

## Data Mining Penyakit Terbanyak Berdasarkan *Decision Tree* Algoritma C4.5 Di RSUD Pandan Arang Boyolali

Hafiddhudin Al Mubarak<sup>1</sup>, Trismianto Asmo Sutrisno<sup>2</sup>, Sri Sugiarsi<sup>3</sup>

<sup>1,2,3</sup>Program Studi D4 Manajemen Informasi dan Kesehatan, STIKes Mitra Husada Karanganyar, Indonesia

### Article Info

#### Article history:

Received Oct 02, 2023

Revised Nov 14, 2024

Accepted Feb 22, 2025

#### Keywords:

Data Mining

Most Diseases

Decision Tree

Algorithm C4.5

RapidMiner

### ABSTRACT

Top 10 disease data is data on diseases that often appear and occur most often. Data collection of the top 10 diseases is used to determine disease patterns in the community. Decision tree classification method is a model that maps observations of an item so that a conclusion is obtained about the target value of an item described in the form of a tree model. The C4.5 algorithm is method for creating a decision tree based on the training data that has been provided. The purpose of this research is to result in the classification of the most diseases based on the C4.5 decision tree algorithm of Pandan Arang Boyolali Hospital using excel and RapidMiner. This research design is descriptive observational. The data collection technique in this study was observation, with a total net data of 3449 data. The results of this research, there are no differences in results between excel and RapidMiner which results in the classification of the most diseases, which are blocks A00-B99, Z00-Z99, I00-I99, O00-O99, J00-J99, and E00-E90, and 18 rules are obtained. Evaluation of the performance of the C4.5 algorithm decision tree classification model using confusion matrix produces an accuracy value of 25.14%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Hafiddhudin Al Mubarak,  
Program Studi D4 Manajemen Informasi Kesehatan,  
STIKes Mitra Husada Karanganyar,  
Papahan, Tasikmadu, Karanganyar, Jawa Tengah.  
Email: hafiddhudin247@gmail.com

## 1. PENDAHULUAN

Berdasarkan Permenkes 24 tahun 2022, rekam medis adalah dokumen yang berisikan data identitas pasien, pemeriksaan, pengobatan, tindakan, dan pelayanan lain yang telah diberikan kepada pasien. Salah satu pengolahan informasi rekam medis yaitu pelaporan yang termasuk pelaporan internal fasilitas pelayanan kesehatan, dan pelaporan eksternal yang salah satunya ke kementerian kesehatan melalui SIRS. Permenkes No. 1171/MenKes/Per/VI/2011, 10 besar penyakit merupakan data yang harus dilaporkan pada SIRS Rev VI. Data 10 besar penyakit adalah data penyakit yang sering muncul dan paling sering terjadi. Pendataan 10 besar penyakit digunakan untuk mengetahui pola penyakit yang ada di masyarakat.

Data mining adalah penggalian dan analisis, dengan menggunakan peranti otomatis atau semi otomatis, dari sejumlah besar data yang bertujuan untuk menemukan pola yang memiliki arti [1]. Salah satu kegunaan data mining adalah klasifikasi, yaitu menggeneralisasi struktur yang diketahui untuk diaplikasikan pada data-data baru [2]. Metode klasifikasi *decision tree* atau pohon keputusan merupakan salah satu model yang memetakan observasi-observasi suatu item sehingga diperoleh suatu kesimpulan tentang nilai target dari suatu item yang digambarkan dalam bentuk pohon model. Algoritma C4.5 adalah salah satu metode untuk membuat *decision tree* berdasarkan *training* data yang telah disediakan [3].

Dalam Penelitiannya [4] dilakukan klasifikasi dan analisa pola penyebaran penyakit dengan metode algoritma machine learning klasifikasi C4.5. Proses akhir klasifikasi dilakukan dengan menggunakan perangkat lunak WEKA. Kesimpulan dari penelitian ini yaitu kode penyakit yang didominasi pada tahun 2014 dan 2015 adalah kode G, P, J dengan kelompok usia dewasa tua sampai paruh baya.

Penelitian yang dilakukan [5] menggunakan teknologi data mining yang berkaitan dengan pengklasifikasian diagnosa penyakit, yaitu algoritma C4.5. Pengklasifikasian dilakukan dalam bentuk website. Kesimpulan dari penelitian ini yaitu pohon keputusannya berbentuk rule pada sistem yang telah dibuat.

Berdasarkan studi pendahuluan terhadap data kunjungan pasien yang berobat pada bulan Desember 2022, didapatkan penyakit terbanyak yaitu J00-J99, Z00-Z99, S00-T98, O00-O99, E00-E90, I00-I99. Sedangkan klasifikasi 10 besar penyakit di pelaporan SIMRS rumah sakit yaitu N00-N19, Z00-Z99, K00-K93, U00-U99, A00-B99, O00-O99, I00-I99, R00-R99, dan P00-P96. Tujuan penelitian ini untuk mengetahui bagaimanakah klasifikasi penyakit terbanyak berdasarkan *decision tree* Algoritma C4.5 di RSUD Pandan Arang Boyolali.

