 ISSN: 2684-8473

Deteksi Penyakit Jantung Menggunakan Metode Klasifikasi *Decision Tree* dan Regresi Logistik

Fahren Bukhari[1], Sri Nurdiati[2*], Mohamad Khoirun Najib[3], Rizki Nurul Amalia[4]

Departemen Matematika, IPB University, Jl. Meranti Kampus, Babakan, Dramaga, Bogor, 16680, Indonesia
E-mail: [1] fahrenbu@apps.ipb.ac.id; [2] nurdiati@apps.ipb.ac.id; [3] mkhoirun_najib@apps.ipb.ac.id*

INFORMASI ARTIKEL

Histori Artikel
Diterima : 03 April 2023
Direvisi : 14 April 2023
Diterbitkan : 30 April 2023

Kata Kunci:
Decision tree
Deteksi penyakit jantung
Klasifikasi data
Regresi logistik

ABSTRAK

Penyakit jantung merupakan salah satu penyakit paling umum dan kritis yang membahayakan kehidupan manusia. Selain diagnosis klinis, pembelajaran mesin dan pendekatan berbasis pembelajaran mendalam sangat penting dalam diagnosis penyakit jantung, seperti *decision tree* dan regresi logistik. Penelitian ini bertujuan membandingkan kedua metode klasifikasi tersebut untuk mendeteksi adanya penyakit jantung berdasarkan beberapa indikator. Data yang digunakan adalah data penyakit jantung yang dikeluarkan oleh *University of California, Irvine* (UCI) *Machine Learning Repository*. Berdasarkan hasil yang diperoleh, model *decision tree* yang terbentuk menempatkan variabel *thal* (tipe detak jantung pasien) sebagai simpul akar, dikarenakan nilai *entropy* yang paling tinggi. Model *decision tree* memiliki akurasi terhadap data uji sebesar 75%. Sementara itu, model regresi logistik menempatkan variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* sebagai variabel-variabel yang berpengaruh nyata. Model regresi logistik memiliki akurasi terhadap data uji sebesar 87%. Dari akurasi dari kedua model tersebut, regresi logistik lebih akurat untuk mendeteksi adanya penyakit jantung dibandingkan model *decision tree*.

I. Pendahuluan

Seiring dengan perkembangan zaman, kemajuan dalam proses pengumpulan data dan teknologi penyimpanan yang cepat dan akurat memungkinkan organisasi menghimpun jumlah data yang sangat luas. Hal ini mengakibatkan penggunaan alat dan teknik analisis data secara manual tentunya tidak dapat digunakan untuk mengekstrak informasi dari data yang sangat besar. Dalam menciptakan efisiensi pengumpulan data yang besar, diperlukan metode baru yang dapat menjawab kebutuhan tersebut. *Data mining* merupakan suatu teknologi yang dapat memproses data dalam volume besar untuk mengubah data mentah menjadi informasi yang berguna dalam membuat suatu keputusan bisnis yang penting. Pada dasarnya *data mining* mempunyai tujuh fungsi yaitu *description, classification, clustering, association, sequencing, forecasting,* dan *prediction* [1].

Klasifikasi merupakan teknik menemukan model yang dapat menjelaskan konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu obyek data. Dalam klasifikasi, data dibagi menjadi dua yaitu data latih dan uji [2]. Untuk mendapatkan model, harus dilakukan analisis terhadap data latih, sedangkan data uji digunakan untuk mengetahui tingkat akurasi dari model yang dihasilkan. Akurasi prediksi digunakan sebagai ukuran untuk membenarkan seberapa efisien algoritma tersebut dan cara algoritma klasifikasi data dapat melakukan klasifikasi instan dengan akurasi tinggi ke ruang fitur (atribut) yang benar.

Banyak teknik *data mining* telah diusulkan oleh para peneliti untuk memecahkan berbagai masalah klasifikasi dan pengelompokan data. Hosmer dan Lemeshow [3] mengatakan bahwa model regresi logistik merupakan metode yang digunakan untuk menganalisis hubungan antara satu variabel dependen dan beberapa variabel independen, dengan variabel dependennya berupa data biner, yaitu bernilai 1 untuk menyatakan benar dan bernilai 0 untuk menyatakan salah. Sementara itu, *decision tree* digunakan untuk mempelajari klasifikasi dan prediksi pola dari data serta menggambarkan relasi dari variabel atribut x dan variabel target y dalam bentuk pohon [4].

Decision tree dan regresi logistik merupakan metode yang sering digunakan dalam masalah klasifikasi. Asmianto *et al.* [5] telah membandingkan *decision tree*, *support vector machine*, dan *k-nearest neighbor* untuk memprediksi penyakit jantung. Sementara itu, Ambrish *et al.* [6] menggunakan teknik regresi logistik untuk prediksi penyakit kardiovaskular. Penyakit jantung merupakan salah satu penyakit paling umum dan kritis yang

membahayakan kehidupan manusia. Selain diagnosis klinis, pembelajaran mesin dan pendekatan berbasis pembelajaran mendalam sangat penting dalam diagnosis penyakit jantung [7].

