

**PERBANDINGAN KLASIFIKASI DATA MENGGUNAKAN
DECISION TREE DAN REGRESI LOGISTIK
(STUDI KASUS : UCI *HEART DISEASE*)**

RIZKI NURUL AMALIA



**DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2023**

@Hak cipta milik IPB University

IPB University





PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan Klasifikasi Data menggunakan *Decision Tree* dan Regresi Logistik (Studi Kasus : UCI *Heart Disease*)” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Maret 2023

Rizki Nurul Amalia
G54180082



ABSTRAK

RIZKI NURUL AMALIA. Perbandingan Klasifikasi Data menggunakan *Decision Tree* dan Regresi Logistik (Studi Kasus : *UCI Heart Disease*). Dibimbing oleh FAHREN BUKHARI dan SRI NURDIATI.

Data mining menjadi sebuah inovasi yang dapat membantu pengumpulan data dalam jumlah besar. *Data mining* dapat digunakan oleh perusahaan untuk mengubah data mentah menjadi informasi yang berguna dalam pengambilan keputusan bisnis yang penting. *Data mining* mempunyai 7 fungsi yang salah satunya adalah fungsi klasifikasi data. Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menjelaskan konsep atau kelas data dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Penelitian ini membahas perbandingan metode *Decision Tree* dan Regresi Logistik dalam klasifikasi data. Data yang digunakan pada penelitian ini adalah “*UCI Heart Disease*” yang bersumber dari *website kaggle.com*. Hasil penelitian didapatkan persentase akurasi hasil dari metode *Decision Tree* adalah 75% sedangkan untuk metode Regresi Logistik adalah 87%.

Kata kunci: *decision tree*, klasifikasi data, regresi logistik,

ABSTRACT

RIZKI NURUL AMALIA. The Comparison of Data Classification using *Decision Tree* and Logistic Regression (Case Study : *UCI Heart Disease*). Supervised by FAHREN BUKHARI and SRI NURDIATI.

Data mining is an innovation that can help one collecting large amounts of data. Data mining can be used by companies to turn raw data into information that is useful in making important business decisions. Data mining has 7 functions, one of those function is a data classification. Classification is a technique used to find models in order to explain concepts or data classes with the goal of being able to estimate the class of an object whose label is unknown. This study learns the comparison of *Decision Tree* and Logistics Regression methods in data classifications. The data used in this study is “*UCI Heart Disease*” which downloaded from *kaggle.com*. The result of the research shows that the percentage of accuracy of the *Decision Tree* method is 75% while for the Logistic Regression method is 87%.

Keywords: *decision tree*, data classification, logistic regression



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2023
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB

**PERBANDINGAN KLASIFIKASI DATA MENGGUNAKAN
DECISION TREE DAN REGRESI LOGISTIK
(STUDI KASUS : UCI *HEART DISEASE*)**

RIZKI NURUL AMALIA

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Matematika pada
Program Studi Matematika

**DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2023**

@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.




Judul Skripsi : Perbandingan Klasifikasi Data menggunakan *Decision Tree* dan
Regresi Logistik (Studi Kasus : *UCI Heart Disease*)


Nama : Rizki Nurul Amalia
NIM : G54180082

Disetujui oleh

Pembimbing 1:
Dr. Ir. Fahren Bukhari, M.Sc

Pembimbing 2:
Prof. Dr. Ir. Sri Nurdianti, M.Sc

digitally signed @ design.ipb.ac.id

C17FFB86-BBA9-4558-9A2C-8E739869A872

digitally signed @ design.ipb.ac.id

C17FFB86-BBA9-4558-9A2C-8E739869A872

Diketahui oleh

Ketua Departemen:
Dr. Ir. Endar Hasafah Nugrahani , MS.
NIP. 196312281989032001

digitally signed

design.ipb.ac.id

Tanggal Ujian: 1 Februari 2023

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan November 2021 sampai bulan September 2022 ini ialah Klasifikasi Data, dengan judul “Perbandingan Klasifikasi Data menggunakan *Decision Tree* dan Regresi Logistik (Studi Kasus : UCI *Heart Disease*)”.

Pengerjaan karya ilmiah ini tidak luput dari bantuan dan dukungan dari berbagai pihak sehingga dapat diselesaikan dengan baik. Untuk itu, penulis ingin menyampaikan banyak terima kasih kepada berbagai pihak, terutama beberapa pihak berikut ini:

1. Ungkapan terima kasih juga disampaikan kepada Mama, Papa, Kakak, dan Tata, serta keluarga besar yang telah memberikan dukungan dan kasih sayangnya selama ini. Tanpa dukungan dan kasih sayang keluarga, penulis belum tentu dapat melewati tantangan selama ini.
2. Terima kasih penulis ucapkan kepada Dosen Pembimbing sekaligus Pembimbing Akademik penulis, yaitu Bapak Dr.Ir. Fahren Bukhari, M.Sc., kepada Dosen Pembimbing 2 ibu Prof. Dr. Ir. Sri Nurdiati, M.Sc., dan kepada Moderator Seminar dan Dosen Penguji bapak Mochamad Tito Julianto, S.Si, M.Kom., yang telah membimbing dan banyak memberi saran kepada penulis agar karya ilmiah ini tersusun dengan baik.
3. Terima kasih kepada teman-teman dari Departemen Matematika, teman seperjuangan. Berkat teman-teman seperjuangan, penulis dapat melewati masa perkuliahan di IPB dengan baik dan menyisakan kenangan yang indah bagi penulis.
4. Terima kasih kepada Dela, Riska, dan Valeria yang telah banyak membantu dan selalu ada bagi penulis sehingga dapat melalui masa perkuliahan dan merasakan suka dan duka selama perkuliahan
5. Terima kasih kepada Manda, Reisyah, dan Velia yang telah menjadi teman yang baik dalam menemani selama proses penulisan karya ilmiah ini.

Segala bentuk doa, dukungan, dan perhatian yang telah diberikan oleh berbagai pihak, penulis ucapkan terima kasih. Semoga karya ilmiah ini dapat bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan di Indonesia.

Bogor, Maret 2023

Rizki Nurul Amalia
G54180082



DAFTAR ISI

DAFTAR TABEL	x
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	x
I. PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan Penelitian	2
II. TINJAUAN PUSTAKA	3
2.1 <i>Decision Tree</i>	3
2.2 Regresi Logistik	8
2.3 <i>Confusion Matrix</i>	13
III. METODE PENELITIAN	15
3.1 Pengambilan Dataset	15
3.2 Pembagian Data <i>Training</i> dan Data <i>Testing</i>	16
3.3 Implementasi menggunakan <i>Decision Tree</i>	16
3.4 Implementasi menggunakan Regresi Logistik	16
3.5 Membandingkan Persentase Akurasi	16
IV. HASIL DAN PEMBAHASAN	17
4.1 Klasifikasi Data menggunakan <i>Decision Tree</i>	17
4.2 Klasifikasi Data menggunakan Regresi Logistik	21
V. SIMPULAN DAN SARAN	27
5.1 Simpulan	27
5.2 Saran	27
DAFTAR PUSTAKA	28
LAMPIRAN	30
RIWAYAT HIDUP	44



2.1	Tabel keputusan bermain atau tidak	5
2.2	Penyusunan jumlah kasus tiap atribut	5
2.3	<i>Entropy</i> keputusan bermain atau tidak	6
2.4	<i>Gain</i> bermain atau tidak	7
2.5	<i>Entropy</i> dan <i>gain</i> keputusan bermain atau tidak dengan kelembaban (tinggi) sebagai akar	7
2.6	Tabel perhitungan <i>odds ratio</i>	13
2.7	Tabel <i>Confusion Matrix</i>	13
4.1	Parameter objek klasifikasi <i>Decision Tree</i>	19
4.2	Perbandingan hasil <i>y</i> prediksi dan <i>y test Decision Tree</i>	20
4.3	<i>Confusion Matrix</i> metode <i>Decision Tree</i>	20
4.4	Tabel penaksiran parameter awal	23
4.5	Penaksiran parameter setelah pengurangan variabel	23
4.6	Uji <i>Wald</i> Variabel Regresi Logistik	25
4.7	<i>Odds Ratio</i> Variabel	25
4.8	<i>Confusion Matrix</i> Regresi Logistik	26