## 2. METODE PENELITIAN

Jenis penelitian yang digunakan adalah penelitian observasional deskriptif, yaitu penelitian yang menggambarkan klasifikasi penyakit terbanyak berdasarkan *decision tree* algoritma C4.5 menggunakan excel dan *RapidMiner*. Penelitian ini dilakukan di RSUD Pandan Arang Boyolali pada tahun 2023 dengan mengambil data triwulan 1 yang terdiri dari 3581 data rekam medis rawat inap. Ada 4 atribut yang digunakan dalam penelitian ini, yaitu usia, jenis kelamin, bulan, dan kode diagnosis. Kemudian dilakukan klasifikasi dan transformasi data 4 atribut tersebut. Pertama usia, klasifikasi usia dibagi menjadi tiga kelompok yaitu bayi dan anak-anak (<15 tahun), muda dan dewasa (> 15 tahun dan < 50 tahun), kemudian tua (> 50 tahun). Kedua, jenis kelamin dibagi menjadi laki-laki dan perempuan. Ketiga, bulan dibagi dalam triwulan I. Keempat, kode diagnosis yang dibagi menjadi 22 blok, yaitu A00-B99, C00-D48, D50-D89, E00-E90, F00-F99, G00-G99, H00-H59, H60-H95, I00-I99, J00-J99, K00-K93, L00-L99, M00-M99, N00-N99, O00-O99, P00-P96, Q00-Q99, R00-R99, S00-T98, V01-Y98, Z00-Z99, U00-U99 [6].

## 3. HASIL DAN PEMBAHASAN

### 3.1 Hasil Penelitian

#### Mencari *Entropy* Total dan *Gain* Tertinggi

Sebelum dilakukan pengklasifikasian menggunakan *RapidMiner*, dilakukan perhitungan manual menggunakan excel. Perhitungan dimulai dengan mencari *entropy* total dari seluruh atribut.

Entropy (total keputusan):

$$\begin{aligned} & \left(-\frac{441}{3099} + \text{Log}_2\left(\frac{441}{3099}\right)\right) + \left(-\frac{141}{3099} + \text{Log}_2\left(\frac{141}{3099}\right)\right) + \left(-\frac{225}{3099} + \text{Log}_2\left(\frac{225}{3099}\right)\right) + \left(-\frac{256}{3099} + \text{Log}_2\left(\frac{256}{3099}\right)\right) + \left(-\frac{31}{3099} + \text{Log}_2\left(\frac{31}{3099}\right)\right) + \\ & \left(-\frac{62}{3099} + \text{Log}_2\left(\frac{62}{3099}\right)\right) + \left(-\frac{7}{3099} + \text{Log}_2\left(\frac{7}{3099}\right)\right) + \left(-\frac{50}{3099} + \text{Log}_2\left(\frac{50}{3099}\right)\right) + \left(-\frac{212}{3099} + \text{Log}_2\left(\frac{212}{3099}\right)\right) + \left(-\frac{202}{3099} + \text{Log}_2\left(\frac{202}{3099}\right)\right) + \\ & \left(-\frac{282}{3099} + \text{Log}_2\left(\frac{282}{3099}\right)\right) + \left(-\frac{22}{3099} + \text{Log}_2\left(\frac{22}{3099}\right)\right) + \left(-\frac{45}{3099} + \text{Log}_2\left(\frac{45}{3099}\right)\right) + \left(-\frac{202}{3099} + \text{Log}_2\left(\frac{202}{3099}\right)\right) + \left(-\frac{285}{3099} + \text{Log}_2\left(\frac{285}{3099}\right)\right) + \\ & \left(-\frac{80}{3099} + \text{Log}_2\left(\frac{80}{3099}\right)\right) + \left(-\frac{5}{3099} + \text{Log}_2\left(\frac{5}{3099}\right)\right) + \left(-\frac{72}{3099} + \text{Log}_2\left(\frac{72}{3099}\right)\right) + \left(-\frac{199}{3099} + \text{Log}_2\left(\frac{199}{3099}\right)\right) + \left(-\frac{2}{3099} + \text{Log}_2\left(\frac{2}{3099}\right)\right) + \\ & \left(-\frac{14}{3099} + \text{Log}_2\left(\frac{14}{3099}\right)\right) + \left(-\frac{181}{3099} + \text{Log}_2\left(\frac{181}{3099}\right)\right) \end{aligned}$$

$$\text{Entropy (total)} = 3,888209639$$

Hasil dari perhitungan nilai entropy total tersebut, akan digunakan untuk menghitung nilai *gain* terhadap atribut lainnya. Setelah mendapat nilai *entropy* total, dilanjutkan menghitung masing-masing nilai *entropy* setiap atribut.