Oleh karena itu, penelitian ini bertujuan untuk membandingkan kedua metode klasifikasi tersebut untuk mendeteksi adanya penyakit jantung berdasarkan beberapa indikator. Data yang digunakan adalah data penyakit jantung yang dikeluarkan oleh *University of California, Irvine* (UCI) *Machine Learning Repository*.

II. Data dan Metode

A. Data

Data yang digunakan merupakan data penyakit jantung "*UCI Heart Disease*" yang diambil dari laman kaggle.com. Data terdiri atas 13 variabel independen dan satu variabel dependen sebagai target klasifikasi. Rincian 13 variabel independen ditunjukkan pada Tabel 1 berikut. Sementara itu, variabel dependen dari dataset adalah *condition* dengan dua nilai atribut yaitu, nilai 1 berarti positif penyakit jantung dan 0 negatif penyakit jantung.

Tabel 1. Rincian 13 variabel independen untuk mengklasifikasikan ada tidaknya penyakit jantung.

No.	Nama	Keterangan
1	*Age*	Umur pasien
2	*Sex*	Jenis kelamin pasien. Atribut ini memiliki dua nilai, yakni nilai 1 untuk laki-laki dan 0 untuk perempuan.
3	*Cp*	Tipe nyeri dada yang diderita pasien. Atribut ini memilik empat nilai, yaitu 0: *asymptomatic* (tanpa gejala), 1: *atypical angina* (nyeri dada yang tidak bisa diprediksi), 2: *non-anginal pain* (gejala di luar penyakit jantung), dan 3: *typical angina* (nyeri dada yang memiliki gejala biasa).
4	*Trestbps*	*resting blood pressure* yaitu tekanan darah pasien ketika dalam keadaan istirahat. Satuan yang dipakai adalah mm·Hg.
5	*Chol*	*Cholesterol* yaitu kadar kolesterol dalam darah pasien, dengan satuan mg/dl.
6	*Fbs*	*fasting blood sugar* yaitu kadar gula darah pasien. Atribut fbs ini memiliki dua nilai yaitu 1: kadar gula darah pasien lebih dari 120 mg/dl, dan 0: kadar gula darah pasien kurang dari sama dengan 120 mg/dl.
7	*Restecg*	*resting electrocardiographic* yaitu kondisi ECG pasien ketika dalam keadaan istirahat. Atribut ini memiliki tiga nilai yaitu 1: untuk keadaan normal, 2: untuk keadaan ST-T *wave abnormality* yaitu keadaan gelombang *inversions T* dan/atau *ST* meningkat maupun menurun lebih dari 0.5 mV, dan 3: untuk keadaan *ventricular* kiri mengalami hipertropi.
8	*Thalach*	rata-rata detak jantung pasien dalam satu menit.
9	*Exang*	keadaan pasien akan mengalami nyeri dada apabila berolahraga. Atribut ini memiliki dua nilai yaitu 0: tidak nyeri, dan 1: menyebabkan nyeri.
10	*Oldpeak*	penurunan *ST* akibat olahraga.
11	*Slope*	*slope* dari puncak *ST* setelah berolah raga. Atribut ini memiliki 3 nilai yaitu 0: *downsloping*, 1: flat, dan 2: *upsloping.*
12	*Ca*	banyaknya pembuluh darah yang terdeteksi melalui proses pewarnaan *flourosopy.*
13	*Thal*	detak jantung pasien. Atribut ini memiliki 3 nilai yaitu 0: *fixed defect*, 1: normal, dan 2: *reversal defect*

B. Decision Tree

Decision tree mengacu pada penggunaan struktur pohon untuk mewakili himpunan keputusan atau klasifikasi data berdasarkan karakteristik data yang berbeda. *Decision tree* merupakan algoritma yang umum digunakan untuk pengambilan keputusan. *Decision tree* akan mencari solusi permasalahan dengan menjadikan kriteria sebagai simpul yang saling berhubungan membentuk struktur seperti pohon [8]. Pada *decision tree* terdapat tiga simpul, yaitu simpul akar, internal, dan daun. Simpul akar merupakan simpul teratas. Simpul ini ditentukan oleh atribut terbaik. Selanjutnya simpul akar memiliki cabang yang disebut simpul internal yang dapat dibagi menjadi cabang lagi jika masih belum mendapatkan nilai luaran. Terakhir adalah simpul daun. Hasil luaran dari klasifikasi didapatkan dari simpul ini dan tidak akan terbagi menjadi cabang lagi.

Salah satu metode pembagian cabang suatu *decision tree* yaitu Algoritma C4.5. Algoritma C4.5 menggunakan metode *divide* and *conquer* untuk membangun pohon yang sesuai. Algoritma C4.5 menggunakan data latih untuk menumbuhkan pohon. Nilai ukuran ketidakpastian (*entropy*) dan ukuran efektivitas suatu atribut dalam mengklasifikasikan data (*gain*) adalah rumus utama dalam algoritma C4.5 [9]. Nilai *entropy* untuk algoritma C4.5 dapat dihitung menggunakan persamaan

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

dengan S merupakan himpunan kasus, n merujuk pada jumlah partisi atribut S, dan p_i merupakan proporsi dari partisi ke-i pada kasus S. Sementara itu, nilai gain dihitung menggunakan persamaan

$$\text{Gain } (S, A) = \text{Entropy } (S) - \sum_{j=1}^{n} \frac{|A_j|}{|S|} \times \text{Entropy } (A_j) \quad (2)$$

dengan A_j merupakan partisi dari atribut A.