DAFTAR GAMBAR

2.1	Diagram <i>Decision Tree</i>	3
2.2	Cabang pertama pohon keputusan	7
2.3	Model akhir pohon keputusan bermain atau tidak	8
4.1	<i>Import</i> paket untuk membaca data yang digunakan di <i>Jupyter Notebook</i>	17
4.2	Pemisahan data <i>training</i> dan data <i>testing</i>	18
4.3	Membuat objek klasifikasi <i>decision tree</i>	18
4.4	Menentukan <i>y</i> prediksi	19
4.5	Mengubah <i>cp</i> , <i>restecg</i> , <i>slope</i> , dan <i>thal</i> ke bentuk <i>dummy</i>	21

DAFTAR LAMPIRAN

1	Koding Implementasi metode <i>Decision Tree</i>	31
2	Rules <i>Decision Tree</i>	35
3	Koding Implementasi metode Regresi Logistik	37
4	Tabel Chi-Square	42



I. PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan zaman, kemajuan dalam proses pengumpulan data dan teknologi penyimpanan yang cepat dan akurat memungkinkan organisasi menghimpun jumlah data yang sangat luas, sehingga 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 tentunya diperlukan metode baru yang dapat menjawab kebutuhan tersebut. *Data mining* merupakan sebuah teknologi yang dapat memproses data dalam volume besar yang digunakan oleh perusahaan untuk mengubah data mentah menjadi informasi yang berguna untuk membuat suatu keputusan bisnis yang sangat penting. Pada dasarnya *data mining* mempunyai 7 fungsi yaitu *Description*, *Classification*, *Clustering*, *Association*, *Sequencing*, *Forecasting*, dan *Prediction* (Mustika *et al.* 2021). Penelitian ini menggunakan metode *data mining* berupa klasifikasi.

Klasifikasi merupakan teknik yang digunakan untuk menemukan model agar dapat menjelaskan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Dalam mengklasifikasi, data dibagi menjadi dua yaitu data latih dan data uji (Anggriyani *et al.* 2022). Untuk mendapatkan model, harus dilakukan analisis terhadap data *training*, sedangkan data *testing* digunakan untuk mengetahui tingkat akurasi dari model yang dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu obyek data. Akurasi prediksi digunakan sebagai pengukuran untuk membenarkan seberapa efisien algoritma tersebut dan bagaimana klasifikasi data algoritma dapat melakukan klasifikasi instan dengan akurasi tinggi ke ruang fitur (atribut) yang benar.

Banyak teknik *data mining* telah diterapkan untuk memecahkan masalah klasifikasi dan pengelompokan data. Hosmer dan Lemeshow (2000) mengatakan bahwa model regresi logistik merupakan metode regresi logistik 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. Menurut Ye (2014) *Decision Tree* digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variabel atribut *x* dan variabel target *y* dalam bentuk pohon. Oleh karena itu peneliti tertarik untuk membandingkan kedua metode klasifikasi dan memilih “Perbandingan Klasifikasi Data menggunakan *Decision Tree* dan Regresi Logistik (Studi Kasus : UCI Heart Disease)” sebagai judul tugas akhir.

1.2 Tujuan Penelitian

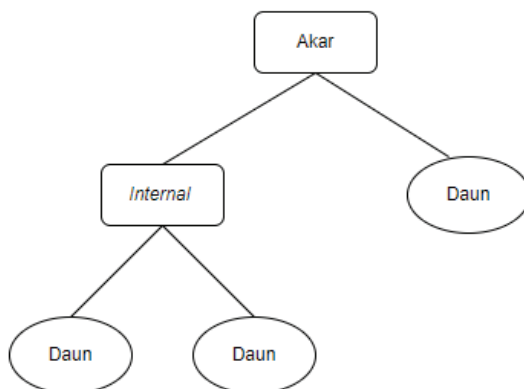
Tujuan dari penelitian ini adalah :

1. mengimplementasi klasifikasi data menggunakan metode *Decision Tree*,
2. mengimplementasi klasifikasi data menggunakan metode Regresi Logistik, dan
3. membandingkan persentase hasil klasifikasi data metode *Decision Tree* dan Regresi Logistik.

II. TINJAUAN PUSTAKA

2.1 Decision Tree

Decision Tree mengacu pada penggunaan struktur pohon untuk mewakili set keputusan atau klasifikasi data berdasarkan karakteristik data yang berbeda. Algoritma *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 seperti struktur pohon (Babic *et al.* 2000). Pada *Decision Tree* terdapat 3 simpul, yaitu simpul akar, *internal*, dan daun. Simpul akar merupakan simpul teratas, simpul ini ditentukan merupakan atribut terbaik. Selanjutnya simpul akar memiliki cabang yang disebut simpul *internal*, simpul ini dapat membagi menjadi cabang lagi jika masih belum mendapatkan nilai *output*. Terakhir adalah simpul daun, dimana hasil *output* dari klasifikasi didapatkan di simpul ini dan tidak akan terbagi menjadi cabang lagi. Diagram untuk *Decision Tree* ditunjukkan oleh Gambar 2.1



Gambar 2.1 Diagram *Decision Tree*

Pengerjaan menggunakan *Decision Tree* dapat digunakan salah satu metodenya yaitu Algoritma C4.5. Algoritma C4.5 adalah metode penyelesaian *Decision Tree* yang menggunakan metode *divide and conquer* untuk membangun pohon yang sesuai. *Decision Tree* Algoritma C4.5 menggunakan *dataset* pelatihan untuk menumbuhkan pohon dan menguji pohon yang dihasilkan. Nilai ukuran ketidakpastian (*entropy*) dan ukuran efektifitas suatu atribut dalam mengklasifikasikan data (*gain*) adalah rumus utama dalam pengerjaan *Decision Tree* Algoritma C4.5 (Larose D dan Larose C 2014).

Metode *Decision Tree* Algoritma C4.5 ini 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 untuk data numerik Algoritma C4.5 mengubah data numerik ke 2 kategori terlebih dahulu menggunakan *threshold*. Misalkan atribut dari data memiliki nilai a_0, a_1, \dots, a_n yang telah diurutkan dari nilai terkecil ke nilai terbesar, kemudian dicari nilai terbaik menggunakan rumus *threshold*

$t = \frac{a_n + a_{n+1}}{2}$ dan dipilih nilai *threshold* dengan *gain* terbaik untuk ditentukan menjadi nilai batas, $t \leq a$ dan $t > a$ (Quinlann 1993).

Tahapan membangun pohon keputusan menggunakan Algoritma C4.5 adalah sebagai berikut (Merawati dan Rino 2019) :

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, disesuaikan pada nilai *gain* paling tinggi dari atribut-atribut yang ada.

Nilai *entropy* untuk *Decision Tree* Algoritma C4.5 dapat dihitung menggunakan persamaan 2.1

$$Entropy(S) = \sum_{i=1}^n (-p_i) \times \log_2(p_i), \quad (2.1)$$

keterangan :

- S = himpunan kasus,
- n = jumlah partisi atribut S ,
- p_i = proporsi dari partisi ke- i kasus S terhadap S .

Sedangkan, untuk menghitung nilai *gain* digunakan persamaan 2.2

$$Gain(S, A) = Entropy(S) - \sum_{j=1}^n \frac{|A_j|}{|S|} \times Entropy(A_j), \quad (2.2)$$

keterangan :

- S = himpunan kasus,
- A_j = partisi dari atribut A .

Untuk memudahkan pemahaman terkait *Decision Tree*, diberikan satu kasus berikut ini. Pada suatu kasus digambarkan seseorang memiliki dua pilihan, yaitu untuk bermain tenis atau tidak bermain tenis. Pilihan tersebut didasarkan atas beberapa kondisi, yaitu Cuaca (Panas, Berangin, atau Hujan), Suhu (Panas, Normal, atau Dingin), Kelembaban (Tinggi atau Normal), dan Berangin (Ya atau Tidak). Kondisi tersebut digambarkan pada Tabel 2.1 untuk menentukan keputusan bermain atau tidak (Santosa 2007).

