Penerapan Algoritma Decision Tree C5.0 Untuk Memprediksi Tingkat Kematian Pasien Penyakit Gagal Jantung

Application of The C5.0 Decission Tree Algorithm to Predict Patient Mortality Rate Heart Failure Disease

Suraji[1*], Abd. Charis Fauzan[2], Harliana [3]
[1,2,3] Program Studi Ilmu Komputer, Fakultas Ilmu Eksakta, Universitas Nahdlatul Ulama Blitar
*e-mail: [*1]surajicomp@gmail.com, [2]abdcharis@unublitar.ac.id, [3]harliana@unublitar.ac.id*

Abstrak

Penyakit jantung salah satu penyakit paling umum di antara penyakit lainnya, penyakit jantung dapat menyerang semua orang tanpa mengenal jenis kelamin, usia, atau faktor lainnya. Penelitian ini bertujuan untuk menerapkan algoritma C5.0 untuk memprediksi keberlangsungan hidup penderita penyakit gagal jantung serta memprediksi kematian pasien gagal jantung. Sedangkan untuk mengetahui akurasi yang dihasilkan oleh algoritma C5.0 dalam memprediksi akan digunakan confusion matrix. Penelitian ini akan menggunakan 300 dataset heart failure clinical records. Berdasarkan hasil yang didapatkan bahwa akurasi yang dihasilkan oleh algoritma decision tree C5.0 adalah 86,6% dengan perbandingan jumlah data latih dan data testing adalah 80 : 20. Sedangkan untuk presisi yang dihasilkan adalah 89,655 dan sensitifitas yang dihasilkan adalah 25 dengan spesifisitas 96,15

Kata kunci*—Algoritma Decision Tree, C.5.0, prediksi, gagal jantung*

Abstract

Heart disease is one of the most common diseases among other diseases, heart disease can affect everyone regardless of gender, age or other factors. This study aims to apply the C5.0 algorithm to predict survival in heart failure patients and predict death in heart failure patients. Meanwhile, to find out the accuracy produced by the C5.0 algorithm in predicting the confusion matrix will be used. This research will use 300 datasets of heart failure clinical records. Based on the results obtained, the accuracy produced by the decision tree algorithm C5.0 is 86.6% with a ratio of the amount of training data and testing data is 80 : 20. Meanwhile, the resulting precision is 89.655 and the resulting sensitivity is 25 with a specificity of 96, 15

Keywords—*Decision Tree Algorithm, C5.0, prediction, heart failure*

PENDAHULUAN

Penyakit jantung merupakan salah satu penyakit pembunuh nomor satu di Dunia, hal ini terbukti dengan data dari WHO dan CDC yang menyatakan bahwa hampir 11,3% atau setara dengan 26,6 juta jiwa orang dewasa di berbagai negara berkembang terdiagnosa penyakit jantung[1]. Di Indonesia sendiri jumlah penderita penyakit jantung pada tahun 2020 naik sekitar 24% bila dibandingkan tahun 2005 atau setara dengan 82 juta jiwa penderita [2]. Apabila dilihat dari definisinya penyakit jantung merupakan semua penyakit yang berhubungan dengan gangguan kinerja jantung yang mengacu pada fungsi hati[1]. Namun definisi ini berbeda jika disandingkan dengan penyakit kardiovaskular, dimana penyakit kardiovaskular adalah penyakit yang berhubungan dengan 2 fungsi utama organ tubuh yaitu pembuluh darah dan jantung[1].

Diagnosis dini yang tepat dan inovasi dalam dunia medis untuk membantu menurunkan tingkat kematian pasien gagal jantung mungkin perlu dilakukan. Maka berdasarkan hal tersebut penelitian ini akan melakukan prediksi terhadap tingkat kematian pasien gagal jantung melalui