Algoritma C4.5 mampu mengatasi data kategorik dan numerik. Untuk data kategorik, algoritma C4.5 memilih salah satu kategori sebagai atribut yang terbaik menggunakan nilai gain tertinggi, sedangkan algoritma C4.5 mengubah data numerik menjadi dua kategori terlebih dulu menggunakan suatu batas tertentu untuk data numerik [10]. Tahapan membangun *decision tree* menggunakan algoritma C4.5 adalah sebagai berikut [11]:

1. memilih atribut dengan gain tertinggi sebagai akar,
2. membuat cabang pada setiap nilai,
3. membagi kasus dalam cabang, dan
4. mengulangi proses sampai semua kasus pada setiap cabang mempunyai kelas yang sama untuk menentukan atribut sebagai akar yang disesuaikan pada nilai gain paling tinggi dari atribut-atribut yang ada.

Decision tree pada penelitian ini menggunakan paket dari Python untuk sains data dan *machine learning* yaitu *Scikit-Learn*. Pada bagian pembuatan objek klasifikasi *decision tree*, ada beberapa parameter yang digunakan ditunjukkan pada Tabel 2.

Tabel 2. Rincian beberapa parameter yang digunakan pada pembuatan objek klasifikasi *decision tree*.

Parameter	Keterangan
Criterion	Kriteria yang digunakan apakah "gini" untuk kriteria pemisahan gini impurity atau "entropy" dimana kriteria pemisahannya adalah information Gain.
Splitter	Strategi untuk memilih split. Pilihan yang dapat digunakan antara "best" atau "random"
Max_depth	Batas maksimal kedalaman pohon keputusan. Jika ingin tanpa batas maka pilihan None, jika terbatas diinput angka integer sesuai keinginan
Min_samples_split	Batas minimal pembagian data pada pohon keputusan. Jika tanpa batas pilihan None, Jika terbatas diinput angka integer sesuai keinginan
Min_samples_leaf	Batas minimal pemecahan cabang pada pohon keputusan. Jika tanpa batas pilihan None, Jika terbatas diinput angka integer sesuai keinginan
Random_state	Kontrol untuk keacakan estimator
Max_leaf_nodes	Pencabangan pohon keputusan dengan jumlah simpul maksimum. Jika None, maka jumlah yang tidak terbatas dimungkinkan.

C. Regresi Logistik

Analisis regresi pada dasarnya merupakan suatu ilmu mengenai hubungan antara variabel dependen dengan satu atau lebih variabel independen, dengan maksud untuk memprediksi dan memperkirakan nilai-nilai variabel dependen berdasarkan nilai variabel independen yang telah diketahui (Ghozali dan Imam 2005). Regresi logistik adalah bagian dari analisis regresi yang digunakan ketika variabel dependen merupakan variabel dikotomi. Variabel dikotomi adalah variabel yang hanya memiliki dua kemungkinan nilai, yaitu sukses yang ditunjukkan dengan angka 1 dan gagal yang ditunjukkan dengan angka 0.

Model regresi logistik merupakan model yang berdistribusi Bernoulli, yaitu distribusi dari variabel acak yang hanya mempunyai dua kategori. Jika data hasil pengamatan memiliki p buah variabel independen X yaitu $X_1, X_2, X_3, \dots X_p$ dan satu variabel dependen Y dengan tiap data akan diperiksa ketepatannya sehingga nilai Y sebanyak $y_1, y_2, y_3, \dots y_n$, mempunyai dua kemungkinan nilai yaitu 0 dan 1.

Regresi logistik mempunyai tujuan untuk menduga pola keterkaitan antara variabel x dengan $\pi(x_i)$. Nilai $\pi(x_i)$ adalah nilai probabilitas suatu kejadian yang disebabkan oleh variabel x sehingga kemungkinan luaran yang diperoleh dari fungsi logistik bernilai 0 atau 1, yang diberikan oleh

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (3)$$

dengan

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right)$$

Adapun langkah-langkah dari regresi logistik dalam penelitian ini adalah

1). Estimasi Parameter

Metode untuk mengestimasi parameter regresi logistik adalah *Maximum Likelihood Estimation* (MLE). Metode maksimum *likelihood* menghasilkan nilai untuk parameter yang tidak diketahui dengan memaksimalkan kemungkinan memperoleh kumpulan data teramati, sehingga hampiran yang dihasilkan adalah yang paling mendekati dengan data yang diamati [12]. Fungsi *likelihood* untuk regresi logistik diberikan oleh

$$L(\beta) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (4)$$

2). Uji Serentak

Statistik uji serentak digunakan untuk menguji pengaruh variabel independen yang ada di dalam model secara serentak. Uji serentak dilakukan menggunakan *likelihood ratio* sebagai berikut:

$$G = -2 \ln \left(\frac{l_0}{l_p} \right) \quad (5)$$

dengan l_0 dan l_p masing-masing merupakan *likelihood* tanpa dan dengan variabel independen. Tahapan uji serentak adalah sebagai berikut.

