Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072379
Vol. 7, No. 5, Oktober 2020, hlm. 1035-1040 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PENINGKATAN AKURASI KLASIFIKASI ALGORITMA C4.5 MENGGUNAKAN TEKNIK BAGGING PADA DIAGNOSIS PENYAKIT JANTUNG

Erwin Prasetyo*[1], Budi Prasetiyo[2]

[1,2]Universitas Negeri Semarang
Email: [1]erwinprasetyo@students.ac.id, [2] bprasetiyo@mail.unnes.ac.id
*Penulis Korespondensi

(Naskah masuk: 15 Agustus 2019, diterima untuk diterbitkan: 07 Oktober 2020)

Abstrak

Perkembangan teknologi yang begitu pesat menjadikan kebutuhan akan suatu informasi semakin meningkat, sehingga keakuratan suatu informasi menjadi suatu hal yang sangat penting, Terutama keakuratan informasi yang dibutuhkan dalam memprediksi penyakit dalam bidang medis. Dalam proses pengumpulan suatu informasi dibutuhkan metode tertentu, sehingga informasi yang telah diproses menjadi sebuah pengetahuan menggunakan suatu metode tertentu disebut dengan penambangan data atau istilah lainnya adalah *data mining*. Umumnya *data mining* digunakan untuk memprediksi suatu penyakit yang bersumber dari data rekam medis pasien, khususnya penyakit jantung. Data penyakit jantung diambil dari *dataset UCI Machine Learning Repository*. Tujuan dari penulis melakukan penelitian ini yaitu untuk mengetahui penerapan teknik *bagging* pada algoritma C4.5, mengetahui hasil akurasi dalam algoritma C4.5, dan membandingkan tingkat akurasi dari penerapan teknik *bagging* pada algoritma C4.5. *Dataset* yang diklasifikasikan dengan algoritma C4.5 memperoleh akurasi sebesar 72,98%. Hasil akurasi ini dapat ditingkatkan dengan menerapkan teknik bagging menghasilkan akurasi sebesar 81,84%, sehingga terjadi peningkatan akurasi sebesar 8,86% dari penerapan teknik *bagging* pada Algoritma C4.5.

Kata kunci: *Heart Disease, Bagging, Algoritma C4.5, Akurasi*

INCREASED CLASSIFICATION ACCURACY C4.5 ALGORITHM USING BAGGING TECHNIQUES IN DIAGNOSING HEART DISEASE

Abstract

The quick development of technology makes the need for information increase, so that the accuracy of the information becomes a very important thing, especially the accuracy of the information needed in predicting diseases in the medical field. In the process of gathering information certain methods are needed, so information that has been processed into knowledge using a certain method is called data mining or other terms is data mining. Data mining is generally used to predict a disease originating from patient medical record data, especially heart disease. Heart disease data is taken from the UCI Machine Learning Repository dataset. The purpose of the authors conducting this research is to determine the application of bagging techniques on the C4.5 algorithm, determine the accuracy of the results in the C4.5 algorithm, and compare the level of accuracy of the application of bagging techniques on the C4.5 algorithm. The dataset classified by the C4.5 algorithm obtained an accuracy of 72.98%. The results of this accuracy can be improved by applying bagging techniques resulting in an accuracy of 81.84%, resulting in an increase in accuracy of 8.86% from the application of bagging techniques in the C4.5 Algorithm.

Keywords: *Heart Disease, Bagging, C4.5 Algorithm, Accuracy*

1. PENDAHULUAN

Pesatnya perkembangan teknologi membuat kebutuhan akan informasi yang akurat sangat dibutuhkan dalam kehidupan sehari-hari dan di masa mendatang. Metode tradisional untuk menganalisis informasi yang ada, sebagian besar informasi tidak dapat diperbaiki. Oleh karena itu, informasi dapat diperbaiki atau diproses menghasilkan suatu pengetahuan dengan istilah *data mining* (Das, 2010). *Data mining* dapat diterapkan di berbagai bidang, salah satunya di bidang kesehatan. Pengambilan keputusan berdasarkan data dan informasi yang akurat akan menghasilkan keputusan dan prediksi penyakit yang menjadi target (Rohman, et al., 2017). Suatu penyakit yang memiliki data rekam medis pasien dapat diprediksi menggunakan istilah *data*

mining. Selain prediksi, metode dalam *data mining* adalah klasifikasi. Suatu hasil prediksi yang akurat khususnya pada penyakit jantung dibutuhkan data seperti jenis nyeri dada, kolesterol serum, jumlah pembuluh utama serta atribut lainnya. (Muzakir & Wulandari, 2016).

