

PENINGKATAN KINERJA DECISION TREE C4.5 MELALUI PENDEKATAN MEAN DAN MEDIAN PADA DATA NUMERIK

Arjon Samuel Sitio¹⁾, Fricles A Sianturi²⁾

^{1,2}Universitas Tjut Nyak Dhien, Sumatera Utara, Indonesia
email: arjonsitio@yahoo.com

Abstract

Decision Tree C4.5 is one of the most widely used classification algorithms due to its simplicity and effectiveness in handling various types of data. However, the accuracy of the algorithm can be affected by the presence of missing or inconsistent numerical values. This study aims to improve the performance of the Decision Tree C4.5 algorithm by incorporating mean and median improvisations in handling numerical datasets. The research methodology involves preprocessing numerical datasets by replacing missing values with mean and median calculations before applying the Decision Tree C4.5 algorithm. The modified algorithm is then tested on multiple datasets to evaluate its impact on classification accuracy. The experimental results indicate that the proposed improvisation significantly enhances the accuracy and reliability of the Decision Tree C4.5 model compared to the standard implementation. The findings suggest that the integration of mean and median techniques can effectively improve decision tree-based classification in numerical datasets.

Keywords: *Decision Tree C4.5, mean, median, classification, numerical dataset, accuracy.*

1. PENDAHULUAN

Klasifikasi merupakan sebuah metode yang biasa digunakan untuk memprediksi atau memperkirakan suatu class pada sebuah dataset. Klasifikasi adalah pengelompokan data kategorik berdasarkan kriteria tertentu. Salah jenis klasifikasi yaitu, *decision tree*. Pada penelitian Wibowo [1] yang mengkomparasikan algoritma *decision tree* dengan *naïve bayes* menghasilkan akurasi *decision tree* lebih tinggi dari pada algoritma *naïve bayes*.

Decision tree memiliki model algoritma C4.5, algoritma ini sangat terkenal pada dunia penelitian sesuai dengan pernyataan Azwanti dan Elisa [2]. Kelebihan algoritma ini yaitu, dalam memprediksi menggunakan pohon keputusan sehingga lebih mudah dipahami.

Algoritma *decision tree C4.5* merupakan pengembangan dari algoritma ID3. Algoritma *Decision tree C4.5* atau yang biasa dikenal dengan pohon keputusan yaitu, jenis klasifikasi pada data mining untuk memberikan keputusan atau prediksi pada objek atau class. Dalam proses mengklasifikasi tingkatan dalam membuat pohon

keputusan dimulai dari menentukan akar, kemudian membuat cabang dengan nilai gain tertinggi sampai menemukan cabang yang classnya sama, berdasarkan pernyataan Masulloh dan Fitriyani [3]. Untuk mendapatkan akar dan cabang pada pohon keputusan menggunakan gain tertinggi atau informasi gain.

Menurut Budiman dan Paradani [4], informasi gain dapat diperoleh berdasarkan persamaan *entropy*. *Entropy* dihasilkan dari *entropy* seluruh data dan *entropy* pada masing-masing atribut yang terdapat pada dataset. Hasil dari informasi gain menentukan tempat atribut untuk *node* pada pohon keputusan (*Decision tree*). Informasi gain juga menentukan *rule* algoritma C4.5 dalam membuat keputusan pada *Decision tree*.

Proses menentukan *rule* melibatkan *preprocessing* data untuk menyeleksi sebuah atribut atau fitur dalam dataset. Dataset yang memiliki banyak atribut sehingga rentan terhadap kemiripan data atau redundansi data yang tidak diperlukan. Menurut Cahyani dan Muslim [5], teknik *Preprocessing* data diperlukan untuk dapat



meningkatkan kualitas data dan menghasilkan hasil yang lebih akurat. Pada sebuah dataset tidak hanya berisi atribut kategorikal saja, terkadang dataset juga memiliki atribut numerik. Atribut numerik memiliki domain yang sangat besar, bahkan tidak terbatas dalam data mining.