a. merumuskan hipotesis: H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ atau H_1: paling tidak ada satu $\beta_i \neq 0$, $i = 1,2, \dots, p$
b. menentukan nilai l_0 dan l_p,
c. menghitung statistik uji menggunakan (5),
d. membandingkan nilai G dengan nilai $\chi^2_{(\alpha,db)}$ tabel dengan $db = k - 1$ dan k merupakan banyaknya variabel independen, dan
e. menafsirkan tolak H_0 jika $G > \chi^2_{(\alpha,db)}$ atau terima H_0 jika $G < \chi^2_{(\alpha,db)}$.

3). Uji Parsial

Uji Wald merupakan teknik pengujian yang digunakan dalam uji parsial. Uji tersebut bertujuan untuk mengetahui apakah setiap variabel independen berpengaruh terhadap model atau tidak. Uji Wald dapat diperhitungkan dengan cara membandingkan parameter yang ditaksir dengan galat baku dari parameter tersebut [3]. Hal ini dapat dirumuskan sebagai berikut

$$W = \left(\frac{\beta_i}{SE(\beta_i)} \right)^2 \quad (6)$$

dengan β_i merupakan estimasi parameter, dan $SE(\beta_i)$ adalah galat baku (*standard error*) yang bersesuaian. Adapun tahapan uji Wald adalah sebagai berikut:

a. merumuskan hipotesis, H_0: $\beta_i = 0$ atau H_1: $\beta_i \neq 0$ dengan $i = 1,2, \dots, p$
b. menentukan nilai β_i dan $SE(\beta_i)$,
c. menghitung uji parsial menggunakan rumus (6),
d. membandingkan tiap parameter W dari $\chi^2_{(db,1)}$ tabel, dan
e. menafsirkan tolak H_0 jika $W > \chi^2_{(\alpha,1)}$ atau terima H_0 jika $W < \chi^2_{(\alpha,1)}$.

4). Odds Ratio

Odds ratio adalah interpretasi variabel independen yang dikategorikan ke dalam 2 kategori yang dinyatakan dengan kode 0 atau 1. Secara umum, *odds ratio* merupakan sekumpulan peluang yang dibagi oleh peluang lainnya. Dalam hal ini, kategori pertama dibandingkan terhadap kategori kedua berdasarkan nilai odds ratio (φ) yang menyatakan kategori pertama berpengaruh φ kali dari kategori kedua terhadap peubah respon [13].

Tabel 3. Contoh perhitungan *odds ratio*.

Dependen	**Independen**	
	X_1	X_0
Y = 1	a	b
Y = 0	c	d

Rumus untuk penghitungan nilai *odds ratio* dapat dituliskan:

$$\varphi = \frac{a/c}{b/d} \quad (7)$$

Hasil dari perhitungan odds ratio akan digunakan sebagai perbandingan dari variabel penjelas, seperti variabel X_1 sebesar φ kali lebih tinggi dibandingkan X_0 terhadap pengaruh $Y = 1$.

D. Confusion Matrix

Confusion Matrix merupakan sebuah metode untuk evaluasi model klasifikasi menggunakan tabel matriks. Tabel matriks yang digunakan untuk mencari *confusion matrix* dituliskan seperti Tabel 4. *Confusion matrix* menghasilkan nilai akurasi dari implementasi metode klasifikasi data. Akurasi menyatakan jumlah data yang diklasifikasikan benar setelah dilakukan proses pengujian [14]. Nilai akurasi tersebut diberikan oleh persamaan berikut.

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (8)$$

Tabel 4. Contoh *confusion matrix*.