Setelah menghitung *entropy* setiap atribut seperti sebelumnya, kemudian menghitung *gain* seperti berikut ini:

$$\text{Gain (total, bulan)} =$$

$$3,88821 - \left( \left( \frac{972}{3099} \times 0 \right) + \left( \frac{966}{3099} \times 3,90726 \right) + \left( \frac{1161}{3099} \times 3,88833 \right) \right)$$

$$\text{Gain (total, bulan)} = 1,21355$$

Kemudian menghitung semua *gain* setiap atribut, dan menentukan *gain* tertinggi. *Gain* tertinggi ini akan dijadikan sebagai *root* dalam *decision tree*.

**Tabel 1. Perhitungan Gain Tertinggi**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Total	3099	441	...	161	3,88821	
Bulan						1,21355
Januari	972	141	...	55	0	
Februari	966	118	...	54	3,90726	
Maret	1161	182	...	52	3,88833	
Usia						3,26671
Bayi dan Anak-anak	616	129	...	140	3,12668	
Muda dan Dewasa	1236	190	...	12	0	
Tua	1247	122	...	9	0	
Jenis Kelamin						0,08417
Laki-laki	1381	237	...	82	3,76672	
Perempuan	1718	204	...	79	3,83404	

Berdasarkan tabel 1 didapatkan *gain* tertinggi pada atribut usia, sehingga atribut usia akan menjadi *root*, atribut usia terbagi menjadi tiga, sehingga perlu dihitung lebih lanjut.

### Mencari Node 1.1-Node 1.1.2

Perhitungan dilakukan untuk mencari *node* 1.1 cabang dari atribut usia bayi dan anak-anak, kemudian dicari *gain* tertinggi.

**Tabel 2. Perhitungan Node 1.1**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Bayi dan Anak-anak	616	129	...	140	3,12667	
Bulan						3,12667
Januari	180	37	...	46	0	
Februari	192	37	...	44	0	
Maret	244	55	...	50	0	
Jenis Kelamin						3,12667
Laki-laki	341	75	...	74	0	
Perempuan	275	54	...	66	0	

Berdasarkan tabel 2, didapatkan semua *gain* mendapatkan nilai yang sama, sehingga untuk perhitungan selanjutnya bisa menggunakan atribut manapun. Perhitungan selanjutnya yaitu menggunakan atribut Jenis kelamin yang terbagi menjadi dua.

Perhitungan dilakukan untuk mencari *node* 1.1.1 cabang dari atribut jenis kelamin laki-laki usia bayi dan anak-anak, kemudian dicari *gain* tertinggi.

**Tabel 3. Perhitungan Node 1.1.1**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Bayi dan Anak-anak, Laki laki	341	75	...	74	0	
Bulan			...			0
Januari	100	23	...	26	0	
Februari	113	24	...	25	0	
Maret	128	28	...	23	0	

Perhitungan dilakukan untuk mencari *node* 1.1.2 cabang dari atribut jenis kelamin perempuan usia bayi dan anak-anak, kemudian dicari *gain* tertinggi.

**Tabel 4. Perhitungan Node 1.1.2**

	Jumlah kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Bayi dan Anak-anak, Perempuan	275	54	...	66	0	
Bulan			...			0
Januari	80	14	...	20	0	
Februari	79	13	...	19	0	
Maret	116	27	...	27	0	

**Mencari Node 1.2-Node 1.2.2**

Perhitungan dilakukan untuk mencari *node* 1.2 cabang dari atribut usia muda dan dewasa, kemudian dicari *gain* tertinggi.

**Tabel 5. Perhitungan Node 1.2**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Muda dan Dewasa	1236	190	...	12	0	
Bulan						0
Januari	409	64	...	5	0	
Februari	387	53	...	6	0	
Maret	440	73	...	1	0	
Jenis Kelamin						0
Laki-laki	407	93	...	5	0	
Perempuan	829	97	...	7	0	

Berdasarkan tabel 5 didapatkan semua *gain* telah mencapai batas minimum yaitu 0, sehingga untuk perhitungan selanjutnya bisa menggunakan atribut manapun. Perhitungan selanjutnya yaitu menggunakan atribut Jenis kelamin yang terbagi menjadi dua.

Perhitungan dilakukan untuk mencari *node* 1.2.1 cabang dari atribut jenis kelamin laki-laki usia muda dan dewasa, kemudian dicari *gain* tertinggi.