Informasi Artikel:
Submitted: Mei 2022, **Accepted:** November 2022, **Published:** November 2022
ISSN: 2685-4902 (media online), Website: http://jurnal.umus.ac.id/index.php/intech

pendekatan algoritma C5.0. Adapun tujuan dari penelitian ini adalah mengetahui nilai akurasi yang dihasilkan oleh Algoritma C5.0 dalam memprediksi tingkat kematian pasien yang mengalami gagal jantung. Dalam hal ini algoritma C5.0 dipilih karena nilai akurasinya yang cukup tinggi yaitu sekitar 93,75% bila dibandingkan dengan Naïve Bayes dalam memprediksi banjir di Indonesia[3], 86,67% bila dibandingkan dengan KNN dalam memprediksi kredit macet pada koperasi [4], dan 96,85% bila dibandingkan dengan regresi linier dalam memprediksi kelulusan mahasiswa [5]. Selain itu penulis memilih algoritma C5.0 karena menghasilkan informasi pada rule berdasarkan node yang dihasilkan, dimana pemilihan node ini akan didasarkan pada nilai gain dan entropynya[6]. Untuk mengetahui akurasi tersebut akan dilakukan pengujian secara confusion matrix, dimana pengujian ini menghasilkan nilai yg akurat dalam menghitung diagnose penyakit koi [7], memprediksi penyakit jantung [8], dan perhitungan akurasi pada breast cancer [9].

METODE PENELITIAN

Tahapan metode penelitian yang dilakukan dalam penelitian ini terangkum pada Gambar 1. Langkah awal peneliti akan melakukan studi literatur terhadap kelebihan dan kekurangan berbagai algoritma klasifikasi dalam memprediksi sesuatu. Menurut berbagai literatur yang penulis baca maka penulis memutuskan untuk menggunakan algoritma C5.0 dalam melakukan prediksi tingkat kematian pasien gagal jantung. Tahapan selanjutnya yaitu pengumpulan dataset, dataset yang digunakan dalam penelitian ini peneliti ambil dari UCI yaitu mengenai *heart failure clinical records* yang terdapat pada link https://archive.ics.uci.edu/ml/machine-learning-databases/00429/ Adapun atribut yang akan digunakan dalam penelitian ini terangkum pada Table 1.

Tabel 1. Dataset heart failure clinical records

No.	Fitur atau Atribut	Deskripsi
1	*Usia*	Usia
2	*Anemia*	Jumlah sel darah merah
3	Creatinine phosphokinase	Jumlah keratin fosfokinase
4	*Diabetes*	Jumlah kadar gula
5	*Ejectionfraction*	Kemampuan jantung
6	*High blood pressure*	Tekanan darah
7	*Trombosit*	Jumlah keeping darah
8	*Serum creatinine*	Zat limbah dalam darah
9	*Serum sodium*	Jumlah kadar serum sodium
10	*Gender*	Jenis kelamin
11	*Smoking*	Merokok atau tidak
12	*Time*	Waktu yang dialami penderita
13	*Death*	Kematian

Selanjutnya akan dilakukan tahapan preprocessing yang meiliputi pengolahan data sekunder dan transformasi data. Pada transformasi data akan ditetapkan rumus guna mengetahui interval untuk mengetahui pengumpulan data dalam berbagai kondisi. Selanjutnya akan dilakukan perhitungan secara algoritma C5.0 dalam menentukan entropi dan gain untuk menghasilkan pohon keputusan yang dibangun. Rumus untuk mencari entropi tersebut terdapat pada persamaan (1)[10]

$$Entropy\ (S) = -P_{(+)} \log_2 P_{(+)} - P_{(-)} \text{-} \log_{2P(-)} \quad \text{persamaan (1)}$$

Dimana:

S : sampel yang digunakan untuk pembelajaran (data)

P (-) : Kuantitas solusi positif atau negatif untuk kriteria yang relevan, menggunakan data sampel.
P (+) : Jumlah solusi atau data negatif yang tidak disertakan dalam satu file untuk kriteria yang relevan.
Entropi (S) = 0 jika setiap instance dari S terletak di cluster yang sama.
Entropi (S) = 1 jika jumlah S positif dan negatif sama.
0> entropi (S)> 1 jika jumlah sampel positif dan negatif S tidak sama.

