ekstrem tersebut sehingga menghasilkan keputusan yang tidak stabil. *Bagging* memperbaiki kondisi ini melalui proses agregasi banyak pohon, sehingga variansi model menurun yang artinya model menjadi kurang sensitif terhadap fluktuasi nilai input dan tidak mudah berubah hanya karena pergeseran kecil pada data pelatihan. Contohnya, pada *dataset* HDC, *Bagging* menghasilkan lebih banyak *True Negative* (TN) dibanding CART tunggal karena banyak pohon dalam *ensemble* menyepakati batas keputusan yang lebih stabil pada kelas negatif, sehingga FP dapat ditekan.

Namun demikian, penurunan variansi ini dibarengi dengan kecenderungan meningkatnya *False Negative* (FN). Peningkatan FN terjadi pada beberapa kondisi di mana nilai fitur pasien cenderung berada pada ambang batas (*borderline*), misalnya nilai detak jantung dan tekanan darah yang mendekati titik *cut-off*. Dalam kasus seperti ini, pohon-pohon dalam *Bagging* cenderung lebih konservatif karena mayoritas pohon menghasilkan prediksi negatif. Sementara itu, CART tunggal justru memiliki sensitivitas lebih tinggi sehingga lebih sering memprediksi kelas positif, meskipun berisiko menghasilkan FP lebih banyak. Dengan demikian, karakteristik data *borderline* menjadi faktor dominan yang menyebabkan FN meningkat pada model *Bagging*.

Dari sisi stabilitas prediksi, *Bagging* menunjukkan keunggulan yang semakin terlihat pada *dataset* gabungan yang memiliki heterogenitas paling tinggi. CART tunggal cenderung membentuk struktur pohon yang berbeda-beda pada setiap lipatan *cross-validation*, sehingga variansinya besar dan performa sulit dipertahankan secara konsisten. *Bagging* mampu menurunkan variansi tersebut melalui mekanisme *bootstrapping* dan agregasi mayoritas, sehingga prediksi menjadi lebih stabil dan akurasi meningkat pada sebagian besar kelas.

Jika dibandingkan dengan eksperimen lain pada *platform Kaggle* yang menggunakan *dataset* serupa, model dengan metode *ensemble* seperti *Random Forest* dan *Gradient Boosting* biasanya menghasilkan akurasi pada kisaran 82–90%. Model *Bagging* CART yang digunakan pada penelitian ini berada pada rentang yang kompetitif, terutama pada *dataset* HDC dan *dataset* gabungan. Keunggulan model dalam penelitian ini terletak pada interpretabilitas yang tetap terjaga, karena setiap pohon pada *Bagging* masih berbasis CART, serta stabilitas prediksi yang meningkat akibat berkurangnya variansi. Dengan demikian, dari sisi konsep, *Bagging* berhasil meningkatkan keandalan CART pada data medis yang bersifat heterogen tanpa mengorbankan transparansi model.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa *Bagging* CART unggul dari sisi akurasi keseluruhan, stabilitas prediksi, dan kemampuan menekan FP. CART tunggal tetap memiliki keunggulan dalam sensitivitas terhadap kelas positif, namun kurang stabil pada dataset yang memiliki distribusi fitur yang bervariasi. *Trade-off* ini menunjukkan bahwa pemilihan model harus disesuaikan dengan konteks klinis, apakah lebih diprioritaskan pengurangan salah deteksi pasien sehat (FP) atau peningkatan deteksi pasien sakit (TP).