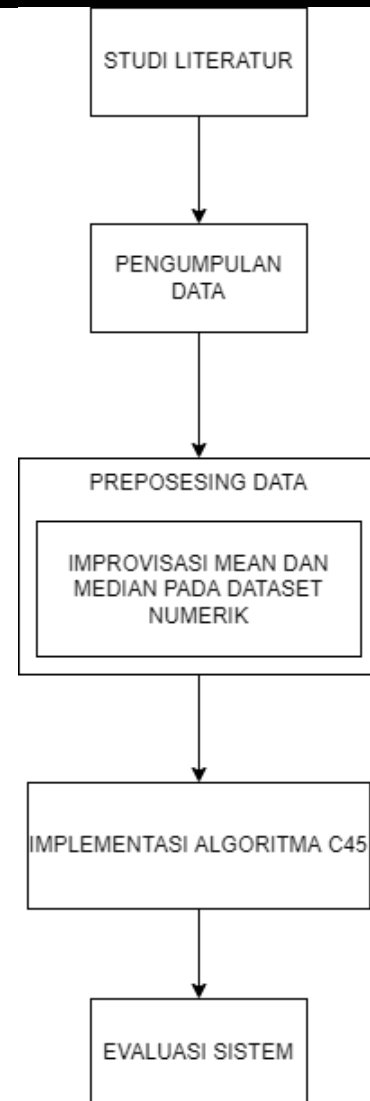
Menurut Ferchichi [6], algoritma C4.5 dalam menangani data kuantitatif seringkali kurang efisien dan efektif. Sehingga untuk meminimalisir kehilangan informasi dan kompleksitas waktu dapat mengimprovisasi dataset pada atribut numerik saat *preprocessing* data. Improvisasi dilakukan dengan menggunakan *mean* dan *median* pada *preprocessing* atribut numerik agar mendapat nilai ambang batas untuk mengimplementasikan algoritma C4.5. *Mean* digunakan untuk mencari nilai rata-rata dan *median* untuk mencari nilai tengah pada seluruh data yang sebelumnya sudah dilakukan pengurutan data dari yang terkecil ke terbesar.

Penelitian mengenai algoritma *decision tree* C4.5 sebelum sudah dilakukan untuk mendiagnosa penyakit liver dengan nilai akurasi 72,67% [7]. Penelitian tersebut menghasilkan bahwa dalam menentukan penyakit liver hanya menggunakan 2 variabel dari 11 variabel dataset.

2. METODE PENELITIAN

Rencana dalam penelitian ini dilakukan berdasarkan diagram yang ditunjukkan pada Gambar 1. Penelitian ini dilakukan mulai dari studi literatur yang digunakan sebagai referensi penulis. Pengumpulan data yaitu pencarian dataset yang digunakan sebagai penelitian. *Preprocessing* data dapat dilakukan melalui tahap data cleaning, data transformasi, data reduction. *Preprocessing* penelitian ini ditambahkan dengan improvisasi *mean* dan *median* pada dataset yang memiliki atribut numerik.

Implementasi algoritma C4.5 dilakukan setelah melakukan proses sebelumnya dan diakhiri dengan proses evaluasi sistem. Evaluasi sistem digunakan untuk mengetahui performa sebuah algoritma yang telah digunakan.



Gambar 1. Alur Penelitian

3.1. Studi Literatur

Studi literatur digunakan sebagai bahan referensi dalam melakukan sebuah penelitian. Referensi diperoleh melalui jurnal atau penelitian terdahulu, buku, dll.

3.2. Pengumpulan Data

Pengumpulan data dilakukan dengan mencari dataset yang dapat digunakan untuk klasifikasi. Dataset pada penelitian ini diambil dari www.kaggle.com yaitu dataset prediksi penyakit gagal jantung. Data yang diperoleh kemudian dilakukan *preprocessing* terlebih dahulu agar dapat dilakukan perhitungan algoritma C4.5.