**Tabel 6. Perhitungan Node 1.2.1**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Muda dan Dewasa, Laki laki	407	93	...	5	0	
Bulan			...			0
Januari	141	33	...	2	0	
Februari	116	24	...	2	0	
Maret	150	36	...	1	0	

Perhitungan dilakukan untuk mencari *node* 1.2.2 cabang dari atribut jenis kelamin perempuan usia muda dan dewasa, kemudian dicari *gain* tertinggi

**Tabel 7. Perhitungan Node 1.2.2**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Muda dan Dewasa, Perempuan	829	97	...	7	0	
Bulan						0
Januari	268	31	...	3	0	
Februari	271	29	...	4	0	
Maret	290	37	...	0	0	

### Mencari Node 1.3-Node 1.3.2

Perhitungan dilakukan untuk mencari *node* 1.3 cabang dari atribut usia tua, kemudian dicari *gain* tertinggi.

**Tabel 8. Perhitungan Node 1.3**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Tua	1247	122	...	9	0	
Bulan			...			0
Januari	383	40	...	4	0	
Februari	387	28	...	4	0	
Maret	477	54	...	1	0	
Jenis Kelamin						0
Laki-laki	633	69	...	3	0	
Perempuan	614	53	...	6	0	

Berdasarkan tabel 8 didapatkan semua *gain* telah mencapai batas minimum yaitu 0, sehingga untuk perhitungan selanjutnya bisa menggunakan atribut manapun. Perhitungan selanjutnya yaitu menggunakan atribut Jenis kelamin yang terbagi menjadi dua.

Perhitungan dilakukan untuk mencari *node* 1.3.1 cabang dari atribut jenis kelamin laki-laki usia tua, kemudian dicari *gain* tertinggi.

**Tabel 9. Perhitungan Node 1.3.1**

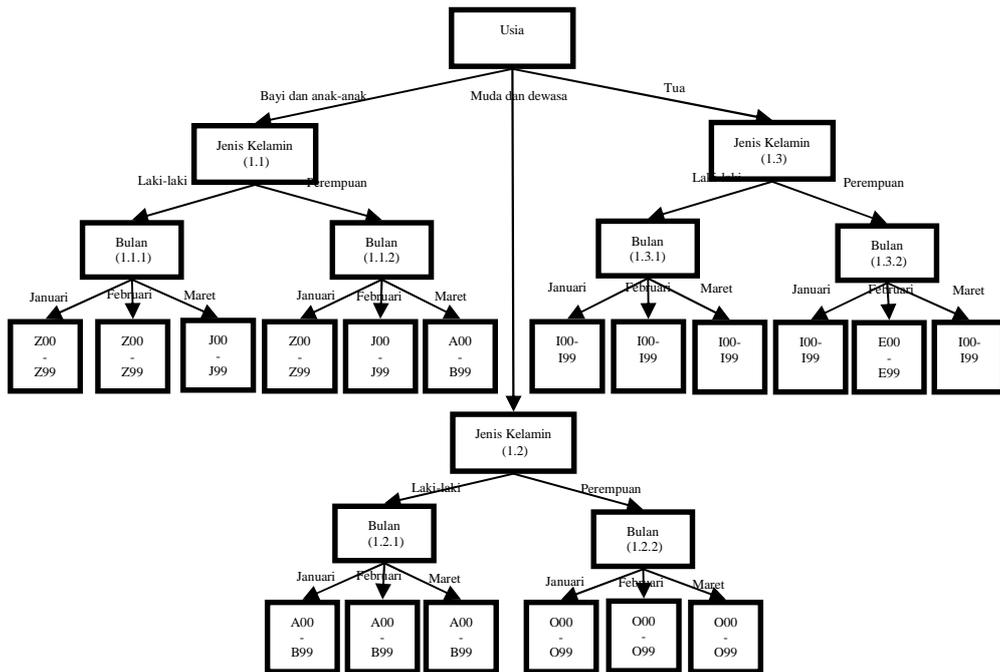
	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Tua, Laki laki	633	69	...	3	0	
Bulan						0
Januari	195	28	...	1	0	
Februari	196	16	...	2	0	
Maret	242	25	...	0	0	

Perhitungan dilakukan untuk mencari *node* 1.3.2 cabang dari atribut jenis kelamin perempuan usia tua, kemudian dicari *gain* tertinggi.