Tabel 2.1 Tabel keputusan bermain atau tidak

No	Cuaca	Suhu	Kelembaban	Berangin	Bermain
1	Cerah	Panas	Tinggi	Salah	Tidak
2	Cerah	Panas	Tinggi	Benar	Tidak
3	Berawan	Panas	Tinggi	Salah	Ya
4	Hujan	Normal	Tinggi	Salah	Ya
5	Hujan	Dingin	Normal	Salah	Ya
6	Hujan	Dingin	Normal	Benar	Ya
7	Berawan	Dingin	Normal	Benar	Ya
8	Cerah	Normal	Tinggi	Salah	Tidak
9	Cerah	Dingin	Normal	Salah	Ya
10	Hujan	Normal	Normal	Salah	Ya
11	Cerah	Normal	Normal	Benar	Ya
12	Berawan	Normal	Tinggi	Benar	Ya
13	Berawan	Panas	Normal	Salah	Ya
14	Hujan	Normal	Tinggi	Benar	Tidak

Untuk membangun pohon keputusan berdasarkan kasus tersebut, langkah pertama adalah menghitung *entropy* dari keputusan bermain atau tidak dengan Rumus 2.1. Untuk kepentingan penghitungan nilai *entropy*, Tabel 2.1 disusun kembali sesuai dengan jumlah kasus masing-masing atribut yang ditunjukkan pada Tabel 2.2,

Tabel 2.2 Penyusunan jumlah kasus tiap atribut

Atribut	Partisi Atribut	Jumlah Kasus (S)	Bermain	
			Tidak	Ya
Keputusan		14	4	10
Cuaca	Berawan	4	0	4
	Cerah	5	1	4
	Hujan	5	3	2
Suhu	Dingin	4	0	4
	Panas	4	2	2
	Normal	6	2	4
Kelembaban	Tinggi	7	4	3
	Normal	7	0	7
Berangin	Benar	8	2	6
	Salah	6	2	4

Setelah menyusun jumlah kasus pada Tabel 2.2, dihitung *entropy* dari tiap kasus dan atribut menggunakan Rumus 2.1. Kasus pertama yang akan dihitung nilai *entropy* nya adalah *entropy* dari bermain ditunjukkan sebagai berikut :

- $$\begin{aligned} \text{Entropy (Keputusan)} &= \left(-\frac{\text{jumlah nilai Tidak}}{\text{jumlah kasus}} \right) \times \log_2 \left(\frac{\text{jumlah nilai Tidak}}{\text{jumlah kasus}} \right) + \left(-\frac{\text{jumlah nilai Ya}}{\text{jumlah kasus}} \right) \times \\ &\quad \log_2 \left(\frac{\text{jumlah nilai Ya}}{\text{jumlah kasus}} \right) \\ &= \left(-\frac{4}{14} \right) \times \log_2 \left(\frac{4}{14} \right) + \left(-\frac{10}{14} \right) \times \log_2 \left(\frac{10}{14} \right) \\ &= 0,863120569 \end{aligned}$$
- $$\text{Entropy (Cuaca = Berawan)} = \left(-\frac{0}{4} \right) \times \log_2 \left(\frac{0}{4} \right) + \left(-\frac{4}{4} \right) \times \log_2 \left(\frac{4}{4} \right) = 0$$

Selanjutnya dihitung setiap *entropy* dari atribut yang ada, menggunakan cara yang sama. Hasil perhitungan *entropy* dari tiap kasus dan atribut yang ada ditunjukkan pada Tabel 2.3.

Tabel 2.3 *Entropy* keputusan bermain atau tidak

Atribut	Partisi Atribut	Jumlah Kasus (S)	Bermain		<i>Entropy</i>
			Tidak	Ya	
Keputusan		14	4	10	0,863120569
Cuaca	Berawan	4	0	4	0
	Cerah	5	1	4	0,721928095
	Hujan	5	3	2	0,970950594
Suhu	Dingin	4	0	4	0
	Panas	4	2	2	1
	Normal	6	2	4	0,918295934
Kelembaban	Tinggi	7	4	3	0,985228136
	Normal	7	0	7	0
Berangin	Benar	8	2	6	0,811278124
	Salah	6	2	4	0,918295834

Langkah selanjutnya mengitung nilai *gain* untuk menentukan atribut mana yang akan dijadikan akar dari pohon keputusan dengan salah satu contoh perhitungan sebagai berikut :

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

$$\begin{aligned} \text{Gain (Keputusan, Cuaca)} &= \text{Entropy (Keputusan)} - \left(\left(\frac{\text{Jumlah Berawan}}{\text{Jumlah Keputusan}} \times \text{entropy(Berawan)} \right) + \right. \\ &\quad \left(\frac{\text{Jumlah Hujan}}{\text{Jumlah Keputusan}} \times \text{entropy(Hujan)} \right) + \left(\frac{\text{Jumlah Cerah}}{\text{Jumlah Keputusan}} \times \right. \\ &\quad \left. \left. \text{entropy (Cerah)} \right) \right) \end{aligned}$$

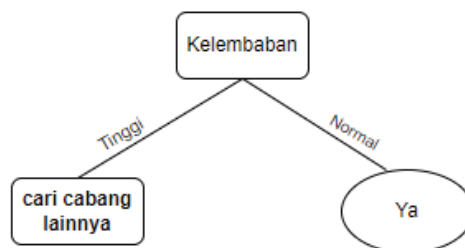
$$\begin{aligned} \text{Gain (Keputusan, Cuaca)} &= 0,863120569 - \left(\left(\frac{4}{14} \times 0 \right) + \left(\frac{5}{14} \times 0,721928095 \right) + \left(\frac{5}{14} \times 0,970950594 \right) \right) \\ &= 0,258521037 \end{aligned}$$

Selanjutnya dihitung nilai *gain* dari semua atribut menggunakan cara yang sama dan hasilnya dituliskan pada Tabel 2.4,

Tabel 2.4 *Gain* Bermain atau tidak

Atribut	Partisi Atribut	Jumlah Kasus (S)	Bermain		<i>Entropy</i>	<i>Gain</i>
			Tidak	Ya		
Keputusan		14	4	10	0,863120569	
Cuaca	Berawan	4	0	4	0	0,258521037
	Cerah	5	1	4	0,721928095	
	Hujan	5	3	2	0,970950594	
Suhu	Dingin	4	0	4	0	0,183850925
	Panas	4	2	2	1	
	Normal	6	2	4	0,918295934	
Kelembaban	Tinggi	7	4	3	0,985228136	0,370506501
	Normal	7	0	7	0	
Berangin	Benar	8	2	6	0,811278124	0,005977711
	Salah	6	2	4	0,918295834	

Tabel 2.4 menunjukkan Kelembaban dipilih sebagai akar dari *Decision Tree* karena memiliki nilai *gain* tertinggi, sehingga struktur pohon dapat digambarkan pada Gambar 2.2.



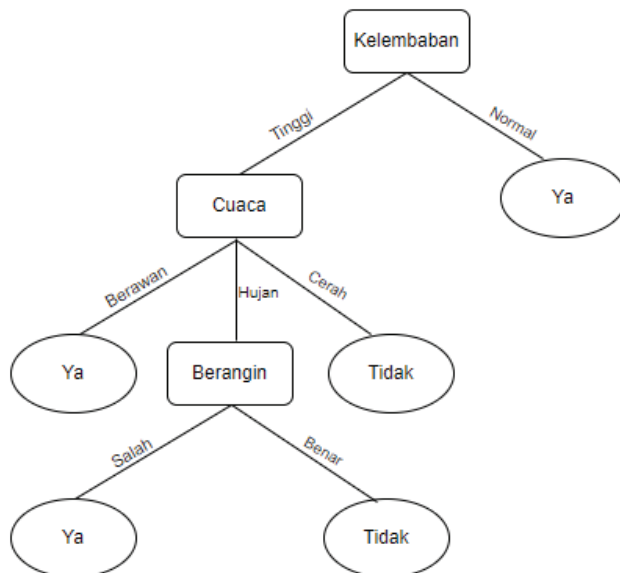
Gambar 2.2 Cabang pertama keputusan

Untuk melanjutkan pembuatan struktur pohon, Gambar 2.2 menggambarkan Kelembaban Tinggi sebagai akar karena belum didapatkan keputusan bermain atau tidak, sehingga proses perhitungan dilanjut seperti langkah pertama. Kelembaban Tinggi sebagai akar ditunjukkan dalam Tabel 2.5.