Prediksi jantung menggunakan *decision tree C4.5* sudah banyak digunakan dalam sebuah penelitian. Salah satu penelitian tentang data tersebut yaitu, penelitian yang dilakukan oleh Alham [8], hasil dari penelitian tersebut yaitu, akurasi confusion matrik yang diperoleh dalam pengujian data menggunakan algoritma C4.5 adalah 94,4%.

Data kemudian dilakukan pengolahan data dengan menggunakan algoritma C4.5. Hasil dari perhitungan algoritma C4.5 selanjutnya dilakukan evaluasi sistem.

Dataset pada Tabel 1 memiliki atribut kategorik dan numerik. Dataset yang telah diambil kemudian dipartisi menjadi tiga skenario yaitu 70% data trainging dan 30% data *testing*, 75% data *training* 25% data *testing*, dan 80% data *training* dan 20% data *testing*. Data *traning* yaitu data yang digunakan untuk melatih sedangkan, data *testing* digunakan untuk mengetahui performa algoritma yang digunakan. Pada data *training* yang atributnya bertipe numerik dilakukan *preprosesing* dengan menghitung *mean* dan *median* untuk menemukan nilai ambang batas.

3. HASIL DAN PEMBAHASAN

Tabel 1. Dataset Prediksi Gagal Jantung

Age	Sex	Chest Pain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	MaxHR	Exercise Angina	Old peak	ST Slope	Heart Disease
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1,5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1,5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1,5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	ATA	120	267	0	Normal	160	N	3	Flat	1



43	F	TA	100	223	0	Normal	142	N	0	Up	0
44	M	ATA	120	184	0	Normal	142	N	1	Flat	0
49	F	ATA	124	201	0	Normal	164	N	0	Up	0

3.1. Preprocessing Data

Preprocessing data dilakukan guna untuk menyiapkan data matang agar siap diolah pada tahap selanjutnya. Pada dataset yang bertipe numerik dalam tahap preprocessing dengan menemukan nilai *mean* dan *median* untuk menemukan nilai ambang batas.

$$Mean = \frac{1}{n} \sum_{i=1}^n Xi \quad (1).$$

Perhitungan *mean* pada improvisasi dataset pernah dilakukan oleh ferchichi [6] yang digunakan untuk menghitung atribut kontinyu. Improvisasi menggunakan *mean* menghasilkan akurasi lebih baik pada pohon keputusan dan kompleksitas model. *Mean* adalah nilai rata-rata dalam sebuah atribut X, yang digunakan untuk mengukur pusat himpunan N. *Preprocessing* dalam menentukan *mean* dilakukan pada setiap atribut pada dataset dengan menggunakan rumus Persamaan (1).

Pada penelitian ini improvisasi dataset numerik tidak hanya menggunakan *mean* tetapi juga menggunakan *median*. Berdasarkan buku suyanto [9], *median* adalah nilai tengah dari sebuah atribut X untuk mengukur pusat dari himpunan N dengan ketentuan data harus diurutkan terlebih dahulu. Dalam mencari nilai *median* pada preprocessing seperti pada tahap menentukan *mean* yaitu dengan menghitung nilai *median* setiap atribut yang terdapat pada dataset yang digunakan.

Menentukan nilai *median* dalam mencari nilai ambang batas menggunakan rumus Persamaan (2) untuk data yang jumlahnya genap.

$$Median = \left(X_n + \left(X_n + 1 \right) \right) / 2 \quad (2).$$

Sedangkan untuk data yang jumlahnya ganjil dalam menentukan nilai *median* menggunakan rumus dengan Persamaan (3).

$$Median = \frac{X_{n+1}}{2} \quad (3)$$

Tabel 2. Nilai ambang batas *mean* dan *median*.