Penyakit jantung telah menjadi tantangan utama di berbagai industri pelayanan kesehatan. Penyakit ini merupakan salah satu alasan paling utama dari penyebab kematian di seluruh dunia dalam dekade terakhir (R & RV, 2016). Penyakit jantung merupakan pembunuh nomor satu di dunia. Setiap tahun, lebih dari 2 juta orang Amerika terkena penyakit jantung. Lebih dari 800.000 orang meninggal karena penyakit tersebut (Thomas R. Frieden, 2011). Permasalahan yang ada pada data jantung yaitu *imbalance class* atau kelas tidak seimbang. Kelas dikatakan tidak seimbang jika kelas positif hanya diwakili oleh beberapa tupel, sedangkan mayoritas tupel mewakili kelas negatif (Rohman, et al., 2017). Parameter kelas dikatakan tidak seimbang menurut Susan Lomax jika distribusi datanya 50/50 atau 40/60. Secara nyata, data yang ditambang dari *database* adalah data yang tidak seimbang (Siringoringo, 2018). *Bagging* merupakan sebuah metode sederhana namun efektif, sehingga hal tersebut membuat *bagging* banyak digunakan pada aplikasi di dunia nyata yang penggunaannya digabungkan dengan metode lain atau disebut juga metode *ensemble* (Liang, et al., 2011).

Penelitian yang dilakukan oleh Mirqotussa'adah (2017, p. 135) menghasilkan hasil akurasi yang meningkat dengan menggunakan algoritma C4.5 dengan menerapkan *discretization* dan teknik *bagging* pada *pima Indian dataset* sebesar 6,26 %. Dengan akurasi awal 68,61% setelah diterapkan *discretization* dan teknik *bagging* menjadi 74,87%. Sedangkan menurut Saifudin & Wahono (2015) dalam penelitannya menunjukkan bahwa *bagging* lebih baik daripada *adaboost*, karena dapat meningkatan sensivitas dan G-Mean secara signifikan pada *dataset software matrix*.

Dengan melihat penelitian sebelumnya bahwa penerapan teknik *bagging* dapat meningkatkan hasil akurasi klasifikasi, Sehingga dengan adanya penerapan teknik *bagging* dan algoritma dalam proses klasifikasi diharapkan dapat memperbaiki akurasi klasifikasi pada *dataset* medis, khususnya *Cleveland heart disease dataset*.

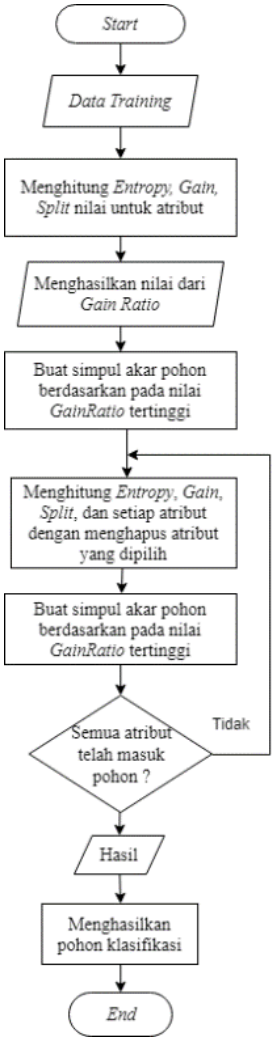
Berdasarkan uraian diatas diharapakan dengan adanya penerapan metode *bagging* dalam algoritma C4.5 dapat meningkatkan akurasi klasifikasi khususnya pada *dataset* penyakit jantung.