Tabel 2.5 *Entropy* dan *gain* keputusan bermain atau tidak dengan kelembaban (tinggi) sebagai akar

Atribut	Partisi Atribut	Jumlah Kasus (S)	Bermain		<i>Entropy</i>	<i>Gain</i>
			Tidak	Ya		
Kelembaban (Tinggi)		7	4	3	0,985228136	
Cuaca	Berawan	4	0	4	0	0,69951385
	Cerah	5	1	4	1	
	Hujan	5	3	2	0	
Suhu	Dingin	4	0	4	0	0,020244207
	Panas	4	2	2	1	
	Normal	6	2	4	2	
Berangin	Benar	8	2	6	1	0,020244207
	Salah	6	2	4	0,918295834	

Tabel 2.5 menunjukkan Cuaca merupakan atribut selanjutnya yang dipilih karena memiliki nilai *gain* tertinggi. Proses pengerjaan terus dilakukan berulang sampai semua cabang memiliki hasil keputusan, sehingga didapatkan pohon keputusan bermain atau tidak digambarkan pada Gambar 2.3.



Gambar 2.3 Model akhir keputusan bermain atau tidak

2.2 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, dimana distribusi Bernoulli adalah distribusi dari peubah 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. Jika variabel berdistribusi Bernoulli dengan parameter $\pi(x_i)$, maka fungsi distribusi peluang menjadi :

$$f(y_i) = [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}, y_i = 0, 1 \quad (2.3)$$

dengan

$i = 1, 2, \dots, n$,

n = banyaknya jumlah kasus,

jika dimasukkan nilai $y_i = 0$ diperoleh $f(0) = 1 - \pi(x_i)$ dan untuk nilai $y_i = 1$ diperoleh $f(1) = \pi(x_i)$. Menurut Hosmer dan Lemeshow (2000), model umum regresi logistik dengan p buah variabel prediktor dibentuk dengan nilai $\pi(x) = E(Y = 1|x)$, $\pi(x)$ dinotasikan sebagai berikut :

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (2.4)$$

dengan $g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$;

dimana β_p = Koefisien variabel ke- p

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 *output* yang diperoleh dari fungsi logistik bernilai 0 atau 1. Apabila $\pi(x_i)$ nilai harapan dan kategori satu terjadi maka $0 \leq \pi(x_i) \leq 1$. Nilai transformasi logit dari nilai $\pi(x_i)$ diperoleh bentuk sebagai berikut (Rizki *et al.* 2015)

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

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}$$

$$e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} = \pi(x)(1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p})$$

$$e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} = \pi(x) + \pi(x)e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

$$\pi(x) = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} - \pi(x)e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

$$\pi(x) = (1 - \pi(x))e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

$$\frac{\pi(x)}{(1 - \pi(x))} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}$$

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.5)$$

dengan $g(x)$ merupakan fungsi hubungan dari model regresi logistik yang disebut fungsi hubungan logit. Pada regresi logistik dilakukan beberapa uji,

yaitu penaksiran parameter, uji serentak, uji parsial, dan *odds ratio* untuk interpretasi model.

2.2.1. Penaksiran Parameter

Metode untuk mengestimasi parameter regresi logistik adalah dengan menggunakan metode MLE (*Maximum Likelihood Estimation*). Metode ini akan memberikan landasan untuk pendekatan dengan regresi logistik. Pengertian paling umum metode *maksimum likelihood* adalah menghasilkan nilai untuk parameter yang tidak diketahui yang memaksimalkan kemungkinan memperoleh kumpulan data teramati, sehingga penaksir yang dihasilkan adalah yang paling mendekati dengan data yang diamati (Safitri *et al.* 2019). Jika peubah dependen (Y). dikodekan sebagai 0 maka diberikan pernyataan untuk persamaan regresi logistik memberikan probabilitas $Y = 0$ dengan syarat x yang dilambangkan sebagai $P(Y = 0|x)$. Jika peubah dependen (Y) dikodekan sebagai 1 maka diberikan pernyataan untuk persamaan regresi logistik memberikan probabilitas $Y = 1$ dengan syarat x yang dilambangkan sebagai $P(Y = 1|x)$. Kemudian, untuk pasangan pada kasus ke- i (x_i, y_i), dengan $y_i = 1$, kontribusi untuk fungsi *likelihood* adalah $\pi(x_i)$, dan untuk pasangan dengan $y_i = 0$, kontribusi untuk fungsi *likelihood* adalah $1 - \pi(x_i)$ dengan kuantitas $\pi(x_i)$ menunjukkan nilai $\pi(x)$ dihitung pada x_i , maka fungsi *maksimum likelihood* adalah:

$$L(\beta) = \prod_{i=1}^n [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i}, \beta = \pi(x_i), \quad (2.6)$$

keterangan :

- y_i = pengamatan pada variabel respon ke- i
- $\pi(x_i)$ = peluang untuk variabel prediktor ke- i

Untuk mempermudah perhitungan, maka dilakukan penaksiran parameter dengan cara memaksimumkan fungsi logaritma kemungkinannya (*loglikelihood*), yaitu :

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \\ \ln(L(\beta)) &= \ln \left(\prod_{i=1}^n [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \right) \\ &= \sum_{i=1}^n \ln ([\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i}) \\ &= \sum_{i=1}^n (\ln[\pi(x_i)]^{y_i} + \ln[1 - \pi(x_i)]^{1-y_i}) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \\
&= \sum_{i=1}^n \{y_i \ln[\pi(x_i)]\} + \sum_{i=1}^n \{(1 - y_i) \ln[1 - \pi(x_i)]\} \quad (2.7)
\end{aligned}$$

Untuk mendapatkan nilai penaksiran koefisien regresi logistik ($\hat{\beta}$) dilakukan dengan membuat turunan pertama $L(\beta)$ terhadap $\pi(x_i)$ dan disamakan dengan nol (Herrhyanto, 2003). Dengan menggunakan cara tersebut, maka akan didapatkan fungsi sebagai berikut ;

$$\begin{aligned}
\ln(L(\beta)) &= \sum_{i=1}^n \{y_i \ln[\pi(x_i)]\} + \left(\sum_{i=1}^n 1 - \sum_{i=1}^n y_i \right) \ln[1 - \pi(x_i)] \\
\ln(L(\beta)) &= \sum_{i=1}^n \{y_i \ln[\pi(x_i)]\} + \left(n - \sum_{i=1}^n y_i \right) \ln[1 - \pi(x_i)]
\end{aligned}$$

Selanjutnya $L(\beta)$ diturunkan terhadap $\pi(x_i)$ menghasilkan :

$$\begin{aligned}
\frac{d}{d\beta} (\ln L(\beta)) &= \frac{\sum_{i=1}^n y_i}{\pi(x_i)} + \frac{n - \sum_{i=1}^n y_i}{1 - \pi(x_i)} (-1) \\
&= \frac{\sum_{i=1}^n y_i}{\pi(x_i)} - \frac{n - \sum_{i=1}^n y_i}{1 - \pi(x_i)}
\end{aligned}$$

Turunan dari $L(\beta)$ kemudian dibuat sama dengan 0, sehingga :

$$\begin{aligned}
\frac{\sum_{i=1}^n y_i}{\pi(x_i)} - \frac{n - \sum_{i=1}^n y_i}{1 - \pi(x_i)} &= 0 \\
\frac{(1 - \pi(x_i)) \sum_{i=1}^n y_i}{\pi(x_i)(1 - \pi(x_i))} - \frac{\pi(x_i)(n - \sum_{i=1}^n y_i)}{\pi(x_i)(1 - \pi(x_i))} &= 0 \\
(1 - \pi(x_i)) \sum_{i=1}^n y_i - (\pi(x_i)(n - \sum_{i=1}^n y_i)) &= 0 \\
\sum_{i=1}^n y_i - \sum_{i=1}^n y_i \pi(x_i) - n\pi(x_i) + \sum_{i=1}^n y_i \pi(x_i) &= 0 \\
\sum_{i=1}^n y_i - n\pi(x_i) &= 0 \\
n\pi(x_i) &= \sum_{i=1}^n y_i \\
\pi(x_i) &= \frac{\sum_{i=1}^n y_i}{n} = \bar{y} \quad (2.8)
\end{aligned}$$

Karena $\beta = \pi(x_i)$ maka didapatkan $(\hat{\beta})$ yang merupakan penduga kemungkinan maksimum.