IV. SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, penerapan teknik *Bagging* pada algoritma CART terbukti mampu meningkatkan kinerja model dalam mendiagnosis penyakit jantung. Evaluasi dilakukan terhadap tiga *dataset* individual dan *dataset* gabungan, yaitu *Heart Disease*, *Heart Disease Cleveland*, dan *Heart Disease Prediction*. Hasil pengujian menunjukkan bahwa *Bagging* secara konsisten memberikan peningkatan akurasi berdasarkan rata-rata nilai 10-*fold cross-validation*. Pada *dataset Heart Disease*, akurasi model meningkat dari 72,89% menjadi 78% setelah menggunakan Bagging. Pada *dataset Heart Disease Cleveland*, akurasi CART yang sebelumnya sebesar 81,89% meningkat menjadi 85,78% setelah diterapkan *Bagging*. Peningkatan serupa terjadi pada *dataset Heart Disease Prediction*, di mana akurasi meningkat dari 77,44% menjadi 82,44%. Pada *dataset* gabungan yang memiliki jumlah data lebih besar dan lebih heterogen, *Bagging* kembali menunjukkan keunggulan dengan meningkatkan akurasi model dari 76% menjadi 79,11%. Selain peningkatan akurasi, analisis *confusion matrix* menunjukkan bahwa *Bagging* mampu meningkatkan *specificity* model melalui penurunan jumlah prediksi *false positive*. Namun, peningkatan ini diikuti oleh kecenderungan meningkatnya *false negative* yang menunjukkan sedikit penurunan sensitivitas terhadap kelas positif. Secara keseluruhan, teknik *Bagging* menghasilkan performa yang lebih stabil dan akurat dibandingkan CART tunggal, terutama pada data yang lebih kompleks. Temuan ini menunjukkan bahwa *Bagging* dapat menjadi pendekatan yang efektif dalam meningkatkan performa algoritma pohon keputusan untuk mendukung proses diagnosis penyakit jantung.

DAFTAR PUSTAKA

[1] WHO, "Cardiovascular diseases (CVDs)," World Health Organization. Accessed: Feb. 19, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Kemkes, "Penyakit Jantung Koroner Didominasi Masyarakat Kota," Sehat Negeriku Sehatlah Bangsaku. Accessed: Feb. 19, 2025. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/umum/20210927/5638626/penyakit-jantung-koroner-didominasi-masyarakat-kota/
[3] R. W. Aisya, L. Dharmawati, and D. P. Dyah K, "Hubungan Kebiasaan Konsumsi Makanan Cepat Saji Dan Kejadian Penyakit Jantung Koroner Pada Pasien Rawat Jalan Di Rsud Dr. Moewardi," *J. Med. Indones.*, vol. 2, no. 2, pp. 21–28, 2021.
[4] M. Ilham Aziz, A. Z. Fanani, and A. Affandy, "Analisis Metode Ensemble Pada Klasifikasi Penyakit Jantung Berbasis Decision Tree," *J. Media Inform. Budidarma*, vol. 7, no. 1, p. 1, 2023.
[5] M. J. A. Berry and D. S. Linoff, *Data Mining Techniques*, Second Edi. Canada: WILEY, 2018.
[6] H. Dahan, S. Cohen, L. Rokach, and O. Maimon, *Proactive Data Mining with Decision Tree*, 1st ed. New York: Springer, 2014.
[7] N. Ye, *Data Mining: Theories, Algorithms, and Examples*, Human Fact. London New York: CRC Press, 2013.
[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts, Models, Methods, and Algorithms*, Third. Elsevier, 2012.
[9] S. Innassuraiya, T. Widiharih, and I. T. Utami, "Analisis Klasifikasi Menggunakan Metode Regresi Logistik Biner dan Bootstrap Aggregating Classification And Regression Trees (Bagging CART) (Studi Kasus: Nasabah Koperasi Simpan Pinjam Dan Pembiayaan Syariah (KSPPS))," *J. Gaussian*, vol. 11, no. 2, pp. 183–194, 2022.
[10] A. S. R. Siregar, Y. S. Siregar, and M. Khairani, "Implementation Of The Data Mining Cart Algorithm In The Characteristic Pattern Of New