2. METODE PENELITIAN

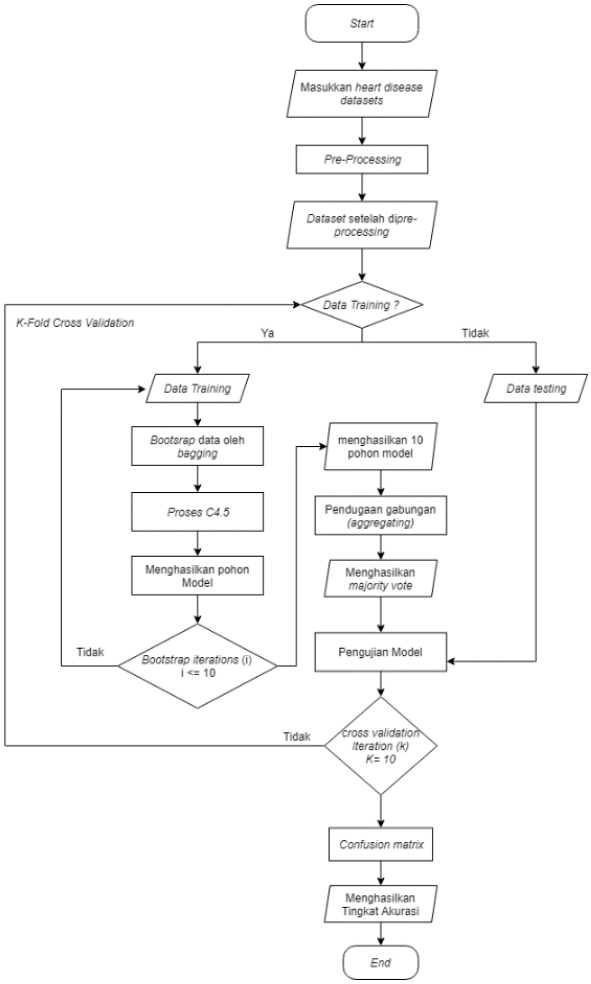
Metode penelitian yang digunakan oleh penulis yaitu menggunakan sebuah metode pada tahap *pre-processing* dengan cara melakukan proses *missing value handling* untuk membuat *dataset* lebih seimbang yang kemudian mengklasifikasikan *dataset* menggunakan algoritma C4.5 yang dikombinasikan dengan teknik *bagging* dengan tujuan meningkatkan hasil akurasi klasifikasi *heart disease dataset*, sehingga dapat diketahui perbandingan akurasi klasifikasi dari algoritma C4.5.

2.1 Algoritma C4.5

Algoritma C4.5 sebagai versi perbaikan ID3 merupakan sebuah algoritma yang diperkenalkan oleh Quinlan. Akan tetapi kelemahan hanya atribut bertipe kategorikal (nominal atau ordinal) saja yang bias di induksi oleh *decision tree*, sedangkan untuk menangani atribut bertipe numerik interval atau rasio tidak dapat menggunakan algoritma ID3. Sehingga dapat diketahui kelebihan algoritma C4.5 daripada algoritma ID3 antara lain, dapat menangani atribut dengan tipe numerik, memangkas pohon keputusan, dan menurunkan *rule set*. Penentuan fitur atau atribut sebagai pemecah simpul pada algoritma C4.5 menggunakan kriteria *gain* dalam pohon yang diinduksi (Prasetyo, 2014). Diagram alur dari algoritma C4.5 dapat dilihat seperti pada Gambar 1. Sedangkan diagram alur dari algoritma C4.5 dengan teknik *bagging* diilustrasikan sebagai Gambar 2.



Gambar 1. Diagram Alur Algoritma C4.5



Gambar 2. Diagram Alur Algoritma C4.5 dengan *Bagging*

Proses Algoritma C4.5 mencakup menghitung nilai entropi, mendapatkan dan memisahkan data latih dari masing-masing atribut untuk menghasilkan *gain ratio*.berikut merupakan persamaan untuk menghitung entropi, gain, split, dan *gain ratio* yang dapat dilihat pada persamaan 1, 2, 3 dan 4 berikut.

$$Entropy(S) = \sum_{i=1}^{n} -p_i * log_2 p_i \quad (1)$$

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

$$SplitEntropy_A(S) = -\sum_{i+1}^{n} \frac{|Si|}{|S|} * log_2 \frac{|Si|}{|S|} \quad (3)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(A)} \quad (4)$$