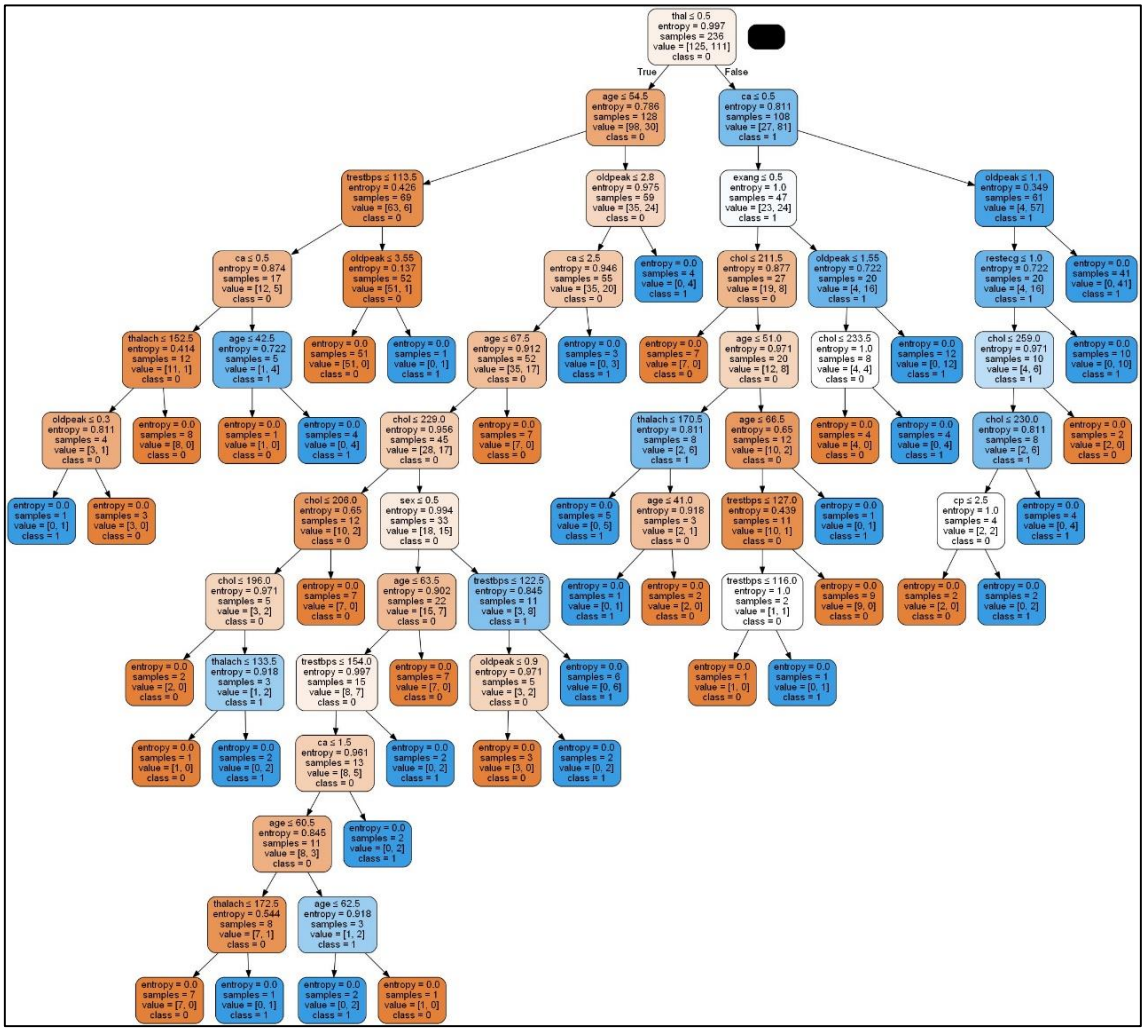
Label Prediksi	**Label Aktual**	
	1	***0***
1	TP (*True Positive*)	FP (*False Positive*)
0	FN (*False Negative*)	TN (*True Negative*)

III. Hasil dan Pembahasan

Pada metode klasifikasi data berbasis *machine learning*, langkah pertama sebelum implementasi metode adalah membagi data menjadi data latih dan uji. Data latih digunakan untuk melatih model klasifikasi dalam memprediksi keputusan, sedangkan data uji digunakan untuk mengukur akurasi dari metode klasifikasi yang digunakan. Pada penelitian ini, perbandingan data latih dan uji yang digunakan sebesar 80% data latih dibanding 20% data uji. Dengan demikian, dari 297 entri pada data "*UCI Heart Disease*", 236 entri data pertama digunakan sebagai data latih dan sisanya digunakan sebagai data uji.

A. Decision Tree

Dengan data latih yang telah ditetapkan, model *decision tree* dilatih dengan nilai parameter objek klasifikasi *decision tree* seperti pada Tabel 5 dan hasil *decision tree* ditunjukkan pada Gambar 1.



Gambar 1. Hasil *decision tree* klasifikasi deteksi penyakit jantung pada data UCI *heart disease*.

Tabel 5. Nilai parameter objek klasifikasi *decision tree*.

Parameter	Nilai
Criterion	entropy
Splitter	best
Max_depth	None
Min_samples_split	2
Min_samples_leaf	1
Random_state	None
Max_leaf_nodes	None

Berdasarkan hasil *decision tree* yang diperoleh, variabel independen yang menjadi simpul akar dalam memprediksi penyakit jantung adalah variabel *Thal* (detak jantung pasien) dengan *entropy* sebesar 0.997. Cabang dari simpul akar ini dibagi menjadi dua, yaitu berdasarkan *Thal* ≤ 0.5 (*Thal* bernilai 0 atau *fixed defect*) dan *Thal* > 0.5 (*Thal* bernilai 1 atau normal dan *Thal* bernilai 2 atau *reversal defect*). Simpul internal pertama setelah simpul akar ditentukan oleh umur dari pasien jika *Thal* bernilai 0, sedangkan jika *Thal* bernilai 1 atau 2, maka simpul internal pertama ditentukan oleh *ca* (banyaknya pembuluh darah yang terdeteksi melalui proses pewarnaan *flourosopy*).