2.2.2 Uji Serentak

Pengujian serentak menggunakan *Likelihood Ratio* terhadap parameter model dilakukan untuk memeriksa peranan variabel-variabel independen yang ada dalam model terhadap variabel dependennya. Statistik uji serentak digunakan untuk menguji pengaruh variabel independen di dalam model secara serentak. Uji serentak dilakukan dengan rumus sebagai berikut :

$$G = -2 \ln \left[\frac{l_0}{l_p} \right], \quad (2.9)$$

keterangan :

- l_0 = *likelihood* tanpa variabel independen, dan
- l_p = *likelihood* dengan variabel independen.

Tahapan pengujian sebagai berikut :

1. rumusan hipotesis
 $H_0: \beta_1 = \beta_2 = \dots = \beta_p$
 H_1 : paling tidak ada satu $\beta_i \neq 0, i = 1, 2, \dots, p$
2. menentukan nilai l_0 dan l_p ,
3. menghitung statistik uji menggunakan rumus (2.9),
4. memperoleh suatu keputusan yang optimal dengan membandingkan nilai G dengan nilai $\chi^2_{(\alpha, db)}$ tabel dimana $db = k - 1$ dan k merupakan banyaknya variabel independen, dan
5. menafsirkan tolak H_0 jika $G > \chi^2_{(\alpha, db)}$ atau terima H_0 jika $G < \chi^2_{(\alpha, db)}$.

2.2.3 Uji Parsial

Uji parsial atau *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 (Hosmer dan Lemeshow 2000), hal ini dapat dirumuskan sebagai berikut:

$$W = \left(\frac{\beta_i}{SE(\beta_i)} \right)^2, \quad (2.10)$$

keterangan :

- β_i = estimasi parameter, dan
- $SE(\beta_i)$ = *standard error*.

Tahapan pengujian sebagai berikut :

1. rumusan hipotesis,
 $H_0: \beta_i = 0$
 $H_1: \beta_i \neq 0$
dengan $i = 1, 2, \dots, p$

2. menentukan nilai β_i dan $SE(\beta_i)$,
3. menghitung uji parsial menggunakan rumus (2.10),
4. membandingkan tiap tiap parameter W dari $\chi^2_{(db,1)}$ tabel, dan
5. menafsirkan tolak H_0 jika $W > \chi^2_{(\alpha,1)}$ atau terima H_0 jika $W < \chi^2_{(\alpha,1)}$.

2.2.4. Odds Ratio

Odds ratio adalah interpretasi koefisien penjelas (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 (Zainal *et al.* 2014).

Tabel 2.6 Tabel 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:

$$\phi = \frac{a/c}{b/d} \quad (2.11)$$

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

2.3. Confusion Matrix

Confusion Matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matriks. Tabel matriks yang digunakan untuk mencari *Confusion Matrix* dituliskan seperti Tabel 2.7. Hasil evaluasi dengan menggunakan *Confusion Matrix* menghasilkan nilai akurasi dari implementasi metode klasifikasi data. Akurasi menyatakan jumlah data yang diklasifikasikan benar setelah dilakukan proses pengujian (Sokolova dan Lapalme 2009).

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+FN+TN} \times 100\%,$$

Tabel 2. 7. Tabel *Confusion Matrix*

Actual Label	Predicted Label	
	0	1
0	TP	FP
1	FN	TN

keterangan :

- $TP = \text{True Positive}$ yaitu jumlah data aktual berlabel positif yang terklasifikasi dengan benar oleh sistem prediksi,
- $TN = \text{True Negative}$ yaitu jumlah data aktual berlabel negatif yang terklasifikasi dengan benar oleh sistem prediksi,
- $FN = \text{False Negative}$ yaitu jumlah data aktual berlabel positif tapi diklasifikasi model sebagai negatif, dan
- $FP = \text{False Positive}$ yaitu jumlah data aktual berlabel negatif tapi diklasifikasi model sebagai positif.

@Hak cipta milik IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

III. METODE PENELITIAN

Ide dasar dalam penelitian ini adalah membandingkan dua metode klasifikasi data, yaitu *Decision Tree* dan Regresi Logistik untuk menentukan metode klasifikasi yang lebih baik dalam pengerjaan kasus sesuai *dataset* yang digunakan. Pada bab ini dijelaskan tahapan yang digunakan dalam penelitian tugas akhir.

3.1. Pengambilan *Dataset*

Dataset yang digunakan pada penelitian ini merupakan data penyakit jantung “ *UCI Heart Disease* “ yang diambil dari *website* Kaggle.com. Data terdiri atas 13 variabel independen dan 1 variabel dependen sebagai target klasifikasi.

Variabel independen sebagai berikut:

1. *Age*: Umur pasien.
2. *Sex*: Jenis kelamin pasien, atribut ini memiliki 2 nilai, yakni nilai 1 untuk laki-laki dan nilai 0 untuk perempuan.
3. *Cp*: Tipe nyeri dada yang diderita pasien. Atribut ini memiliki 4 nilai, yaitu Nilai 0: *asymptomatic* (tanpa gejala), Nilai 1: *atypical angina* (nyeri dada yang tidak bisa diprediksi), Nilai 2: *non-anginal pain* (gejala di luar penyakit jantung), Nilai 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 hanya memiliki 2 nilai yaitu 1 jika kadar gula darah pasien lebih dari 120 mg/dl, dan 0 jika 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 3 nilai yaitu nilai 1 untuk keadaan normal, nilai 2 untuk keadaan *ST-T wave abnormality* yaitu keadaan dimana gelombang *inversions* T dan atau ST meningkat maupun menurun lebih dari 0,5 mV dan nilai 3 untuk keadaan dimana *ventricular* kiri mengalami hipertropi.
8. *Thalach*: rata-rata detak jantung pasien dalam satu menit.
9. *Exang*: keadaan dimana pasien akan mengalami nyeri dada apabila berolah raga, 0 jika tidak nyeri, dan 1 jika menyebabkan nyeri.
10. *Oldpeak*: penurunan ST akibat olahraga.
11. *Slope*: *slope* dari puncak ST setelah berolah raga. Atribut ini memiliki 3 nilai yaitu 0 untuk *downsloping*, 1 untuk *flat*, dan 2 untuk *upsloping*.
12. *Ca*: banyaknya pembuluh darah yang terdeteksi melalui proses pewarnaan *flourosopy*.
13. *Thal*: detak jantung pasien. Atribut ini memiliki 3 nilai yaitu 0 untuk *fixed defect*, 1 untuk normal dan 2 untuk *reversal defect*.

Variabel dependen dari *dataset* adalah *condition* dimana nilai 1 berarti positif penyakit jantung dan 0 negatif penyakit jantung.

@Hak cipta milik IPB University

3.2. Pembagian Data *Training* dan Data *Testing*

Pada penelitian metode klasifikasi data, langkah pertama sebelum mengimplementasi adalah membagi data menjadi data *training* dan data *testing*. Data *training* berguna sebagai bentuk model klasifikasi dalam memprediksi keputusan, sedangkan data *testing* digunakan untuk mengukur sejauh mana akurasi melakukan klasifikasi dengan tepat. Pada penelitian ini perbandingan data *training* dan data *testing* yang digunakan sebesar 80:20.