Atribut	Mean	Median
Age	53,20	54
Resting BP	132,45	130
Cholestrol	186,42	216
Max HR	133,81	133
Oldpeak	0,82	0,4

3.2. Algoritma C4.5

Pada algoritma *Decision tree* terdapat beberapa model algoritma, salah satu dari model pohon keputusan yaitu *Algoritma C4.5*. Algoritma C4.5 dalam membuat pohon keputusan menggunakan parameter yang telah ditentukan untuk membuat sebuah keputusan. *Root, Node, dan Relationship* merupakan elemen untuk membuat pohon keputusan. Elemen pada keputusan ditentukan berdasarkan nilai gain tertinggi, untuk menghitung nilai gain memerlukan nilai *entropy* berdasarkan penelitian Yulianti [10].

Tahapan dalam perhitungan algoritma C4.5 menurut Sinaga [11] dengan menentukan nilai gain informasi. Gain informasi diperoleh berdasarkan nilai *entropy*. Menentukan nilai *entropy* menggunakan Persamaan (4). Pencarian nilai *entropy* dengan menghitung *entropy* pada keseluruhan atribut dan *entropy* setiap atribut.

$$Entropy = \sum_{i=1}^n - Pi * \log_2 Pi \quad (4).$$

Keterangan:

N = Jumlah Skenario S

Pi = Proporsi dari Si terhadap S



Menurut Santosa dan Umam [12], algoritma C4.5 merupakan modifikasi atau pengembangan dari model algoritma ID3. Pada dasarnya algoritma C4.5 digunakan untuk mengatasi masalah missing value dan atribut numerik. Dalam membuat aturan atau *rule* pada pohon keputusan sesuai dengan data *training*. Algoritma C4.5 merepresentasikan atau memodelkan sebuah data seperti pohon.

Dalam membuat sebuah pohon keputusan membutuhkan root node, internal node dan leaf node agar menyerupai pohon sesungguhnya [13]. Menurut Patel dan Prajapati [14], *root node* atau akar merupakan induk dari semua *node* dan merupakan node tertinggi pohon pada algoritma *decision tree*. Dalam menentukan *root* atau akar dari pohon keputusan ditentukan berdasarkan *gain* tertinggi atau *gain* informasi.

Gain informasi ditentukan menggunakan *entropy* pada seluruh atribut dan *entropy* setiap atribut sebagai pencarian nilai gain. Rumus

pencarian nilai gain informasi ditunjukkan pada Persamaan (5).

$$Gain = Entropy(s) - \sum_{i=1}^n \frac{|s_i|}{|s|} * Entropy(A) \quad (5).$$

Keterangan:

$Entropy(S)$ = *Entropy* pada seluruh data/kasus.

$Entropy(A)$ = *Entropy* pada setiap atribut.

N = Jumlah parisi atribut A.

|Si| = Jumlah kasus pada Skenario ke-i.

|S| = Jumlah kasus dalam S.

Tabel 3. Data prediksi gagal jantung

Jumlah	Hipotesa		Entropy Total
	0	1	
734	328	402	0,991838651

Tabel 3 merupakan perhitungan untuk memperoleh nilai *entropy* pada seluruh data *training*.