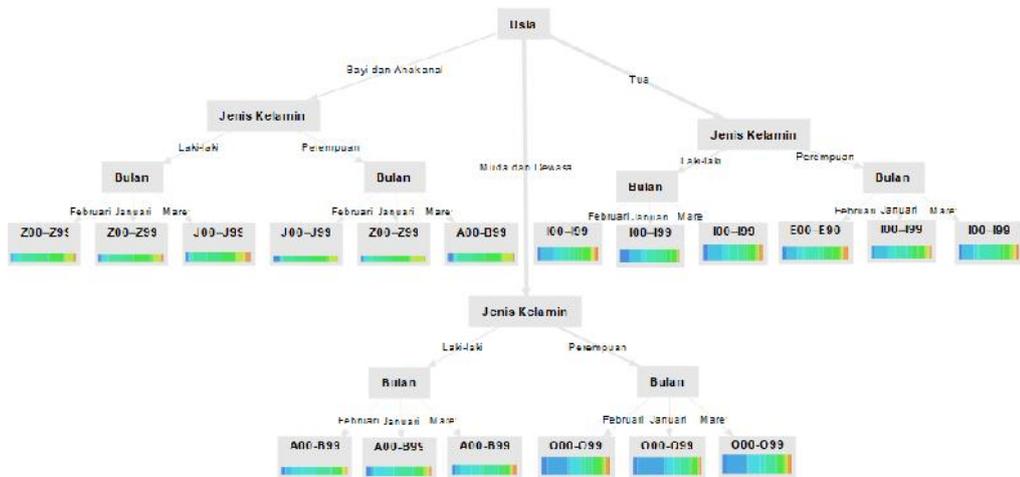
**Tabel 10. Perhitungan Node 1.3.2**

	Jumlah Kasus	A00-B99	...	Z00-Z99	Entropy	Gain
Tua, Perempuan	614	53	...	6	0	
Bulan						0
Januari	188	12	...	3	0	
Februari	191	12	...	2	0	
Maret	235	29	...	1	0	

Dari perhitungan di atas, maka dapat dibuat *decision tree*, seperti berikut ini:



Gambar 1. Hasil *Decision Tree* Secara Manual



Gambar 2. Hasil *Decision Tree* Menggunakan *RapidMiner*

Berdasarkan *decision tree* yang telah terbentuk dari perhitungan *entropy* dan *gain* diperoleh *rule*. *Rule-rule* tersebut yaitu:

**Tabel 11. Rule Klasifikasi**

Rule 1	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Laki-laki Bulan = Januari	MAKA	Kode Diagnosis = Z00-Z99
Rule 2	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Laki-laki Bulan = Februari	MAKA	Kode Diagnosis = Z00-Z99
Rule 3	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Laki-laki Bulan = Maret	MAKA	Kode Diagnosis = J00-J99
Rule 4	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Perempuan Bulan = Januari	MAKA	Kode Diagnosis = Z00-Z99
Rule 5	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Perempuan Bulan = Februari	MAKA	Kode Diagnosis = J00-J99
Rule 6	JIKA	Usia = Bayi dan anak-anak Jenis Kelamin = Perempuan Bulan = Maret	MAKA	Kode Diagnosis = A00-B99
Rule 7	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Laki-laki Bulan = Januari	MAKA	Kode Diagnosis = A00-B99
Rule 8	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Laki-laki Bulan = Februari	MAKA	Kode Diagnosis = A00-B99
Rule 9	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Laki-laki Bulan = Maret	MAKA	Kode Diagnosis = A00-B99
Rule 10	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Perempuan Bulan = Januari	MAKA	Kode Diagnosis = O00-O99
Rule 11	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Perempuan Bulan = Februari	MAKA	Kode Diagnosis = O00-O99
Rule 12	JIKA	Usia = Muda dan dewasa Jenis Kelamin = Perempuan Bulan = Maret	MAKA	Kode Diagnosis = O00-O99
Rule 13	JIKA	Usia = Tua Jenis Kelamin = Laki-laki Bulan = Januari	MAKA	Kode Diagnosis = I00-I99
Rule 14	JIKA	Usia = Tua Jenis Kelamin = Laki-laki Bulan = Februari	MAKA	Kode Diagnosis = I00-I99
Rule 15	JIKA	Usia = Tua Jenis Kelamin = Laki-laki Bulan = Maret	MAKA	Kode Diagnosis = I00-I99
Rule 16	JIKA	Usia = Tua Jenis Kelamin = Perempuan Bulan = Januari	MAKA	Kode Diagnosis = I00-I99
Rule 17	JIKA	Usia = Tua Jenis Kelamin = Perempuan Bulan = Februari	MAKA	Kode Diagnosis = E00-E90
Rule 18	JIKA	Usia = Tua Jenis Kelamin = Perempuan Bulan = Maret	MAKA	Kode Diagnosis = I00-I99

### Analisis Klasifikasi Penyakit Terbanyak Berdasarkan Decision Tree Algoritma C4.5

Penyakit terbanyak dari hasil decision tree algoritma C4.5 yaitu blok A00-B99, Z00-Z99, I00-I99, O00-O99, J00-J99, dan E00-E90. Kemudian untuk mengukur kinerja dari model klasifikasi penyakit terbanyak berdasarkan algoritma decision tree dilakukan uji confusion matrix. Hasil dari uji confusion matrix menunjukkan akurasi nya sebesar 25.14%.