3.3. Implementasi menggunakan *Decision Tree*

Implementasi menggunakan *Decision Tree* dilakukan menggunakan bantuan *software Python*. Penelitian ini akan dibantu menggunakan paket *Scikit Learn* yang sudah tersedia di *Python* untuk mengerjakan beberapa metode klasifikasi data yang salah satu di antaranya adalah metode *Decision Tree*.

3.4. Implementasi menggunakan Regresi Logistik

Implementasi menggunakan Regresi Logistik dilakukan menggunakan *software Python*. Penelitian ini akan dibantu menggunakan paket *Scikit Learn* dan *Statmodel* untuk menentukan model statistika dari Regresi Logistik.

3.5. Membandingkan Persentase Akurasi

Setelah mengerjakan implementasi *Decision Tree* dan Regresi Logistik akan didapatkan dua nilai akurasi dari masing-masing metode. Nilai akurasi dari masing-masing metode kemudian akan dibandingkan dan dituliskan kesimpulan terkait perbandingan nilai akurasi kedua metode.

IV. HASIL DAN PEMBAHASAN

Pada bab ini akan dibahas pengimplementasian metode *Decision Tree* dan Regresi Logistik untuk pengklasifikasian data. Setelah didapatkan hasil klasifikasi dari tiap metode, akan dilakukan perbandingan persentase akurasi untuk menentukan metode yang terbaik dalam klasifikasi pada data yang digunakan

4.1. Klasifikasi Data menggunakan *Decision Tree*

Decision Tree pada penelitian yang dilakukan menggunakan paket dari *Python* untuk data sains dan *machine learning* yaitu *Scikit-Learn*

4.1.1 Implementasi *Decision Tree*

Dalam pengerjaan *Decision Tree* menggunakan *Python* langkah yang paling utama adalah membaca *dataset* yang digunakan dengan cara seperti yang ditunjukkan pada Gambar 4.1.

```
import numpy as np
import pandas as pd
heart = pd.read_csv('data.csv')
```

heart

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
0	59	1	1	140	221	0	0	164	1	0.0	0	0	0	0
1	57	0	3	120	354	0	0	163	1	0.6	0	0	0	0
2	41	0	1	105	198	0	0	168	0	0.0	0	1	0	0
3	45	1	3	115	260	0	2	185	0	0.0	0	0	0	0
4	52	1	0	118	186	0	2	190	0	0.0	1	0	1	0
...
292	62	0	3	138	294	1	0	106	0	1.9	1	3	0	1
293	47	1	2	130	253	0	0	179	0	0.0	0	0	0	0
294	51	1	2	100	222	0	0	143	1	1.2	1	0	0	0
295	54	1	3	124	266	0	2	109	1	2.2	1	1	2	1
296	67	0	2	115	564	0	2	160	0	1.6	1	0	2	0

297 rows x 14 columns

Gambar 4.1 *Import* paket untuk membaca data yang digunakan di *Jupyter Notebook*

Pada pembacaan data digunakan *numpy* dan *pandas* yang berfungsi untuk pengolahan data. Selain itu, digunakan *pd.read_csv* dikarenakan *dataset* yang digunakan disimpan di folder komputer dengan nama *file* “data.csv” yang

berbentuk csv. Selanjutnya dipisahkan data *training* dan data *testing* seperti pada Gambar 4.2.

```
X_train = heart.loc[1:236, heart.columns != 'condition']
X_test = heart.loc[237:297, heart.columns != 'condition']
y_train = heart.loc[1:236, heart.columns == 'condition']
y_test = heart.loc[237:297, heart.columns == 'condition']
```

Gambar 4.2 Pemisahan data *training* dan data *testing*

Setelah dilakukan pemisahan data *training* dan data *testing*, dimulai pengoperasian metode *Decision Tree* menggunakan bantuan modul *tree* dengan formula ditunjukkan pada Gambar 4.3.

```
# Create Decision Tree classifier object
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(criterion='entropy',
                             splitter='best',
                             max_depth=None,
                             min_samples_split=2,
                             min_samples_leaf=1,
                             random_state=None,
                             max_leaf_nodes=None)

clf = clf.fit(X_train,y_train)
```

Gambar 4.3 Membuat objek klasifikasi *decision tree*

Pada bagian membuat objek klasifikasi *Decision Tree*, ada beberapa parameter yang digunakan untuk membangun *Decision Tree*. Penjelasan parameter yang digunakan ditunjukkan pada Tabel 4.1.

Tabel 4.1 Parameter objek klasifikasi *Decision Tree*

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

Setelah dimasukkan formula untuk *decision tree classifier*, akan di dapatkan nilai *y* prediksi atau prediksi hasil dari pemodelan oleh formula *decision tree classifier*. Nilai *y* prediksi didapatkan dengan cara pada Gambar 4.4.

```
#Predict the response for test dataset
y_pred = clf.predict(X_test)

y_pred
array([0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1,
       0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0,
       1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0])
```

Gambar 4.4 Menentukan *y* prediksi

Selanjutnya nilai *y* prediksi yang telah didapatkan dibandingkan dengan *y testing* untuk mendapatkan berapa persen akurasi ketepatan klasifikasi menggunakan *Decision Tree* yang ditunjukkan pada Tabel 4.2



Tabel 4.2 Perbandingan hasil y prediksi dan y test *Decision Tree*

No	Y_pred	Y_test	No.	Y_pred	Y_test	No.	Y_pred	Y_test
1	0	0	21	0	0	41	0	1
2	1	1	22	1	1	42	0	0
3	1	0	23	0	0	43	1	1
4	0	0	24	0	0	44	0	1
5	0	0	25	0	0	45	1	0
6	1	0	26	1	0	46	0	0
7	0	1	27	1	1	47	1	1
8	1	0	28	0	1	48	0	0
9	0	0	29	1	0	49	1	0
10	1	1	30	1	1	50	0	0
11	0	0	31	0	0	51	1	1
12	0	1	32	1	0	52	1	1
13	0	1	33	1	1	53	0	0
14	0	0	34	0	1	54	1	1
15	1	1	35	1	1	55	0	0
16	1	1	36	1	1	56	1	1
17	0	0	37	0	0	57	0	0
18	1	1	38	0	0	58	0	0
19	0	0	39	0	0	59	1	1
20	0	0	40	1	1	60	0	0

4.1.2. Ketepatan Klasifikasi

Tahap terakhir adalah menghitung akurasi dari model yang telah digunakan, yang ditunjukkan dengan *Confusion Matrix* pada Tabel 4.3.

Tabel 4.3 *Confusion Matrix* metode *Decision Tree*

Actual Label	Predicted Label	
	0	1
0	26	8
1	7	19

Dari Tabel 4.3 dapat dihitung nilai akurasi sebagai berikut :

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP+TN}{\text{Total}} \times 100\% \\
 &= \frac{26+19}{60} \times 100\% \\
 &= 75\%
 \end{aligned}$$

4.2 Klasifikasi Data menggunakan Regresi Logistik

Dalam pengerjaan klasifikasi data menggunakan regresi logistik, dilakukan 5 tahap pengerjaan. Tahap pertama penaksiran parameter, uji serentak, uji parsial, interpretasi koefisien dan akurasi model.