Student Admissions," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 5, no. 1, pp. 263–275, 2023.
[11] B. A. Saputra, E. Kurnia, M. Rahmah, and T. Sumarni, "Penerapan Privasi Dan Etika Di Era Digital Dalam Perlindungan Data Pribadi," *Musytari Neraca Manajemen, Akuntansi, dan Ekon.*, vol. 5, no. 9, pp. 55–65, 2024.
[12] A. Fadillah Hermawan, F. Rakhmat Umbara, and F. Kasyidi, "Prediksi Awal Penyakit Stroke Berdasarkan Rekam Medis menggunakan Metode Algoritma CART(Classification and Regression Tree)," *J. MIND*, vol. 7, no. 2, pp. 151–164, 2022.
[13] A. Jananto, S. Sulastri, E. Nur Wahyudi, and S. Sunardi, "Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus Menggunakan Algoritma CART Klasifikasi Data Mining," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 10, no. 1, pp. 71–78, 2021.
[14] R. Kustiawan, A. Adiwijaya, and M. D. Purbolaksono, "A Multi-label Classification on Topic of Hadith Verses in Indonesian Translation using CART and Bagging," *J. Media Inform. Budidarma*, vol. 6, no. 2, p. 868, 2022.
[15] A. N. Ikhsan, A. N. Fadilah, and A. Dafa Iftinani, "Performance Comparison of Decision Tree J48, CART and Naïve Bayes Algorithms for Predicting Chronic Kidney Disease," *Indones. J. Artif. Intell. Data Min.*, vol. 7, no. 1, pp. 64–70, 2024.
[16] J. Yulinda, R. S. Lubis, and R. Aprilia, "Penggunaan Metode Classification Analysis Regression Trees dan Iterative Dichotomizer 3 Dalam Mengklasifikasikan Pasien Hipertensi Di Rumah Sakit Umum Daerah Dr. Pirngadi KotaMedan," *Justek J. Sains dan …*, vol. 6, no. 4, pp. 482–492, 2023.
[17] L. Muflikhah, F. A. Bachtiar, D. E. Ratnawati, and R. Darmawan, "Improving Performance for Diabetic Nephropathy Detection Using Adaptive Synthetic Sampling Data in Ensemble Method of Machine Learning Algorithms," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 1, p. 123, 2024.
[18] A. Nugroho and Y. Religia, "Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 3, pp. 504–510, 2021.
[19] V. I. Yani, A. Aradea, and H. Mubarok, "Optimasi Prakiraan Cuaca Menggunakan Metode Ensemble pada Naïve Bayes dan C4.5," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 3, pp. 607–619, 2022.
[20] L. H. Y. Arini, S. Solimun, A. Efendi, and M. O. Ullah, "CART Classification on Ordinal Scale Data with Unbalanced Proportions using Ensemble Bagging Approach," *JTAM (Jurnal Teor. dan Apl. Mat.*, vol. 8, no. 2, p. 441, 2024.
[21] K. Pujiyanti, "Optimasi Metode CART Menggunakan Metode Bagging Pada Studi Kasus Data Imbalance Berbasis Metode Adasyn," *J. Ilm. Mat.*, vol. 10, no. 1, pp. 34–42, 2023.
[22] S. Indah Nurhafida and F. Sembiring, "Analisis Text Clustering Masyarakat di Twiter mengenai Mcdonald'Sxbts menggunakan Orange Data Mining," *SISMATIK (Seminar Nas. Sist. Inf. dan Manaj. Inform.*, pp. 28–35, 2021.
[23] A. A. Syam, G. H. M, A. Salim, D. F. Surianto, and M. F. B, "Analisis teknik preprocessing pada sentimen masyarakat terkait konflik israel-palestina menggunakan support vector machine," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 3, pp. 1464–1472, 2024.
[24] M. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 393–399, 2021.
[25] A. Fitri Ariani, K. Aulia, and L. O. Ahmad Arafat, "Pengembangan Dashboard Interaktif Menggunakan Looker Studio Untuk Visualisasi Dan Prediksi Harga Komoditas Cabe Di Jawa Timur," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 4, pp. 8067–8074, 2024.
[26] A. Daniswara and I. K. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023.
[27] W. Sudrajat and I. Cholid, "K-Nearest Neighbor (K-NN) Untuk Penanganan Missing Value Pada Data Umkm," *J. Rekayasa Sist. Inf. dan Teknol.*, vol. 1, no. 2, pp. 54–63, 2023.
[28] W. Lan, X. Chen, T. Zou, and C. L. Tsai, "Imputations for High Missing Rate Data in Covariates Via Semi-supervised Learning Approach," *J. Bus. Econ. Stat.*, vol. 40, no. 3, pp. 1282–1290, 2022.
[29] F. Ayuning Tyas, M. Nurayuni, and H. Rakhmawati, "Optimasi Algoritma K-Nearest Neighbors Berdasarkan Perbandingan Analisis Outlier (Berbasis Jarak, Kepadatan, LOF)," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 13, no. 2, pp. 108–115, 2024.
[30] A. K. Jailani, A. Erna, T. Informatika, and U. Dipa, "Deteksi Anomali pada Rasio Jam Belajar dan Aktivitas Sosial terhadapPerforma Akademis Mahasiswa menggunakan Metode Local Outlier Factor (LOF)," *Semin. Ilm. Sist. Inf. dan Teknol. Inf.*, vol. XIV, no. 1, pp. 79–88, 2025.