Tabel 4. Perhitungan gain

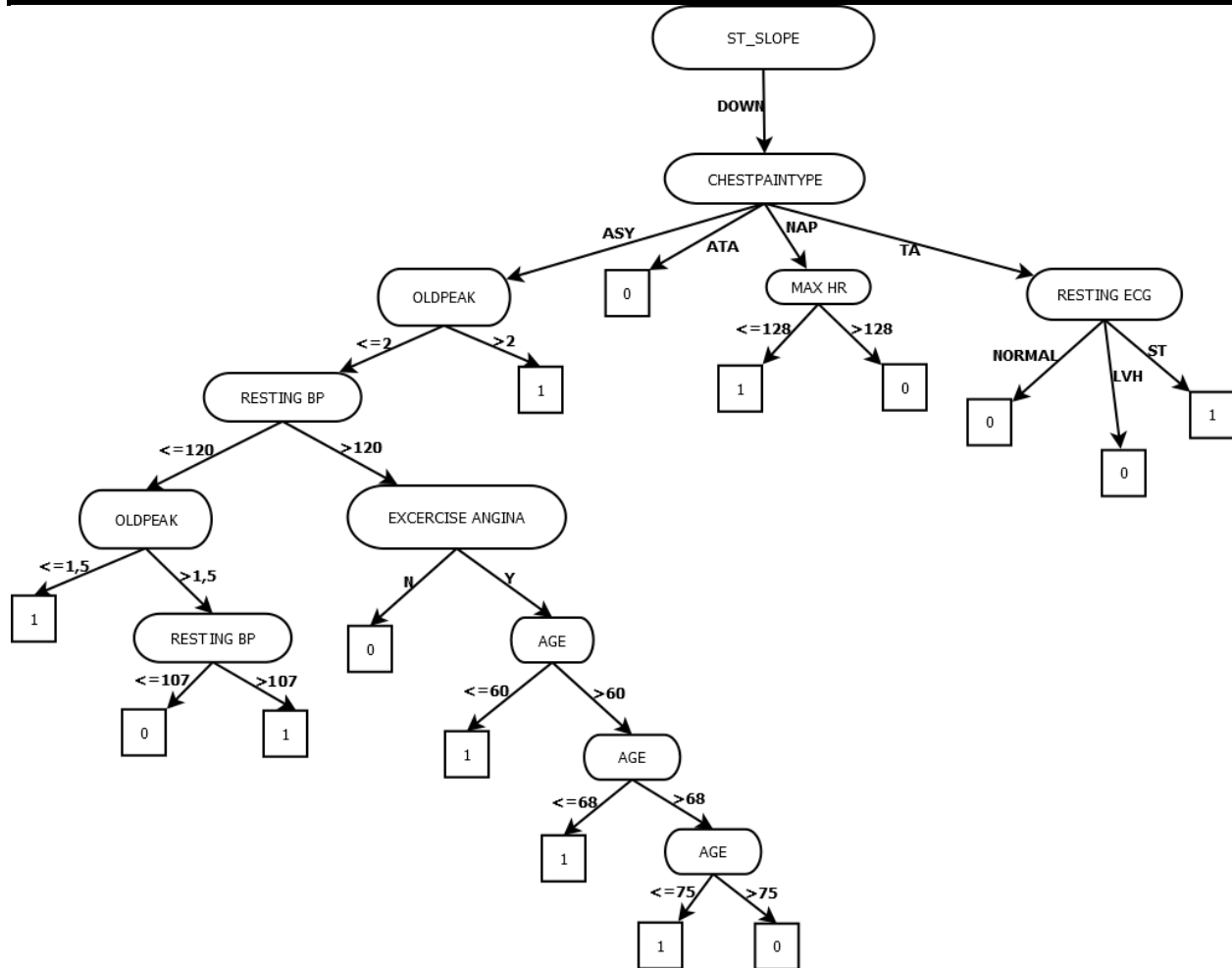
Atribut			S	Si	Si	Entropy	Gain
Total			734	328	406	0,991839	
Age	<=	53,20	349	200	149	0,984541	0,155464
	>		385	128	14	0,702064	
	<=	54,00	389	223	166	0,984456	0,145379
	>		345	105	12	0,690863	
Sex		M	593	217	376	0,9475	0,082905
		F	141	111	30	0,746737	
ChestPainType		ASY	399	87	312	0,756582	0,221307
		ATA	147	128	19	0,555382	
		TA	33	18	15	0,99403	
		NAP	155	95	60	0,9629	
RestingBP	<=	132,45	402	206	196	0,999554	0,105035
	>		332	122	17	0,750282	
	<=	130,00	387	196	191	0,99988	0,11311
	>		347	132	17	0,743611	