Keterangan:

S = himpunan kasus
A = fitur atau atribut
n = jumlah partisi atribut A
$|Si|$ = jumlah kasus pada partisi ke-i
$|S|$ = jumlah kasus dalam S

Cara memilih simpul akar atau *root* pada pada *tree* yaitu dengan membandingkan *gain ratio* tiap atribut sehingga untuk memilih simpul akar (*root*) pada atibut yang memiliki nilai *gain ratio* yang terbesar. Langkah selanjutnya, menghitung nilai entropi, dapatkan, dan pisahkan setiap atribut dengan cara menghapus atribut yang sebelumnya telah dipilih. Setelah membuat simpul akar langkah selanjutnya adalah membuat simpul internal dengan cara yang sama ketika kita membuat sebuah *root* atau akar yaitu dengan cara memilih *gain ratio* terbesar dari tiap atribut. Setelah mendapatkan simpul akar dan simpul internal, ulangi perhitungan dengan cara yang sama sehingga didapatkan kelas pada tiap atribut , sehingga ketika semua atribut sudah memiliki kelas maka langkah selanjutnya adalah membuat pohon keputusan. (Muslim, et al., 2017).

2.2 Bagging

Bagging menurut (Alpaydin, 2010) merupakan sebuah metode *voting* dimana *base-learners* dibuat berbeda dengan *training* mereka melalui *training set* yang sedikit berbeda. Menghasilkan sampel L yang sedikit berbeda dari sampel yang diberikan dilakukan dengan *bootstrap,* ketika diberikan *training set* X ukuran N, maka digambarkan N contoh secara acak dari X dengan penggantian.

Bagging ditemukan oleh Breiman (1996) yang merupakan kependekan dari "*bootstrap agregating*". Han et al. (2012, p. 379) dalam bukunya menyatakan bahwa *bagging* adalah salah satu teknik dari *ensemble method* dengan cara memanipulasi *data training*, *data training* diduplikasi sebanyak *d* kali dengan pengembalian *(sampling with replacement),* yang akan menghasilkan sebanyak *d data training* yang baru, kemudian dari *d data training* tersebut akan dibangun *classifier-classifier* yang disebut sebagai *bagged classifier* (Han, et al., 2012, p. 379).

Metode *bagging* memiliki dua tahapan, pertama *bootstrap,* kedua *aggregating*, tahapan bootstrap pada metode *bagging* dilakukan dengan cara mengambil sampel dari data lstih yang dimiliki, hal itu disebut *resampling*.Tahapan kedua pada metode *bagging* adalah *aggregating.* Proses *aggregating* yaitu menggabungkan banyak nilai prediksi menjadi satu nilai prediksi. Salah satu cara mendapatkan nilai prediksi / dugaan adalah dengan cara suara terbanyak (*majority vote*) khusunya pada kasus klasifikasi, sedangkan untuk kasus regresi menggunakan rata-rata. Menurut Sutton *et al.* (Sutton, 2005, p. 126) tahapan *bagging* dapat diperhatikan sebagai sebagai berikut.

1. Tahapan *bootstrap*
   a. Ambil sampel secara acak sebanyak *n* dari data latih.
   b. Setelah sampel terambil, susun pohon terbaik berdasarkan data latih tersebut.
   c. Ulangi langkah a-b sebanyak B kali sehingga diperoleh B buah pohon klasifikasi.
2. Tahapan *Aggregating*

Tahapan *aggregating* identik dengan *majority vote* yaitu Melakukan prediksi / dugaan dari gabungan B buah pohon klasifikasi.

2.3 *Confusion Matrix*

Pengukuran kinerja klasifikasi umumnya menggunakan *confusion matrix* (Prasetyo, 2014, p.

47). Menurut Gorunescu (Gorunescu, 2011) *Confusion matrix* adalah sebuah *tool* yang dengan bantuannya didapatkan objek yang benar atau salah dengan cara mengevaluasi model klasifikasi. Sehingga untuk menentukan hasil dari *confusion matrix* yaitu dengan membandingkannya dengan kelas asli dari inputan.