**Tabel 12. Hasil Uji Confusion Matrix**

Accuracy: 25.14%									
	true I00	I99	true K00-K93	...	true V01	Y98	true U00	U99	class precision
pred. I00	I99	26	12	...	0		0		22.22%
pred. K00-K93		0	0	...	0		0		0.00%
...		...	...	...	...		...		...
pred. V01	Y98	0	0	...	0		0		0.00%
pred. U00	U99	0	0	...	0		0		0.00%
class recall		72.22%	0.00%	...	0.00%		0.00%		

### 3.2 Pembahasan

Dari hasil perhitungan menggunakan excel dan *RapidMiner* diperoleh 18 *rule*, yang sesuai dengan penelitian [7] yang telah memperoleh 12 *rule*. Dari *rule-rule* tersebut dapat diketahui penyakit terbanyak berdasarkan usia, jenis kelamin, dan bulan.

Jika terdapat sebuah kasus yaitu, apa penyakit yang sering terjadi pada usia tua, berjenis kelamin perempuan, di bulan Maret. Berdasarkan penerapan *rule* 18, JIKA usia “tua”, jenis kelamin “perempuan”, bulan “Maret” MAKA I00-I99. Jadi penyakit terbanyak berdasarkan contoh kasus tersebut adalah I00-I99 yaitu Penyakit pada sistem peredaran darah.

Dari *decision tree* dan *rule* yang terbentuk, Z00-Z99 (*factors influencing health status and contact with health services*) menjadi penyakit terbanyak pada usia bayi dan anak-anak, baik jenis kelamin laki-laki maupun perempuan. Kategori-kategori Z00-Z99 disediakan untuk suatu kejadian apabila lingkungan kejadian terkait bukan sakit/penyakit, cedera atau sebab luar yang terklasifikasi pada kategori-kategori A00-Y89 namun terekam sebagai “*diagnoses*” atau “*problems*” (masalah kesehatan). Dalam hal ini Z00-Z99 digunakan untuk mengkode kelahiran bayi di rumah sakit, yang berarti banyak bayi lahir di RSUD Pandan Arang Boyolali.

Selanjutnya J00-J99 (*diseases of the respiratory system*) juga menjadi penyakit terbanyak pada usia bayi dan anak-anak, baik jenis kelamin laki-laki maupun perempuan. Hal ini sesuai dengan Badan kesehatan dunia *World Health Organization* (WHO) menyatakan pneumonia merupakan penyebab utama kematian balita di dunia dan penelitian [8] yang menyatakan pneumonia merupakan pembunuh utama anak dibawah usia lima tahun (balita) di dunia, dan variabel berhubungan sebab akibat dengan kejadian pneumonia pada anak balita adalah faktor pendidikan, riwayat pemberian ASI, riwayat asma, kepadatan rumah dan ventilasi rumah.

A00-B99 (*certain infectious and parasitic diseases*) menjadi penyakit terbanyak pada usia bayi dan anak-anak, berjenis kelamin perempuan. Menurut profil kesehatan indonesia tahun 2021, penyakit infeksi masih merupakan penyebab kematian terbanyak pada masa post neonatal, diare masih menjadi penyebab kematian terbanyak kedua setelah pneumonia pada masa post neonatal. Menurut [9] faktor risiko kejadian diare pada anak yaitu pengetahuan ibu, status gizi, pemberian ASI-Eksklusif, dan pengolahan makanan dan minuman.

Pada usia muda dan dewasa berjenis kelamin laki-laki, penyakit terbanyaknya yaitu A00-B99 (*certain infectious and parasitic diseases*). Dalam penelitian [10] distribusi gastroenteritis akut paling banyak ditemukan pada kelompok umur dewasa (41-60 tahun), dan kejadian penyakit gastroenteritis akut antara jenis kelamin laki-laki dan perempuan sebenarnya tidak ada perbedaan, baik antara laki-laki dan perempuan memiliki peluang yang sama mengalami penyakit gastroenteritis akut, walaupun memiliki peluang yang sama tapi dari penelitian [10] distribusi penyakit gastroenteritis akut lebih tinggi pada perempuan hanya sebesar 9,8% dari jenis kelamin laki-laki.

Sedangkan pada usia muda dan dewasa berjenis kelamin perempuan penyakit terbanyaknya yaitu O00-O99 (*pregnancy, childbirth and the puerperium*). Bab ini digunakan untuk ibu hamil, bersalin dan nifas. Hal ini juga berkaitan dengan penyakit terbanyak pada usia bayi dan anak-anak, yaitu semakin banyak ibu hamil yang datang ke rumah sakit, semakin banyak pula bayi yang lahir dirumah sakit.