Selanjutnya untuk mencari nilai gain digunakan persamaan (2)[11]:

$$\text{Gain (S,A)} = \textit{Entropy}\ (S) - \sum_{i=1}^{n} \frac{|si|}{|s|} * \textit{Entropy} \qquad \text{persamaan (2)}$$

Dimana:

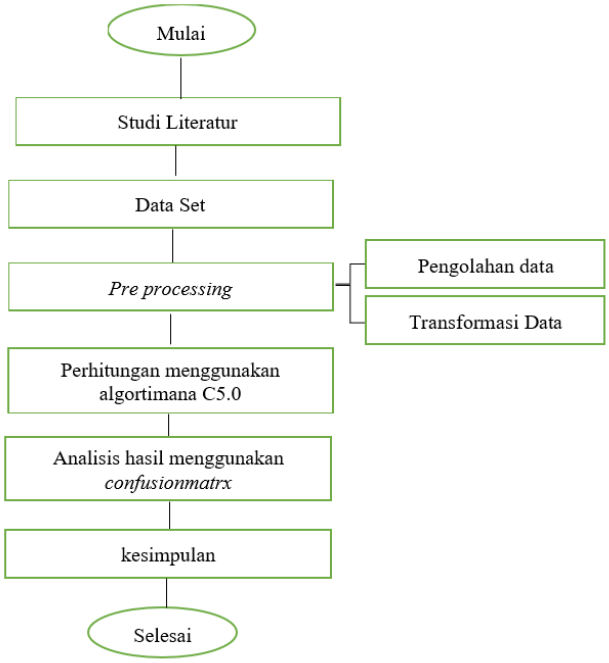
A : Atribut
S : Sampel
n : Jumlah partisis himpunan atribut A
|Si| : Jumlah sampel pada pertisi ke –i
|S| : Jumlah sampel dalam S

Karena Si adalah kumpulan kasus dalam kelas ke-i, dan A adalah komponen yang dipakai, sedangkan n adalah banyaknya kelas pada komponen A maka | Si | adalah banyaknya kasus dalam kategori ke-i, dan | S | adalah banyaknya kasus di S.
Rumus dasar untuk menghitung rasio keuntungan adalah Persamaan (3).

$$\text{Gain Rasio} = \frac{Gain\ (S,A)}{\sum_{i=1}^{m} \ Entropy\ (Si)} \qquad \text{persamaan (3)}$$

Di sini, Gain (S, A) adalah nilai gain dari variable, $\sum_{i=1}^{m} \ Entropy\ (Si)$ Jumlah nilai entropi dalam variable.

Selanjutnya algoritma ini akan diimplementasikan kedalam suatu system. Untuk mengetahui akurasi yang dihasilkan maka akan dilakukan evaluasi / pengujian terhadap system guna penarikan kesimpulan yang akurat. Evaluasi yang digunakan dalam penelitian ini akan menggunakan confusssion matrix.



Gambar 1. Alur penelitan

HASIL DAN PEMBAHASAN

Penelitian ini akan menggunakan 299 data training dan 270 data testing. Selanjutnya peneliti akan memastikan bahwa setiap atribut pada dataset telah siap untuk diproses. Rangkuman mengenai dataset awal sebelum dilakukan proses prepocessing data terangkum pada Tabel 2.

Table 2. Dataset awal

Usia	Anaemia	Creatinine	Diabetes	Ejection	Tekanan	Trom	Serum	Kematian
75	0	582	0	20	1	265000	1,9	1
55	0	7861	0	38	0	263358,03	1,1	1
65	0	146	0	20	0	162000	1,3	1
50	1	111	0	20	0	210000	1,9	1
65	1	160	1	20	0	327000	2,7	1
90	1	47	0	40	1	204000	2,1	1
75	1	246	0	15	0	127000	1,2	1
60	1	315	1	60	0	454000	1,1	1
65	0	157	0	65	0	263358,03	1,5	1
50	0	196	0	45	0	395000	1.6	0

Pada table 2 terlihat data masih belum dikelompokkan, dimana usia masih dituliskan secara berulang sesuai dengan dataset awal pada UCI. Untuk memudahkan proses pengolahan, maka peneliti akan mengelompokkan dataset tersebut sesuai batas minimum dan maksimumnya. Hasil pengelompokkan dataset tersebut terangkum pada table 3.