Confusion matrix merupakan sebuah pencatat dalam hasil kerja klasifikasi yang dijelaskan dalam bentuk tabel. Contoh *confusion matrix* yang melakukan klasifikasi masalah biner (dua kelas) dapat dilihat pada Tabel 1, biner yang dimaksud adalah hanya memiliki 2 kelas, yaitu kelas 0 dan 1. Bisa kita ambil contoh sel *f11* adalah data pada kelas hasil prediksi sesuai data dalam kelas asli, dan f10 adalah data pada kelas hasil prediksi tidak sesuai pada kelas asli dengan kata lain bahwa kelas 1 yang dipetakan secara salah ke kelas 0.

Tabel 1. Matriks konfusi untuk 2 kelas (Prasetyo, 2014: 257)

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Berdasarkan Tabel 1 untuk dapat menghitung tingkat akurasi didapatkan melalui penjumlahan data masing-masing kelas yang diprediksi secara benar yaitu ($f_{11} + f_{00}$) dibagi dengan jumlah keseluruhan data. Sedangkan untuk menghitung laju *error* didapatkan melalui penjumlahan data masing-masing kelas yang diprediksi secara salah yaitu ($f_{10} + f_{01}$) dibagi dengan jumlah keseluruhan data (Prasetyo, 2014: 257). Perhitungan hasil akurasi dapat dilihat pada persamaan 1. Sedangkan untuk menghitung laju *error* (kesalahan prediksi) digunakan persamaan 2.

$$Akurasi = \frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{01}+f_{00}} \quad (1)$$

$$Laju\ error = \frac{f_{10}+f_{01}}{f_{11}+f_{00}+f_{01}+f_{00}} \quad (2)$$

Kinerja dari sebuah algoritma klasifikasi ditentukan dari pengujian model yang dibentuk dengan data uji (Prasetyo, 2014: 258).

2.4 *Heart Disease Dataset*

Dataset penyakit jantung diambil dari UCI Repository of Machine Learning (http://archive.ics.uci.edu/ml/datasets/heart+Disease) .Terdiri dari 303 instance dengan 75 atribut dan 1 kelas yang akan digunakan untuk memprediksi apakah seseorang memiliki penyakit jantung atau tidak, hanya 13 atribut dan 1 kelas yang akan digunakan dalam *dataset* ini. Berikut ini dijelaskan tentang atribut dan deskripsinya pada Tabel 2. Sedangkan untuk contoh *dataset* yang digunakan pada penelitian ini ditunjukkan pada Tabel 3. Contoh dataset yang terdapat *missing value handling* ditunjukkan pada Tabel 5.

Tabel 2. *Dataset* Penyakit Jantung

Atribut	Deskripsi
Age	Umur (tahun)
Sex	Jenis kelamin (1 = pria; 0 = wanita)
Cp	Jenis nyeri dada – Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic
Trestbps	tekanan darah beristirahat (dalam mm Hg saat masuk ke rumah sakit)
Chol	kolesterol serum dalam mg / dl
Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	hasil elektrokardiografi istirahat Nilai 0 : normal Nilai 1 : memiliki kelainan gelombang ST-T (inversi gelombang T dan / atau elevasi ST atau depresi> 0,05 mV) Nilai 2 : menunjukkan kemungkinan hipertrofi ventrikel kiri dengan kriteria Estes
Thalach	denyut jantung maksimal tercapai
exang	olah raga angina (1 = ada; 0 = tidak ada)
oldpeak	depresi ST diinduksi oleh latihan relatif untuk beristirahat
slope	the slope of the peak exercise ST segment – Value 1: upsloping – Value 2: flat – Value 3: downsloping
Ca	jumlah pembuluh darah utama (0-3) yang diwarnai oleh fluoroskopi
thal	3 = normal; 6 = cacat tetap; 7 = cacat reversible
num	diagnosis penyakit jantung (status penyakit angiografi) Nilai 0: <50% diameter menyempit Nilai 1:> 50% penyempitan diameter