Cholesterol	<=	186,42	251	69	182	0,848409	0,213748
	>		483	259	32	0,741548	
	<=	216,00	368	142	226	0,962082	0,130151
	>		366	186	25	0,760745	
FastingBS		0	549	296	253	0,99557	0,079722
		1	185	32	153	0,664461	
RestingECG		Normal	462	220	242	0,998364	0,008985
		LVH	97	47	50	0,99931	
		ST	175	61	114	0,932786	
MaxHR	<=	133,81	368	93	275	0,815557	0,288501
	>		366	235	14	0,590505	
	<=	133,00	368	93	275	0,815557	0,288501
	>		366	235	14	0,590505	
ExerciseAngina		N	425	278	147	0,930337	0,184306
		Y	309	50	259	0,638628	
Oldpeak	<=	0,82	409	263	146	0,940138	0,12743
	>		325	65	28	0,769107	
	<=	0,40	370	246	124	0,920087	0,146655
	>		364	82	28	0,769045	
ST_Slope		UP	318	266	52	0,642669	0,37035
		DOWN	49	10	39	0,730017	
		FLAT	367	52	315	0,588646	

Tabel 4 merupakan hasil perhitungan gain pada setiap atribut yang ada pada dataset. Hasil *gain* informasi yang memiliki nilai tertinggi pada node pertama digunakan sebagai *root* pada pohon keputusan. Pohon keputusan pada prediksi

gagal jantung menghasilkan *rule* pada Gambar 2. *Rule* yang didapat berdasarkan pohon keputusan kemudian diimplementasikan pada evaluasi sistem.



Gambar 2. Pohon keputusan

3.4. Evaluasi Sistem

Evaluasi sistem dilakukan dengan menggunakan *confusion matrix*. *Confusion matrix* sebagai tolak ukur untuk menguji pada metode klasifikasi dengan menggunakan data *testing* yang telah ditentukan. Pada penelitian ini menggunakan 20% data dari dataset.

Tabel 2. Hasil *Testing*

Aktual	Prediksi
0	1
0	0
0	1
0	0
.....
1	0
1	1
1	1

Evaluasi sistem dilakukan untuk memperoleh nilai akurasi, *error rate*, *precision*, dan *recall* dengan menggunakan data *testing* pada Tabel 2.

Tabel 3. *Confusion matrix*

		N	
		1	0
AKTUAL	1	TP	FN
	0	FP	TN

Tabel 3 digunakan untuk merepresentasikan hasil *testing* pada *confusion matrix*.

Menurut Ginting, Kursini, dan Taufiq [15], metode untuk menentukan akurasi dalam sebuah data mining menggunakan *confusion matrix*. Akurasi dari klasifikasi data mining adalah tingkat persentase ketepatan antara prediksi dan kenyataan dalam klasifikasi sebuah data. *Error rate* adalah persentase tingkat

kesalah dalam memprediksi dengan keadaan nyatanya. *Precision* mengilustrasikan tingkat akurasi antara data prediksi dengan data yang diminta. Keberhasilan dalam menggambarkan model agar sebuah informasi dapat ditemukan kembali disebut *recall* atau *sensitivity*.