4.2.1 Pengubahan *dummies* data kategorik

Menurut Sharma dan Garavaglia (1998) dalam model regresi logistik, semua variabel independen dikodekan sebagai variabel *dummy* memberikan kemudahan interpretasi dan kalkulasi dari *odds ratio*, dan meningkatkan stabilitas dan signifikansi dari koefisien. Oleh karena itu, sebelum melakukan pengerjaan, data yang berbentuk kategorik terlebih dahulu diubah ke bentuk *dummies* agar dapat lebih mudah diklasifikasikan. Pengubahan ke bentuk *dummies* dilakukan seperti pada Gambar 4.5.

```
train['cp'].value_counts(dropna=False)
train['restecg'].value_counts(dropna=False)
train['slope'].value_counts(dropna=False)
train['thal'].value_counts(dropna=False)
```

```
0    128
2     90
1     17
Name: thal, dtype: int64
```

```
train = pd.get_dummies(heart[1:236], columns=['cp', 'restecg', 'slope', 'thal'], drop_first=True)
train
```

	age	sex	trestbps	chol	fbs	thalach	exang	oldpeak	ca	condition	cp_1	cp_2	cp_3	restecg_1	restecg_2	slope_1
1	57	0	120	354	0	163	1	0.6	0	0	0	0	1	0	0	0
2	41	0	105	198	0	168	0	0.0	1	0	1	0	0	0	0	0
3	45	1	115	260	0	185	0	0.0	0	0	0	0	1	0	1	0
4	52	1	118	186	0	190	0	0.0	0	0	0	0	0	0	1	1
5	58	0	170	225	1	146	1	2.8	2	1	0	0	1	0	1	1
...

Gambar 4.5 Mengubah *cp*, *restecg*, *slope*, dan *thal* ke bentuk *dummy*

Seperti yang ditunjukkan pada Gambar 4.5, nilai *cp* diubah menjadi *cp_1*, *cp_2* dan *cp_3*, sedangkan *cp_0* dihilangkan karena tidak perlu untuk pemisahan kelas dari *cp*. Penjelasan angka dari *cp* sebagai berikut :

- Ketika *cp*=0, *cp_1*, *cp_2*, dan *cp_3* = 0
- Ketika *cp*=1, *cp_1* = 1, *cp_2* dan *cp_3* = 0
- Ketika *cp*=2, *cp_2* = 1, *cp_1* dan *cp_3* = 0
- Ketika *cp*=3, *cp_3* = 1, *cp_1* dan *cp_2* = 0

Penjelasan untuk variabel *dummy* dari *restecg*, *slope*, dan *thal* sama dengan penjabaran variabel dari *cp*.

4.2.2 Penaksiran Parameter

Tahap awal yang dilakukan untuk pembentukan model Regresi Logistik adalah dengan melakukan penaksiran parameter model. Untuk pengerjaan ini, peneliti menggunakan bantuan modul *statsmodels* pada perangkat lunak python. *Statsmodels* adalah modul *Python* yang menyediakan kelas dan fungsi untuk estimasi banyak model statistik yang berbeda, serta untuk melakukan uji statistik, dan eksplorasi data statistik (Seabold, Skipper, & Perktold, 2010). 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 nilai signifikansi sebesar 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. Tahap penaksiran parameter dikerjakan menggunakan Python dengan formula logit sebagai berikut :

```
smf.logit(dep_var ~ ind_var1 + ind_var2 + ... + ind_varn, data =
df).fit(),
```

keterangan :

- dep_var = variabel dependen,
- ind_var_n = variabel independen; $n=1,2,\dots,n$

Pada formula di atas, smf.logit digunakan untuk memanggil “logit” yang terdapat pada paket *statsmodels.formula*. Pada tanda kurung setelah smf.logit, sebelah kiri tanda “~” menjelaskan variabel dependen (Y) sedangkan sebelah kanan tanda “~” menjelaskan variabel independen (X_1, X_2, \dots, X_n). Setelah dilakukan pengodingan menggunakan formula logit. didapatkan penaksiran parameter yang ditunjukkan pada Tabel 4.4.

Tabel 4.4 Tabel penaksiran parameter awal

	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 4.4, dapat dilihat variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* mempunyai nilai signifikansi kurang dari 0,05, artinya variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* berpengaruh. Sementara itu, untuk variabel lainnya mempunyai nilai signifikansi lebih besar dari 0,05. Dengan demikian, variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* akan di uji menggunakan cara yang sama namun mengurangi variabel yang tidak berpengaruh. Hasil penaksiran parameter setelah pengurangan variabel ditunjukkan pada Tabel 4.5.

Tabel 4.5 Penaksiran parameter setelah pengurangan 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 4.5, ditunjukkan nilai signifikansi variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2* berturut-turut sebesar 0,018, 0,000, 0,001, 0,000, dan 0,000. Nilai signifikansi masing-masing variabel tersebut kurang dari 0,05 maka variabel tersebut mempunyai pengaruh secara signifikan. Dengan demikian,



variabel-variabel tersebut dapat dimasukkan dalam model Regresi Logistik sebagai berikut:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_2 X_2 + \beta_{3.3} X_{3.3} + \beta_{11.1} X_{11.1} + \beta_{12} X_{12} + \beta_{13.2} X_{13.2})}{1 + \exp(\beta_0 + \beta_2 X_2 + \beta_{3.3} X_{3.3} + \beta_{11.1} X_{11.1} + \beta_{12} X_{12} + \beta_{13.2} X_{13.2})}$$

$$\pi(x) = \frac{\exp(-3,7913 + 1,0832X_2 + 1,8258X_{3.3} + 1,3560X_{11.1} + 1,3378X_{12} + 1,7130X_{13.2})}{1 + \exp(-3,7913 + 1,0832X_2 + 1,8258X_{3.3} + 1,3560X_{11.1} + 1,3378X_{12} + 1,7130X_{13.2})}$$

4.2.3 Uji Serentak

Uji yang digunakan untuk menguji signifikansi model secara serentak menggunakan uji *Likelihood Ratio* yang diperoleh dengan cara membandingkan fungsi *Log Likelihood* menggunakan seluruh variabel prediktor dengan fungsi *Log Likelihood* tanpa variabel prediktor. Pada penelitian menggunakan *Python* ini, *Log Likelihood* tanpa variabel prediktor dituliskan dengan *Log-Likelihood* memiliki nilai sebesar -88,311 sedangkan *Log Likelihood* menggunakan variabel prediktor dituliskan dengan *LL-Null* memiliki nilai sebesar -162,53. Berdasarkan Hosmer dan Lemeshow (2002), statistik uji yang digunakan untuk uji rasio *likelihood* adalah sebagai berikut

$$G = (-2 \ln(l_0)) - (-2 \ln(l_1)).$$

$$G = -88,311 - (-162,53)$$

$$G = 74,219$$

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 $G > \chi^2_{(0,05, 5)}$ dengan nilai $74,219 > 11,070$. Hal ini menunjukkan H_0 ditolak pada tingkat signifikansi $\alpha = 0,05$ berarti bahwa paling tidak ada satu variabel prediktor yang memiliki kontribusi yang signifikan terhadap variabel respon. Dengan kata lain bahwa uji *Likelihood Ratio* menyatakan variabel *sex*, *cp*, *chol*, *slope*, *ca*, dan *thal* secara serentak mempunyai pengaruh terhadap *condition*.

4.2.4 Uji Parsial

Setelah melakukan uji serentak maka langkah selanjutnya akan dilakukan pengujian signifikansi untuk masing-masing parameter dalam model dengan cara mengkuadratkan 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 pada penelitian ini ditunjukkan pada Tabel 4.6

Tabel 4.6 Uji *Wald* Variabel Regresi Logistik

	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 4.6 menjelaskan bahwa parameter yang signifikan merupakan koefisien dari variabel *sex*, *cp_3*, *slope_1*, *ca*, dan *thal_2*, dikarenakan variabel-variabel yang mempunyai nilai $W > \chi^2_{(0,05, 1)} =$. 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 *condition*.

4.2.5 . Odds Ratio

Odds ratio menunjukkan besarnya pengaruh masing-masing variabel prediktor yang signifikan. *Odds ratio* dapat juga diartikan sebagai jumlah relatif dimana 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. Nilai *odds ratio* masing-masing variabel ditunjukkan pada Tabel 4.7.

Tabel 4.7 Odds Ratio Variabel

Koefisien	OR
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:

1. Sex (X_2)

Jenis kelamin laki laki 2,954165 kali lebih besar kemungkinan mengalami penyakit jantung dibandingkan jenis kelamin perempuan.