Decision tree yang terbentuk memiliki kedalaman paling rendah yaitu dengan tiga simpul. Ini terjadi dengan dua jalur. Jika pasien memiliki *Thal* bernilai 0, umur lebih dari 54.5 tahun, dan *oldpeak* lebih dari 2.8, maka pasien diprediksi untuk memiliki penyakit jantung. Jalur terpendek lain yang memprediksi pasien untuk memiliki penyakit jantung adalah jika pasien memiliki *Thal* bernilai 1 atau 2, *ca* lebih dari 0.5, dan *oldpeak* lebih dari 1.1. Sementara itu, *decision tree* memiliki kedalaman paling tinggi yaitu dengan 12 simpul.

Model *decision tree* pada Gambar 1 dievaluasi terhadap data uji yang telah ditentukan untuk mengukur akurasi. *Confusion matrix* dari model *decision tree* terhadap data uji ditunjukkan pada Tabel 6. Berdasarkan *confusion matrix* tersebut, akurasi dari model *decision tree* adalah 75%.

Tabel 6. *Confusion matrix* dari model *decision tree* terhadap data uji.

Label Prediksi	Label Aktual	
	1	***0***
1	19	11
0	7	23

B. Regresi Logistik

Menurut Garavaglia dan Sharma ([15]) dalam model regresi logistik, semua variabel independen dengan jenis *multiclass* dikodekan sebagai variabel *dummy* untuk memberikan kemudahan interpretasi dan kalkulasi dari *odds ratio*, serta meningkatkan stabilitas dan signifikansi dari koefisien regresi. Oleh karena itu, sebelum melakukan estimasi parameter, data yang berbentuk kategorik terlebih dahulu diubah ke bentuk *dummies* agar dapat lebih mudah diklasifikasikan. Adapun beberapa variabel yang berbentuk kategorik tersebut adalah *cp*, *restecg*, *slope*, dan *thal*. Contoh pembentukan variabel *dummy* baru berdasarkan variabel *cp* dapat dilihat pada Tabel 7.

Tabel 7. Contoh pembentukan variabel *dummy* baru berdasarkan variabel *cp*.

cp	cp_1	cp_2	cp_3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

Dari empat nilai yang terdapat pada variabel *cp*, terbentuklah tiga variabel *dummy* yaitu *cp_1*, *cp_2*, dan *cp_3*. Jika cp bernilai 0, maka *cp_1*, *cp_2*, dan *cp_3* semuanya bernilai nol. Jika cp bernilai 1, maka *cp_1* bernilai 1 dan sisanya bernilai nol, dan seterusnya seperti yang dapat dilihat pada Tabel 7. Hal yang sama juga berlaku untuk tiga variabel lainnya yaitu *restecg*, *slope*, dan *thal*, yang masing-masing menghasilkan dua variabel *dummy* karena setiap variabel memiliki tiga nilai.

1). Estimasi Parameter

Pada tahap ini, seluruh variabel diuji terlebih dahulu untuk memperoleh variabel independen yang berpengaruh terhadap variabel dependen dengan cara mengeliminasi variabel independen, jika nilai signifikansi ($P < |z|$) variabel tersebut lebih besar dari 0.05. Dengan demikian, jika nilai signifikansi variabel independen kurang dari 0.05, maka variabel-variabel tersebut yang akan dimasukkan ke dalam model regresi logistik. Hasil estimasi parameter menggunakan seluruh variabel independen dapat dilihat pada Tabel 8.

Tabel 8. Hasil estimasi parameter menggunakan seluruh variabel independen.

Variabel	Koefisien	Std Error	z	Sig.
Intercept (β_0)	-7.9666	3.472	-2.294	0.022
Age (β_1)	-0.0046	0.027	-0.167	0.867
Sex (β_2)	1.6428	0.603	2.726	0.006
Cp_1 (β_{3_1})	1.6612	0.869	1.912	0.056
Cp_2 (β_{3_2})	0.8900	0.778	1.144	0.252
Cp_3 (β_{3_3})	2.4441	0.793	3.081	0.002
Trestbps (β_4)	0.0214	0.013	1.656	0.098
Chol (β_5)	0.0099	0.005	1.905	0.057
Fbs (β_6)	-0.5815	0.682	-0.852	0.394
Restecg_1 (β_{7_1})	0.6168	2.323	0.265	0.791
Restecg_2 (β_{7_2})	0.1710	0.430	0.397	0.691
Thalach (β_8)	-0.0168	0.013	-1.323	0.186
Exang (β_9)	0.5215	0.533	0.979	0.328
Oldpeak (β_{10})	0.2598	0.257	1.010	0.312
Slope_1 (β_{11_1})	1.4320	0.529	2.708	0.007
Slope_2 (β_{11_2})	1.1309	1.020	1.109	0.268
Ca (β_{12})	1.6070	0.350	4.591	0.000
Thal_1 (β_{13_1})	-0.1010	0.846	-0.119	0.905
Thal_2 (β_{13_2})	1.4127	0.493	2.867	0.004

Berdasarkan Tabel 8, dapat dilihat variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* mempunyai nilai signifikansi kurang dari 0.05, artinya variabel-variabel tersebut berpengaruh nyata. Sementara itu, variabel lainnya memiliki nilai signifikansi lebih besar dari 0.05. Dengan demikian, variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* akan diuji menggunakan cara yang sama, tetapi variabel yang tidak berpengaruh dihilangkan. Hasil estimasi parameter setelah pengurangan variabel ditunjukkan pada Tabel 9.