Table 3. Hasil pengelompokkan dataset

Usia	Anemia	Creatine	Diabetes	Ejection	Tekanan	Trom	Serum	kematian
60-80	0	200-600	0	<30	1	100000-500000	<2	120-140
<60	0	>1000	0	30-60	0	100000-500000	<2	120-140
60-80	0	<200	0	<30	0	100000-500000	<2	120-140
<60	1	<200	0	<30	0	100000-500000	<2	120-140
60-80	1	<200	1	<30	0	100000-500000	2-5	<120
>80	1	<200	0	30-60	1	100000-500000	2-5	120-140
60-80	1	200-600	0	<30	0	100000-500000	<2	120-140
<60	1	200-600	1	30-60	0	100000-500000	<2	120-140
60-80	0	<200	0	>60	0	100000-500000	<2	120-140

Kemudian dilanjutkan dengan data *transformation* yaitu mengubah data yang masih mentah dijadikan data yang memiliki value dalam menentukan keputusan. Pengubahan ini menggunakan rumus *Multiple Condision Range* ditunjukkan pada tabel 4.

Tabel 4. Nilai *Entropy* total

KEMATIAN	0 (tidak)	174	270	*Entropy*
	1 (ya)	96		0,938932011

Perhitungan untuk mencari nilai menggunakan Algoritma C5.0 dilakukan untuk mencari nilai *entropy*, *information* gain dan gain rasio untuk selanjutnya dipakai dalam pembuatan pohon keputusan. Data yang digunakan untuk bahan pengolahan dan hasil sebagaimana pada Tabel 5.

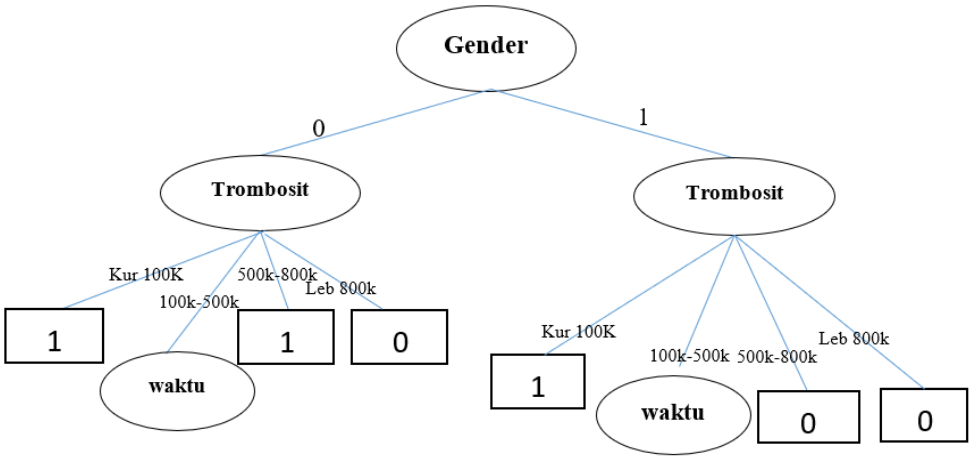
Tabel 5. Hasil Perhitungan Nilai Gain Rasio

	KEMATIAN	0	174	270	*Entropy*	Gain	Gain Rasio
		1	96		0,938932011	0,035567962	0,013597977
	Death Event			Jumlah			
USIA	Kur60	0	99				
		1	44	143	0,89049164		
	60-80	0	39				
		1	71	110	0,93806215		
	Leb80	0	13				
		1	4	17	0,787126586		
				270	2,615680376		

Selanjutnya akan dilakukan perhitungan informasi gain seperti table 6, dimana perhitungan nilai gain rasio dilakukan dengan nilai gain di bagi dengan jumlah entropy yang kemudian mencari nilai gain rasio tertinggi untuk membuat sebuah pohon keputusan seperti gambar 2.

Table 6. Perhitungan informasion gain

	KEMATIAN	0	174	270	*Entropy*	Gain
		1	96		0,938932011	0,035567962
	Death Event			Jumlah		
USIA	Kur60	0	99			
		1	44	143	0,89049164	
	60-80	0	39			
		1	71	110	0,93806215	
	Leb80	0	13			
		1	4	17	0,787126586	
				270	2,615680376	



Gambar 3. Pohon Keputusan

Pengujian

Untuk membuktikan bahwa sistem yang dibuat mampu memprediksi keberlangsungan hidup pasien, peneliti menggunakan pengujian confusion matrix pada hasil perhitungan yang dilakukan. Uji coba dilakukan dengan menggunakan 80:20 pada data testing dan data training yang didapatkan. Table 7 merupakan hasil dari True Positif (TP), True Negatif (TN), False Positif (FP) dan False Negatif (FN) uji coba system.