2. Cp_3 (X_{3_3})

Tipe nyeri dada yang gejalanya biasa dan mudah di prediksi memiliki kemungkinan mengalami penyakit jantung 6,207985 kali lebih besar dari tipe nyeri yang memiliki gejala di luar penyakit jantung.

3. Slope_1 (X_{11_1})

Upsloping ST berpotensi mengalami penyakit jantung 3,880811 kali lebih tinggi dari slope *flat*.

4. Ca (X_{12})

Semakin tinggi 1 nilai Ca pada satu pasien, kecenderungan untuk mengalami penyakit jantung meningkat sebesar 3,810838 kali dari 1 nilai Ca di bawahnya.

5. Thal_2 (X_{13_2})

Penyakit *thalassemia* tipe normal berpotensi mengalami penyakit jantung 5,545758 kali lebih tinggi dari tipe *fixed defect*.

4.2.6 Ketepatan Klasifikasi

Untuk mengetahui tingkat keakurasian dari metode ini, dapat digunakan *Confusion Matrix* yang ditunjukkan pada Tabel 4.8.

Tabel 4.8 *Confusion Matrix* Regresi Logistik

<i>Actual Label</i>	<i>Predicted Label</i>	
	0	1
0	29	5
1	3	23

Dari Tabel 4.8 dapat dihitung nilai akurasi sebagai berikut :

$$\begin{aligned}
 \text{Akurasi} &= \frac{TP+TN}{\text{Total}} \times 100\% \\
 &= \frac{29+23}{60} \times 100\% \\
 &= 87\%
 \end{aligned}$$

Klasifikasi data dengan *dataset UCI Heart Disease* menggunakan metode *Decision Tree* dan Regresi Logistik menunjukkan nilai persentase akurasi untuk Regresi Logistik lebih besar dibandingkan *Decision Tree*. Data yang digunakan dapat berpengaruh terhadap hasil perbandingan.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan penelitian yang telah dilakukan, hasil klasifikasi data pada kasus *dataset UCI Heart Disease* menunjukkan

1. implementasi menggunakan *Decision Tree* menunjukkan hasil persentase akurasi ketepatan sebesar 75% ,
2. implementasi menggunakan Regresi Logistik dengan 5 variabel independen yang digunakan menghasilkan persentase akurasi sebesar 87%, dan
3. persentase akurasi dari Regresi Logistik lebih baik dibandingkan *Decision Tree*.

5.2 Saran

Pada penelitian ini metode klasifikasi yang digunakan hanya 2 metode dari banyaknya metode yang ada. Untuk perbandingan yang lebih baik, disarankan menggunakan metode klasifikasi data seperti *Naïve Bayes*, KNN, atau lainnya. Selain itu untuk penggunaan metode yang sama disarankan mencoba menggunakan bantuan lain selain python atau mencari model (paket) yang dapat mengklasifikasi data kategorik dengan metode *Decision Tree* Algoritma C4.5

DAFTAR PUSTAKA

- Ahmed A.M, Rizaner A, Ulusoy AH. 2018. A Decision Tree Algorithm Combined with Linear. *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*. doi : 10.1109/ICCCEEE.2018.8515759
- Anggryani IR, Kusumawati ED, & Kawulur EI. 2022. Metode Regresi Logistik Biner dan Metode K-Nearest Neighbor pada Klasifikasi Menopause Dini Wanita Distrik Oransbari Provinsi Papua Barat. *Seminar Nasional Matematika, Geometri, Statistika, dan Komputasi*.
- Babic SH, Kokol P, Podgorelec V, ZormanM, Sprogar M, Stiglic MM. 2000. The Art of Building Decision Trees. *J. Med. Syst. Vol. 24, No. 1*, 43–52. doi : <https://doi.org/10.1023/A:1005437213215>
- Cherngs. 2019. *Heart Disease Cleveland* UCI. Kaggle [internet]. [diakses 20 Januari 2022]. Tersedia dari :<https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci>.
- Garavaglia S, Sharma A. 1998. A smart guide to dummy variabels: Four applications and a macro. In *Proceedings of the northeast SAS users group conference* (Vol. 43)
- Ginting VS, Kusriani K, Taufiq E. 2020. Implementasi Algoritma C4. 5 untuk Memprediksi Keterlambatan Pembayaran Sumbangan Pembangunan Pendidikan Sekolah Menggunakan Python. *Inspiration: Jurnal Teknologi Informasi dan Komunikasi*. 10(1): 36-44.
- Herrhyanto, N. 2003. *Statistika Matematis Lanjutan*. Bandung: CV Pustaka Setia.
- Hosmer DW, Lemeshow S. 2000. *Applied Logistic Regression Second Edition*. New York: John Wiley & Sons, Inc.
- Larose DT, & Larose CD. 2014. *Discovering Knowledge in Data : an Introduction to Data Mining, Second Edition*. New Jersey: John Wiley & Sons, Inc.
- Merawati D, & Rino R. 2019. Penerapan Data Mining Penentu Minat Dan Bakat Siswa Smk Dengan Metode C4.5. *Algor*. 1(1): 28-37
- Mustika, Ardilla Y, Manuhutu A, Ahmad N, Hasbi I, Manuhutu MA, Ridwan M, Hozairi, Wardhani AK, Alim S, *et al*. 2021. *Data Mining dan Aplikasinya*. Bandung: Widiana Bhakti Persada
- Quinlan JR. 1993. *C4.5: Programs for Machine Learning*. London: Morgan Kaufmann Publisher, Inc.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

- Rizki F, Widodo DA, Wulandari SP. 2015. Faktor Risiko Penyakit Anemia Gizi Besi pada Ibu Hamil di Jawa Timur Menggunakan Analisis Regresi Logistik. *Jurnal Sains dan Seni ITS*. 4(2): 2337-3520.
- Safitri A, Sudarmin, 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):1-6.
- Santosa B. 2007. *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis (edisi 1)*. Yogyakarta: Graha Ilmu.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* 2011. Scikit Learn : Machine Learning in Python. *Journal of Machine Learning Research*. 12: 2825-2830
- Seabold, Skipper, Perktold J. 2010. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*.
- Sokolova M, Lapalme G. 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45(4): 427-437. doi: <https://doi.org/10.1016/j.ipm.2009.03.002>
- Ye N. 2014. *Data Mining: Theories, Algorithms, and Examples (1st ed.)*. CRC Press. <https://doi.org/10.1201/b15288>
- Zainal MM, Rindengan AJ, & Weku WC. 2014. Penggunaan Association Rule Data Mining Untuk Menentukan Pola Lama Studi Mahasiswa F-MIPA UNSRAT. *Jurnal de Cartesian*. 3(1): 1-8.
- Zhu F, Tang M, Xie L, Zhu H. 2018. A Classification Algorithm of CART Decision Tree based on MapReduce Attribute Weights. *International Journal of Performability Engineering*.



RIWAYAT HIDUP

Penulis dilahirkan di kota Jakarta pada 22 Juni 2000 sebagai anak kedua dari pasangan bapak Pandji Setiawan dan ibu Nurlaela. Pendidikan sekolah menengah atas (SMA) ditempuh di sekolah SMAN 3 Bekasi dan lulus pada tahun 2017/2018. Pada tahun 2018, penulis diterima sebagai mahasiswa program sarjana (S-1) di Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam di IPB.

Selama mengikuti program S-1, penulis aktif menjadi Staf Divisi Humas Passion 3.0 2019, Staf Divisi LO Pesta Sains Nasional 2019, Sekretaris dan Bendahara Divisi Medis *Math League* 2019, Staf Divisi Acara G-Familiarity 2019, Staf Divisi *Ticketing* dan *Marketing Mathematics Challenge* 2020, Sekretaris Biro Bisnis dan Kemitraan Gugus Mahasiswa Matematika Kabinet *Golden Ratio* pada tahun periode 2019/2020, Sekretaris Divisi Danus Matematika Ria IPB 2020, Staf Divisi Penanggung Jawab Kelompok E-Math 2020, Sekretaris umum E-Math 2021, dan Sekretaris umum 2 Gugus Mahasiswa Matematika Kabinet Sigma Karya pada tahun periode 2020/2021.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.