Tabel 9. Hasil estimasi parameter menggunakan seluruh variabel independen.

Variabel	Koefisien	Std Error	z	Sig.
Intercept (β_0)	-3.7913	0.551	-6.886	0.000
Sex (β_2)	1.0832	0.459	2.360	0.018
Cp_3 (β_{3_3})	1.8258	0.389	4.697	0.000
Slope_1 (β_{11_1})	1.3560	0.395	3.431	0.001
Ca (β_{12})	1.3378	0.272	4.925	0.000
Thal_2 (β_{13_2})	1.7130	0.418	4.097	0.000

Berdasarkan Tabel 9, nilai signifikansi variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* kurang dari 0.05 sehingga variabel-variabel tersebut mempunyai pengaruh secara signifikan. Dengan demikian, variabel-variabel tersebut dapat dimasukkan dalam model regresi logistik sebagai berikut:

$$\pi(x) = \frac{\exp(-3{,}7913 + 1{,}0832X_2 + 1{,}8258X_{3\_3} + 1{,}3560X_{11\_1} + 1{,}3378X_{12} + 1{,}7130X_{13\_2})}{1 + \exp(-3{,}7913 + 1{,}0832X_2 + 1{,}8258X_{3\_3} + 1{,}3560X_{11\_1} + 1{,}3378X_{12} + 1{,}7130X_{13\_2})} \quad (9)$$

2). Uji Serentak

Uji signifikansi model secara serentak menggunakan uji rasio *likelihood* yang diperoleh dengan cara membandingkan fungsi log *likelihood* menggunakan seluruh variabel independen dengan fungsi log *likelihood* tanpa variabel independen. Statistik uji rasio *likelihood* dari model regresi logistik yang diperoleh diberikan oleh

$$G = (-2\ln(l_0)) - (-2\ln(l_1)) = 74.219 \quad (10)$$

Pada pengujian ini nilai $\chi^2_{(\alpha,dk)} = \chi^2_{(0.05,5)}$. Berdasarkan tabel χ^2, nilai $\chi^2_{(0.05,5)} = 11.070$, maka nilai statistik uji $G > \chi^2_{(0.05,5)}$ dengan nilai 74.219 > 11.070. Hal ini menunjukkan H_0 ditolak pada tingkat signifikansi $\alpha = 0.05$ berarti bahwa terdapat paling sedikit ada satu parameter $\beta_i \neq 0$ yaitu terdapat satu atau lebih variabel independen yang berpengaruh signifikan terhadap variabel dependen.

3). Uji Parsial

Setelah melakukan uji serentak, langkah selanjutnya dilakukan pengujian signifikansi untuk masing-masing parameter dalam model dengan cara menguadratkan hasil bagi estimasi parameter β_n dengan *standard error* estimasi parameternya. Pengujian ini menggunakan tingkat signifikan $\alpha = 0.05$ dengan aturan keputusan H_0 ditolak pada tingkat signifikan α jika $W > \chi^2_{(0.05,1)}$ atau nilai signifikansinya lebih kecil dari α. Hasil uji parsial ditunjukkan pada Tabel 10 berikut.

Tabel 10. Hasil uji parsial untuk setiap variabel independen yang digunakan.

Variabel	Koefisien	Std Error	W	Sig.
Intercept (β_0)	-3.7913	0.551	47.345	0.000
Sex (β_2)	1.0832	0.459	5.569	0.018
Cp_3 (β_{3_3})	1.8258	0.389	22.0296	0.000
Slope_1 (β_{11_1})	1.3560	0.395	11.7849	0.001
Ca (β_{12})	1.3378	0.272	24.1904	0.000
Thal_2 (β_{13_2})	1.7130	0.418	16.7943	0.000

Tabel 10 menjelaskan bahwa parameter yang signifikan adalah semua variabel independent yang digunakan, karena variabel-variabel tersebut mempunyai nilai $W > \chi^2_{(0.05,1)} = 3.481$. Oleh karena itu variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* diputuskan tolak H_0, sehingga dapat disimpulkan bahwa variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* mempunyai pengaruh terhadap deteksi penyakit jantung.

4). Interpretasi Odds Ratio

Odds ratio menunjukkan besarnya pengaruh masing-masing variabel prediktor yang signifikan. *Odds ratio* dapat juga diartikan sebagai jumlah relatif dengan peluang hasil meningkat (*odds ratio* > 1) atau turun (*odds ratio* < 1). Setelah dilakukan pengolahan data, didapat nilai *odds ratio* masing-masing variabel prediktor yang berpengaruh terhadap variabel respon yang ditunjukkan pada Tabel 11.

Tabel 11. *Odds ratio* untuk setiap variabel independen yang digunakan.