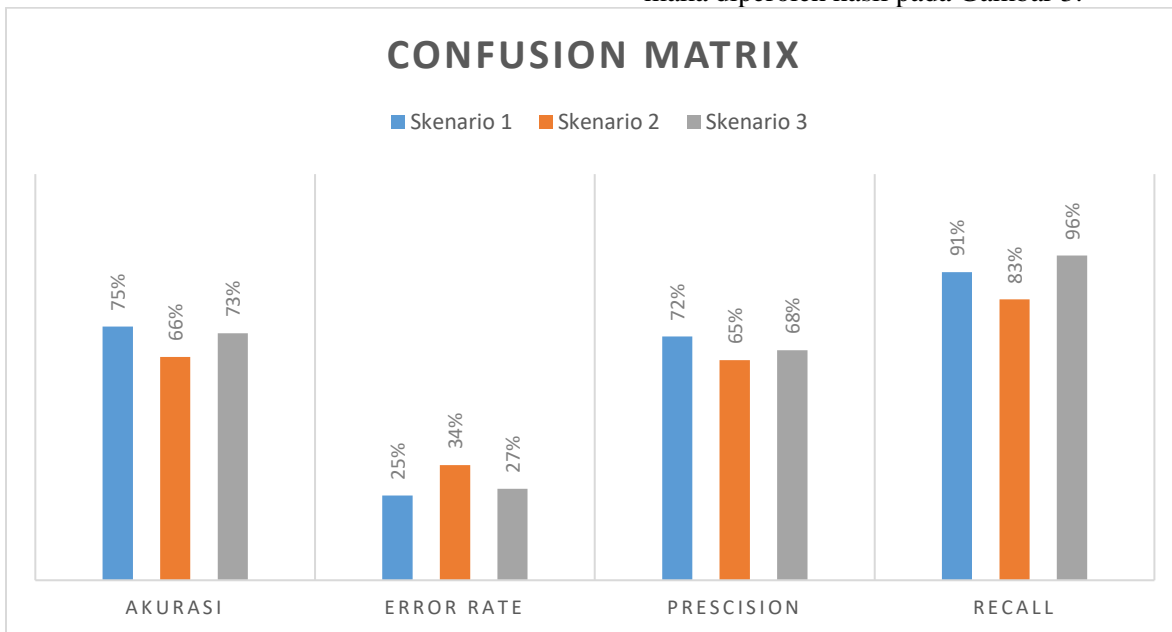
$$\text{akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (6).$$

$$\text{error rate} = \frac{FP+FN}{TP+TN+FP+FN} \times 100\% \quad (7).$$

$$\text{precision} = \frac{TP}{TP+FP} \times 100\% \quad (8).$$

$$\text{recall} = \frac{TP}{TP+FN} \times 100\% \quad (9)$$

Dari Persamaan (6), (7), (8), dan (9) maka diperoleh hasil pada Gambar 3.



Gambar 3. Diagram confusion matrix

Pada Gambar 3 menunjukkan hasil dari evaluasi sistem menggunakan *confusion matrix* dengan dataset dibagi menjadi 3 Skenario yaitu, Skenario 1 menggunakan 70% data *training* dan 30% data *testing* mendapatkan akurasi 75%, Skenario 2 dengan 75% data *training* 25% data *testing* menghasilkan akurasi 66%, Skenario 3 menggunakan 80% data *training* 20% data *testing* menghasilkan akurasi 73%. Perbandingan pada 3 Skenario tersebut menunjukkan bahwa akurasi tertinggi terdapat pada Skenario 1. Sehingga improvisasi *mean* dan *median* pada algoritma *decision tree* C4.5 dapat menggunakan Skenario tersebut.

4. KESIMPULAN

Hasil dari penelitian menggunakan dataset tentang prediksi gagal jantung dengan memSkenario dataset menjadi 3 bagian menunjukkan akurasi tertinggi pada Skenario 1. Dataset pada Skenario tersebut terbagi menjadi 70% data *training* dan 30% data *testing* yang menghasilkan akurasi sebesar 75%. Sehingga penerapan *mean* dan *median* untuk mencari nilai ambang batas pada *preprocessing* algoritma *decision tree* C4.5 dapat menggunakan Skenario tersebut untuk data *training* dan *testing* pada dataset baru.

Preprocessing pada atribut yang bertipe numerik menggunakan *mean* dan *median* mempermudah dalam mencari nilai ambang batas atau pusat pada sebuah atribut. *Mean* dan *median* dalam pencarian nilai ambang batas dapat



meminimalisir kehilangan informasi dan kompleksitas waktu pada atribut numerik.