Tabel 3. *Dataset* Penyakit Jantung

	age	...	Restecg	..	Slope	ca	thal	num
1	63	...	left_vent	...	Down	0	fixed_defect	<50
2	67	...	left_vent	...	flat	3	normal	>50_1
3	67	...	left_vent	...	flat	2	reversable_d	>50_1
4	37	...	normal	...	down	0	normal	<50
5	41	...	left_vent_	...	up	0	normal	<50
6	56	...	normal	...	up	0	normal	<50
7	62	...	left_vent_	...	down	2	normal	>50_1
8	57	...	normal	...	up	0	normal	<50
9	63	...	left_vent_	...	flat	1	reversable_d	>50_1
10	53	...	left_vent_	...	down	0	reversable_d	>50_1
	…	...	…	...	…	…	…	…
294	63	...	normal	...	flat	0	normal	>50_1
295	41	...	normal	...	up	0	normal	<50
296	59	...	left_vent_	...	flat	2	fixed_defect	>50_1
297	57	...	Normal	...	flat	0	reversable_d	>50_1
298	45	...	Normal	...	flat	0	reversable_d	>50_1
299	68	...	Normal	...	flat	2	reversable_d	>50_1
300	57	...	Normal	...	flat	1	reversable_d	>50_1
301	57	...	left_vent_	...	flat	1	normal	>50_1
302	38	...	normal	...	up	?	normal	<50
303	38	...	normal		up	?	normal	<50

Setelah dilakukan penanganan *missing value* dengan mengisi data kosong menggunakan metode modus yaitu dengan cara mengisi atribut dengan nilai yang

sering keluar. Pengisian missing value ditunjukkan pada Tabel 5.

Tabel 4. *Dataset* Penyakit Jantung dengan *missing value*

	age	. . . .	Restecg	. .	Slope	ca	thal	num
	…	..	…	...	…	…	…	…
30	57	..	Normal	...	flat	1	reversable	>50_1
30	57	..	left_vent_hy	...	flat	1	normal	>50_1
30	38	..	normal	...	up	?	normal	<50
30	38	..	normal		up	?	normal	<50

Tabel 5. *missing value handling* dengan modus

	age	. . . .	Restecg	. .	Slope	ca	thal	num
	…	..	…	...	…	…	…	…
30	57	..	Normal	...	flat	1	reversable	>50_1
30	57	..	left_vent_hy	...	flat	1	normal	>50_1
30	38	..	normal	...	up	**1**	normal	<50
30	38	..	normal		up	**1**	normal	<50

3. HASIL DAN PEMBAHASAN

Proses pengujian dalam penelitian ini menggunakan bahasa pemrograman python dengan memanfaatkan perpustakaan seperti *sklearn*, *numpy*, panda. untuk antarmuka pengguna, ia menggunakan html, css (*bootstrap*), dan *javascript*.

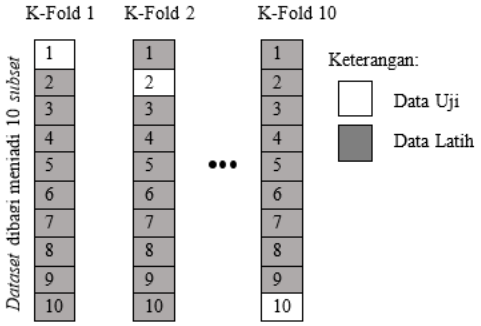
Dalam penelitian ini menerapkan teknik *bagging* dalam algoritma C4.5. *Dataset* penyakit jantung diambil dari sebuah *repository* UCI Machine Learning yang memilik jumlah atribut 13 dan 1 atribut kelas prediksi. Ada 6 atribut dengan tipe numerik dan 8 atribut tipe nominal. Atribut-atribut ini termasuk, usia, *sex*, *cp, chol, fbs, exang, slope, ca, oldpeak,* dan *num*. Data terdiri dari 303 contoh dengan sejumlah nilai hilang 7 contoh. Data diperoleh dalam format .arff. untuk menyederhanakan proses klasifikasi data yang dikonversi ke format .csv. Tahap awal adalah pra-pemrosesan yang merupakan tahap pemrosesan data. *Dataset* diubah dari tipe nominal ke tipe numerik. Transformasi data nominal ke tipe numerik dijelaskan pada Tabel 6.