I00-I99 (*diseases of the circulatory system*) menjadi penyakit terbanyak untuk usia tua berjenis kelamin laki-laki maupun perempuan. Hal ini sesuai dengan [11] kasus hipertensi lumayan banyak dijumpai terutama pada lansia. Angka kematian karena hipertensi di Indonesia sangat tinggi. Hipertensi merupakan penyebab kematian nomor 3 setelah stroke dan tuberkulosis. Terdapat beberapa faktor yang menyebabkan tingginya kasus hipertensi beserta komplikasi yang ditimbulkannya. Faktor-faktor tersebut diantaranya adalah gejala-gejala hipertensi sangat bervariasi dimulai dengan tanpa gejala, sakit kepala ringan ataupun gejala lain yang hampir sama dengan penyakit lainnya.

E00-E90 (*endocrine, nutritional and metabolic diseases*), merupakan penyakit terbanyak pada usia tua berjenis kelamin perempuan. Hal ini sesuai dengan penelitiannya [12] yaitu diabetes mellitus sering muncul setelah seseorang memasuki rentang usia rawan yaitu setelah usia 45 tahun, juga menunjukkan bahwa responden perempuan memiliki risiko untuk terkena DM tipe 2 sebesar 2,777 kali lebih besar dibandingkan dengan responden laki-laki.

Tidak ada perbedaan dari hasil *decision tree* yang terbentuk dari perhitungan *excel* maupun *RapidMiner*. Hal ini sesuai dengan penelitiannya [13] yaitu menggunakan *software* data mining *KNIME* dihasilkan *decision tree* dan aturan yang sesuai dan sama dengan pengujian secara manual. Menggunakan *decision tree* Algoritma C4.5 berhasil mengklasifikasikan 6 kode diagnosis terbanyak dari 22 kode. Hal ini juga sesuai dengan penelitian [14] yang berhasil mengelompokkan 13 jenis kategori dari 21 jenis kategori.

Penyakit yang paling banyak berdasarkan *decision tree* algoritma C4.5 yaitu A00-B99, Z00-Z99, I00-I99, O00-O99, J00-J99, dan E00-E90. Sedangkan penyakit terbanyak berdasarkan SIMRS yaitu N00-N99, A00-B99, Z00-Z99, K00-K93, I00-I99. Terdapat perbedaan penyakit terbanyak yang dihasilkan *decision tree* algoritma C4.5, yaitu N00-N99, O00-O99, J00-J99, K00-K93, dan E00-E99. O00-O99 diderita oleh usia muda dan dewasa dengan jenis kelamin perempuan pada bulan Januari, Februari, dan Maret. J00-J99 diderita oleh usia bayi dan anak-anak dengan jenis kelamin perempuan pada bulan Februari. E00-E99 diderita oleh usia tua dengan jenis kelamin perempuan pada bulan Februari.

Perbedaan tersebut bisa terjadi beberapa faktor, yaitu:

- a. Untuk menghitung penyakit terbanyak berdasarkan *decision tree* algoritma C4.5 hanya menggunakan diagnosis utama, diagnosis sekunder tidak digunakan dalam perhitungan. Data dengan kode utama E10.9 dan kode sekunder N18.5, kode sekunder tidak digunakan dalam perhitungan sehingga jumlah kode diagnosis N18.5 akan berkurang. Begitu juga dengan kode serupa lainnya.
- b. Dengan dilakukan pengelompokan penyakit menjadi 22 blok, diagnosis penyakit akan di kelompokkan ke masing-masing blok yang sesuai. Blok J00-J99 terdiri dari J00, J02,

J02.9, J03.9, dan Blok K00-K93 terdiri dari K01.1, K04.0, maka jumlah diagnosis Blok J00-099 akan lebih banyak.

- c. Penyakit terbanyak yang dihasilkan *decision tree* terbagi menjadi beberapa atribut seperti usia dan jenis kelamin, sehingga penyakit yang di dominasi oleh atribut tertentu akan diklasifikasikan menjadi penyakit terbanyak. Seperti penyakit O00-O99 hanya diderita oleh perempuan usia muda dan dewasa.

Berdasarkan tabel menunjukkan ke akurasian sebesar 25.14%. Hal ini sesuai dengan penelitian [15] yaitu menghitung keakuratan model yang mana akan menampilkan *Accuracy* yang didapat dari *confusion matrix*. Meskipun tidak ada nilai pasti yang digunakan untuk menentukan rendah atau tidak, tetapi secara umum, akurasi yang dianggap rendah yaitu dibawah 70%-80%. Rendahnya akurasi bisa dikarenakan kelas tujuannya terlalu banyak, dan variasi data yang tinggi. Hal ini sesuai dengan penelitiannya [16].