Variabel	Odds ratio
Intercept (β_0)	0.022566
Sex (β_2)	2.954165
Cp (β_{3_3})	6.207985
Slope (β_{11_1})	3.880811
Ca (β_{12})	3.810838
Thal (β_{13_2})	5.545758

Interpretasi odds ratio masing-masing variabel adalah sebagai berikut:

a. *Sex* (X_2)

Jenis kelamin laki-laki 2.95 kali lebih besar kemungkinan mengalami penyakit jantung dibandingkan jenis kelamin perempuan.

b. *Cp_3* (X_{3_3})

Tipe nyeri dada yang gejalanya biasa dan mudah diprediksi memiliki kemungkinan mengalami penyakit jantung 6.21 kali lebih besar dari tipe nyeri yang memiliki gejala di luar penyakit jantung.

c. *Slope_1* (X_{11_1})

Upsloping ST berpotensi mengalami penyakit jantung 3.88 kali lebih tinggi dari *slope flat*.

d. *Ca* (X_{12})

Semakin tinggi 1 nilai Ca pada satu pasien, kecenderungan untuk mengalami penyakit jantung meningkat sebesar 3.81 kali dari 1 nilai Ca di bawahnya.

e. *Thal_2* (X_{13_2})

Penyakit *thallasemia* tipe normal berpotensi mengalami penyakit jantung 5.55 kali lebih tinggi dari tipe *fixed defect*.

5). Akurasi model

Model regresi logistik pada Persamaan 9 dievaluasi terhadap data uji yang telah ditentukan untuk mengukur akurasi. *Confusion matrix* dari model regresi logistik terhadap data uji ditunjukkan pada Tabel 12. Berdasarkan *confusion matrix* tersebut, akurasi dari model regresi logistik adalah 87%.

Tabel 12. *Confusion matrix* dari model regresi logistik terhadap data uji.

Label Prediksi	Label Aktual	
	1	***0***
1	23	5
0	3	29

IV. Simpulan

Berdasarkan hasil yang diperoleh, model *decision tree* yang terbentuk menempatkan variabel *thal* sebagai simpul akar untuk mendeteksi adanya penyakit jantung, dikarenakan nilai *entropy* yang paling tinggi. Model *decision tree* yang terbentuk memiliki kedalaman terpendek 3 simpul dan terpanjang 12 simpul. Model *decision tree* memiliki akurasi terhadap data uji sebesar 75%. Sementara itu, model regresi logistik menempatkan variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* sebagai variabel-variabel yang berpengaruh nyata. Model regresi logistik memiliki akurasi terhadap data uji sebesar 87%. Berdasarkan akurasi dari kedua model, regresi logistik lebih akurat untuk mendeteksi adanya penyakit jantung dibandingkan model *decision tree*.

Daftar Pustaka

[1] Mustika, Ardilla, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Guntoro, Manuhutu, A. M., Ridwan, M., & Hozairi. (2021). *Data Mining dan Aplikasinya*. Widiana Bhakti Persada.

[2] Anggriyani, I. R., Kusumawati, E. D., & Kawulur, E. I. J. J. (2022). etode Regresi Logistik Biner dan Metode K-Nearest Neighbor Pada Klasifikasi Menopause Dini Wanita Distrik Oransbari Provinsi Papua Barat. *UNEJ E-Proceeding*.

[3] Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/0471722146

[4] Ye, N. (2014). *Data Mining: Theories, Algorithms, and Examples*. CRC Press. https://doi.org/10.1201/b15288

[5] Asmianto, Pusawidjayanti, K., Hafiizh, M., & Supeno, I. (2022). Comparative of Classification Algorithm: Decision Tree, SVM, and KNN for Heart Diseases Prediction. *AIP Conference Proceedings*, *2639*. https://doi.org/10.1063/5.0110243

[6] Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, *3*(1), 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

[7] Albert, A. J., Murugan, R., & Sripriya, T. (2023). Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology. *Research on Biomedical Engineering*, *39*(1), 99–113. https://doi.org/10.1007/s42600-022-00253-9

[8] Babič, Š. H., Kokol, P., Podgorelec, V., Zorman, M., Šprogar, M., & Štiglic, M. M. (2000). The art of building decision trees. *Journal of Medical Systems*, *24*(1), 43–52. https://doi.org/10.1023/A:1005437213215

[9] Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition. In *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition* (Vol. 9780470908). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118874059

[10] Quinlan, J. R. (1993). *C4.5 - Programs for Machine Learning*. Morgan Kaufmann Publisher, Inc.

[11] Merawati, D., & Rino. (2019). Penerapan data mining penentu minat Dan bakat siswa SMK dengan metode C4.5. *Jurnal Algor*, *1*(1), 28–37.

[12] Safitri, A., Sudarmin, S., & Nusrang, M. (2019). Model Regresi Logistik Biner pada Tingkat Pengangguran Terbuka di Provinsi Sulawesi Barat Tahun 2017. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *1*(2), 1. https://doi.org/10.35580/variansiunm9354

[13] Mahmudin, M. Z., Rindengan, A., & Weku, W. (2014). Penggunaan Association Rule Data Mining Untuk Menentukan Pola Lama Studi Mahasiswa F-MIPA UNSRAT. *D'CARTESIAN*, *3*(1), 1. https://doi.org/10.35799/dc.3.1.2014.3777

[14] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[15] Garavaglia, S., & Sharma, A. (1998). a Smart Guide To Dummy Variables : Four Applications and a Macro. *Proceedings of the Northeast SAS Users Group Conference*, *43*.