Tabel 8. Hasil perhitungan

TP	1
TN	25
FP	1
FN	3

Berdasarkan table 8, maka akan dilakukan perhitungan akurasi, presisi, dan sensitifitas dari system yang dibuat. Rangkuman mengenai hasil perhitungan tersebut terdapat pada table 9.

Table 9. Confusion Matriks

Akurasi	86,66666667
Presisi	89,65517241
Sensitifitas	25
Spesifisitas	96,15384615

Beradasarkan tabel 9 didapatkan bahwa ketepatan system dalam memprediksi hanya berkisar 86% dengan akurasi prediksi dari model yang dibangun adalah 89%. Sedangkan keberhasilan model dalam menemukan kembali informasi yang dihasilkan yaitu 25%.

KESIMPULAN

Berdasarkan hasil analisa pada 569 dataset dan 13 atribut pada *heart failure clinical records* yang telah dilakukan, didapatkan bahwa Decision Tree C5.0 mampu memprediksi tingkat kematian pasien gagal jantung dengan tingkat akurasi sebesar 86% dengan akurasi prediksi dari model yang dibangun adalah 89%.

DAFTAR PUSTAKA

[1] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[2] A. Purnama, "Edukasi dapat meningkatkan kualitas hidup pasien yang terdiagnosa penyakit jantung koroner," *J. Kesehat. Indones.*, vol. X, no. 2, pp. 66–71, 2020, doi: 10.33657/jurkessia.v10i2.251.

[3] D. Fitrianah, W. Gunawan, and A. P. Sari, "Studi Komparasi Algoritma Klasifikasi C5 . 0 , SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir," *Techno.COM*, vol. 21, no. 1, pp. 1–11, 2022, doi: 10.33633/tc.v21i1.5348.

[4] T. Permana, A. M. Siregar, A. F. N. Masruriyah, and A. R. Juwita, "Perbandingan Hasil Prediksi Kredit Macet Pada Koperasi Menggunakan Algoritma KNN dan C5.0," *Conf. Innov. Appl. Sci. Technol.*, vol. 3, no. 1, pp. 737–746, 2020.

[5] F. J. Zebua, R. P. Br Manalu, and M. N. K. Nababan, "Prediksi Kelulusan Mahasiswa Menggunakan Perbandingan Algoritma C5.0 Dengan Regression Linear," *J. Tek. Inf. dan Komput.*, vol. 4, no. 2, p. 230, 2021, doi: 10.37600/tekinkom.v4i2.400.

[6] M. S. Sungkar and M. T. Qurohman, "Penerapan Algoritma C5.0 Untuk Prediksi Kelulusan Pembelajaran Mahasiswa Pada Matakuliah Arsitektur Sistem Komputer," *J. Media Inform. Budidarma*, vol. 5, no. 3, p. 1166, 2021, doi: 10.30865/mib.v5i3.3116.

[7] M. Ainur Rohman and D. Arifianto, "Penerapan Metode Euclidean Probality dan Confusion Matrix dalam Diagnosa Penyakit Koi," *J. Smart Teknol.*, vol. 2, no. 2, pp. 122–130, 2021, [Online]. Available: file:///C:/Users/dokta/Downloads/4992-14056-2-PB.pdf

[8] H. W. Dhany, "Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung," *Semin. Nas. Inform. Pros. Senat. 2021*, pp. 177–179, 2021.

[9] I. Mubarog, A. Setyanto, and H. Sismoro, "Sistem Klasifikasi Pada Penyakit Breast Cancer Dengan Menggunakan Metode Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 2, p. 109, 2021, doi: 10.24076/citec.2019v6i2.246.

[10] R. Pratiwi, M. N. Hayati, and S. Prangga, "Perbandingan Klasifikasi Algoritma C5.0 Dengan Classification and Regression Tree (Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 2, pp. 273–284, 2020, doi: 10.30598/barekengvol14iss2pp273-284.

[11] M. Fajri, I. T. Utami, and M. Maruf, "Comparison of C4.5 and C5.0 Algorithm Classification Tree Models for Analysis of Factors Affecting Auction," *Indones. J. Stat. Its Appl.*, vol. 6, no. 1, pp. 13–22, 2022, doi: 10.29244/ijsa.v6i1p13-22.