Tabel 6. *Dataset* Penyakit Jantung setelah Transformasi

	age	. . . .	restecg	. . .	slope	ca	thal	num
1	63	. . .	2	. . .	3	0	6	0
2	67	. . . .	2	. . .	2	3	3	1
3	67	. . . .	2	. . .	2	2	7	1
4	37	. . . .	0	. . .	3	0	3	0
5	41	. . . .	2	. . .	1	0	3	0
6	56	. . . .	0	. . .	1	0	3	0
	…	. . . .	…	. . .	…	…	…	…
300	57	. . . .	0	. . .	2	1	7	1
301	57	. . . .	2	. . .	2	1	3	1
302	38	. . . .	0	. . .	1	?	3	0
303	38	. . . .	0	. . .	1	?	3	0

Untuk data kosong (nilai yang hilang) ditangani dengan mengisi menggunakan sebagian besar data (modus). Seperti yang ditunjukkan pada Tabel 4.

Tahap klasifikasi ini menggunakan Algoritma C4.5 dengan metode validasi, yaitu dengan *k-fold cross validation* dengan nilai k = 10 pada *dataset* penyakit jantung asli tanpa proses apa pun. Secara singkat proses 10 *k-cross validation* dapat dilihat pada Gambar 3.



Gambar 3. 10-*K Fold Cross Validation*

Pembagian *dataset* menggunakan *cross validation* dilakukan secara acak. Sehingga diperoleh data latih dan data uji. Data latih diproses dengan Algoritma C4.5 dengan menggunakan model klasifikasi. Model klasifikasi diuji menggunakan data uji. Dengan kata lain kinerja algoritma klasifikasi diukur menggunakan *confusion matrix*. Ditampilkan seperti pada Tabel 7.

Tabel 7. *Confusion Matrix* Algoritma C4.5

classified as →	A	B
a=<50	121	44
b=>50_1	38	100

Hasil klasifikasi ditampilkan dalam bentuk confusion matrix mendapatkan tingkat akurasi sebesar 72,98% diperoleh. ditunjukkan pada Tabel 8.

Tabel 8. Akurasi Algoritma C4.5

Algorithm	Accuracy Results
C4.5	72.98%

Hasil tingkat akurasi klasifikasi yang didapatkan akan dibandingkan dengan tingkat akurasi klasifikasi menggunakan algoritma C4.5 dengan teknik *bagging*. Tahap klasifikasi ini menggunakan algoritma C4.5 dan *bagging* dengan metode validasi, yaitu validasi *k*-fold cross dengan nilai k = 10, dalam *dataset* penyakit jantung asli tanpa proses apa pun. Proses dimulai dengan membagi menjadi data latih dan data uji menggunakan *cross validation*. Pengambilan data dilakukan secara acak. Sedangkan data latih diproses dengan Algoritma C4.5 dan *bagging* untuk mendapatkan model klasifikasi. Model klasifikasi diuji menggunakan data uji. Kinerja algoritma klasifikasi diukur dengan menggunakan *confusion matrix*. Seperti pada Tabel 9.

Tabel 9. *Confusion Matrix* Algoritma C4.5 + teknik *bagging*

classified as →	A	b
a=<50	146	19
b=>50_1	36	102

Hasil akurasi yang diperoleh diperoleh dari akurasi rata-rata 10 iterasi yang dihasilkan dari *k-fold cross validation*. Penerapan algoritma C4.5 dan *bagging*

dalam *dataset* penyakit jantung menghasilkan akurasi 81,84%. Hasil Akurasi algoritma C4.5 dengan penerapan teknik *bagging* dengan *k-fold cross validation* ditunjukkan pada Tabel 10.

Tabel 10. Akurasi Algoritma C4.5

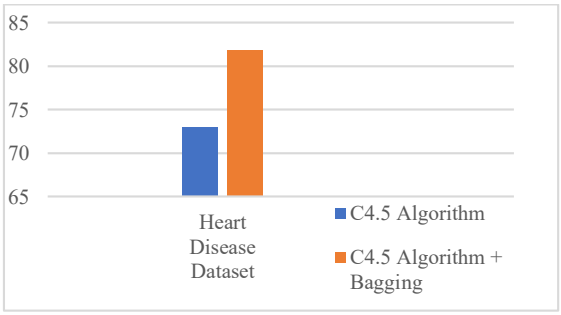
Algorithm	Accuracy Results
C4.5	81,84%

Klasifikasi menggunakan Algoritma C4.5 memperoleh tingkat akurasi sebesar 72,98% sedangkan klasifikasi dengan Algoritma C4.5 dengan menerapkan teknik *bagging* diperoleh hasil 81,84%. Dengan menerapkan teknik *bagging* dalam Algoritma C4.5 dapat meningkatkan tingkat akurasi sebesar 8,86% dari 72,98% menjadi 81,84%. Tabel Perbandingan akurasi dalam memprediksi penyakit jantung ditunjukkan pada Tabel 11.