5. REFERENSI

- [1] I. C. Wibowo, A. C. Fauzan, M. Dwi, P. Yustiana, and F. A. Qhabib, "Komparasi Algoritma Naive Bayes dan *Decision tree* Untuk Memprediksi Lama Studi Mahasiswa," vol. 1, no. 2, pp. 65–74, 2019.
- [2] N. Azwanti, E. Elisa, U. P. Batam, and J. R. S. Kuning, "InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan Analisis Pola Penyakit Hipertensi Menggunakan Algoritma C4 . 5," vol. 2, 2019.
- [3] I. Massulloh and Fitriyani, "Implementasi Algoritma C4.5 Untuk Klasifikasi Anak Berkebutuhan Khusus Di Ibnu Sina Stimulasi Center," *eProsiding Sist. Inf.*, vol. 1, no. 1, pp. 136–144, 2020.
- [4] A. S. Budiman and X. A. Parandani, "Uji Akurasi Klasifikasi Dan Validasi Data Pada Penggunaan Metode Membership Function Dan Algoritma C4.5 Dalam Penilaian Penerima Beasiswa," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 565–578, 2018, doi: 10.24176/simet.v9i1.2021.
- [5] N. Cahyani and M. A. Muslim, "Increasing Accuracy of C4 . 5 Algorithm by Applying Discretization and Correlation-based Feature Selection for Chronic Kidney Disease Diagnosis," vol. 12, no. 1, pp. 25–32, 2020.
- [6] A. Ferchichi, K. Nouira, and A. Cherfi, "MC4.5 *decision tree* algorithm: an improved use of continuous attributes," *Int. J. Comput. Intell. Stud.*, vol. 9, no. 1/2, p. 4, 2020, doi: 10.1504/ijcistudies.2020.10028137.
- [7] I. Setiawati, A. P. Wibowo, and A. Hermawan, "IMPLEMENTASI *DECISION TREE* UNTUK MENDIAGNOSIS PENYAKIT LIVER," *J. Inf. Syst. Manag.*, vol. 1, no. 1, pp. 13–17, 2019.
- [8] S. R. J. I. Alham, "Sistem Diagnosis Penyakit Jantung Koroner Dengan Menggunakan Algoritma C4.5 Berbasis Website (Studi Kasus: RSUD Dr. Soedarso Pontianak)," *Petir*, vol. 14, no. 2, pp. 214–222, 2021, doi: 10.33322/petir.v14i2.1338.
- [9] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, Revisi. Bandung: Informatika Bandung, 2019.
- [10] I. Yulianti, R. A. Saputra, M. S. Mardiyanto, and A. Rahmawati, "Optimasi Akurasi Algoritma C4.5 Berbasis Particle Swarm Optimization dengan Teknik Bagging pada Prediksi Penyakit Ginjal Kronis," *Techno.Com*, vol. 19, no. 4, pp. 411–421, 2020, doi: 10.33633/tc.v19i4.3579.
- [11] N. A. Sinaga, A. T. Purba, K. Akuntansi, P. B. Indonesia, T. Komputer, and P. B. Indonesia, "Penerapan algoritma c.45 untuk memprediksi tingkat kepuasan mahasiswa terhadap politeknik bisnis indonesia," vol. 4, pp. 245–254, 2021.
- [12] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics*, 2nd ed. Yogyakarta: Penebar Media Pustaka, 2018.
- [13] H. D. Fahma and A. C. Fauzan, "Prediksi Keberlangsungan Studi Mahasiswa Fakultas Ilmu Pendidikan dan Sosial Universitas Nahdlatul Ulama Blitar," vol. 1, no. 2, pp. 110–119, 2021.
- [14] H. H. Patel and P. Prajapati, "Study and Analysis of *Decision tree* Based Classification Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 74–78, 2018, doi: 10.26438/ijcse/v6i10.7478.
- [15] V. S. Ginting, Kusriani, and E. Taufiq, "Implementasi algoritma c4.5 untuk memprediksi keterlambatan pembayaran sumbangan pembangunan pendidikan sekolah menggunakan python," vol. 10, pp. 36–44, 2020.