Tabel 11. Perbandingan of Accuracy of Heart Disease Prediction

Classifier	*Pre-processing*	Accuracy
C4.5 Algorithm	-	72,98%
C4.5 Algorithm + Bagging	*Transformation, missing value handling*	81,84%

Hasil akurasi yang diperoleh dari penerapan Algoritma C4.5 dengan teknik *bagging* meningkat sebesar 8.86% dibandingkan dengan hasil akurasi menggunakan algoritma C4.5. Ditunjukkan pada Gambar 4.



Gambar 4. Grafik Perbandingan Akurasi Prediksi Penyakit Jantung

4. KESIMPULAN

Klasifikasi menggunakan Algoritma C4.5 dalam memprediksi penyakit jantung menghasilkan akurasi sebesar 72,98%. Dari hasil ini ditingkatkan dengan menerapkan teknik *bagging* pada Algoritma C4.5, menghasilkan akurasi sebesar 81,84%. Hal ini dapat disimpulkan bahwa dengan menerapkan teknik *bagging* pada Algoritma C4.5 pada prediksi penyakit jantung dapat meningkatkan hasil akurasi sebesar 8,86%.

DAFTAR PUSTAKA

ALPAYDIN, E., 2010. *Introduction of Machine Learning Second Edition.* 2 ed. Cambridge: The MIT Press.

DAS, R., 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert System With Application,* p. 1568–1572.

GORUNESCU, F., 2011. *Data Mining: Concepts, Models and Techniques.* Berlin: Springer US.

HAN, J., KAMBER, M. & JIAN, P., 2012. *Data Mining Concepts and Techniques Third Edition.* Third ed. Waltham: Morgan Kaufmann.

LIANG, G., ZHU, X. & ZHANG, C., 2011. An Empirical Study of Bagging Predictors for Imbalanced Data with Different Levels. *AI: Advances in Artificial Intelligence,* p. 213–222.

M. dkk., 2017. Penerapan Dizcretization dan Teknik Bagging Untuk Meningkatkan Akurasi Klasifikasi Berbasis Ensemble pada Algoritma C4.5 dalam Mendiagnosa Diabetes. *Jurnal Imiah Teknologi Informasi,* 8(2), p. 132.

MUZAKIR, A. & WULANDARI, R. A., 2016. Model Data Mining sebagai Prediksi Penyakit Hipertensi. *Scientific Journal of Informatics,* 3(1), pp. 19-26.

PRASETYO, E., 2014. *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: CV. Andi Offset.

R, J. & RV, S. B., 2016. C4.5 classification algorithm with back-track pruning for accurate prediction of heart disease. *Biomedical Research,* 0(0), pp. 107-111.

ROHMAN, A., SUHARTONO, V. & SUPRIYANTO, C., 2017. PENERAPAN ALGORITMA C4.5 BERBASIS ADABOOST. *Jurnal Teknologi Informasi,* 13(1), pp. 13-19.

SAIFUDIN, A. & WAHONO, R. S., 2015. Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software. *Journal of Software Engineering,* 1(1), pp. 28-37.

SIRINGORINGO, R., 2018. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor. *Jurnal ISD,* 3(1), pp. 44-49.

SUTTON, C., 2005. *Classification and Regression Trees, Bagging, and Boosting.* s.l.:Elsevier.

THOMAS R. FRIEDEN, D. M. B., 2011. The "Million Hearts" Initiative — Preventing Heart Attacks and Strokes. *The New England Journal of Medicine,* 363(1), p. 1–3.

TSAI, C.-J., LEE, C.-I. & Yang, W.-P., 2008. A discretization algorithm based on. *Information Sciences,* 4(1), pp. 714-731.