#### 4. KESIMPULAN

Klasifikasi penyakit terbanyak berdasarkan *decision tree* Algoritma C4.5 menggunakan excel yaitu blok A00-B99, Z00-Z99, I00-I99, O00-O99, J00-J99, dan E00-E90, dengan 18 rule. Klasifikasi penyakit terbanyak berdasarkan *decision tree* Algoritma C4.5 menggunakan *RapidMiner* yaitu blok A00-B99, Z00-Z99, I00-I99, O00-O99, J00-J99, dan E00-E90, dengan 18 rule. Tidak ada perbedaan dari hasil *decision tree* yang terbentuk dari perhitungan *excel* maupun *RapidMiner* dan hasil uji *confusion matrix* menunjukkan akurasi 25.14%.

#### REFERENCES

- [1] M. Lailil, D. E. Ratnawati, and R. R. Mp, *Data Mining*. Malang: UB Press, 2018.
- [2] D. Suyanto, *Data Mining untuk klasifikasi dan klasterisasi data*, 1st ed. Bandung: Informatika Bandung, 2017.
- [3] M. Arhami and M. Nasir, *Data Mining-Algoritma dan Implementasi*, 1st ed. Yogyakarta: Andi, 2020.
- [4] S. M. Turnip and P. Silitonga, “Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5,” *Jurnal Teknik Informatika Unika Santo Thomas*, vol. 3, no. 1, pp. 1–5, 2018.
- [5] Allwine, M. Sakban, and B. G. R. F. Sitepu, “Pengklasifikasian Diagnosa Penyakit Pasien Pengguna BPJS Kesehatan Menggunakan Algoritma C4.5,” *Jurnal Bisantara Informatika*, vol. 6, no. 1, 2022.
- [6] WHO, *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. 2010.
- [7] R. Rafiska, S. Defit, and G. W. Nurcahyo, “Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4.5,” *Jurnal Rekayasa Sistem dan Teknologi Informasi*, vol. 2, no. 1, pp. 391–396, 2018, [Online]. Available: <http://jurnal.iaii.or.id>
- [8] atul Husna, F. Dewi Pertiwi, and A. Saputra Nasution, “Faktor-Faktor Yang Berhubungan Dengan Kejadian Pneumonia Pada Balita Di Puskesmas Semplak Kota Bogor 2020,” *PROMOTOR Jurnal Mahasiswa Kesehatan Masyarakat*, vol. 5, no. 3, p. 273, 2022, [Online]. Available: <http://ejournal.uika-bogor.ac.id/index.php/PROMOTOR>
- [9] V. Sagara, B. Aramico, and V. N. Arifin, “Faktor Risiko Kejadian Diare Pada Anak Usia 1-4 Tahun Di Wilayah Kerja Puskesmas Batoh Kota Banda Aceh Tahun 2022,” *Jurnal Kesehatan Tambusai*, vol. 4, no. 2, pp. 2635–2640, 2023.

[10] W. A. Saputra, K. Mariadi, and G. Somayana, “Karakteristik Penyakit Gastroenteritis Pada Pasien di RSUP Sanglah Denpasar Tahun 2018,” *Jurnal Medika Udayana*, vol. 10, no. 4, 2021.

[11] N. Maulana, “Pencegahan Dan Penanganan Hipertensi Pada Lansia” *Jurnal Peduli Masyarakat*, vol. 4, no. 1, pp. 163–168, 2022, [Online]. Available: <http://jurnal.globalhealthsciencegroup.com/index.php/JPM>

[12] F. Milita, S. Handayani, and B. Setiaji, “Kejadian Diabetes Mellitus Tipe II pada Lanjut Usia di Indonesia (Analisis Riskesdas 2018),” *Jurnal Kedokteran Dan Kesehatan*, vol. 17, no. 1, pp. 9–20, 2021, [Online]. Available: <https://jurnal.umj.ac.id/index.php/JKK>

[13] Y. Mardi, “Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5,” *Jurnal Sains dan Informatika*, vol. 4, no. 1, pp. 40–53, Apr. 2018, doi: 10.22216/jsi.v4i1.3077.

[14] Y. A. Fiandra, S. Defit, and Yuhandri, “Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-10),” *Jurnal Rekayasa Sistem dan Teknologi Informasi*, vol. 2, no. 1, pp. 82–89, 2017, [Online]. Available: <http://jurnal.iaii.or.id>

[15] I. M. Agus Oka Gunawan, I. D. A. Indah Saraswati, I. D. G. Riswana Agung, and I. P. Eka Putra, “Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4.5 Dengan Rapidminer,” *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 2, pp. 73–83, Apr. 2023, doi: 10.47233/jteksis.v5i2.775.

[16] B. Aulifia Riski Putra Wahyu, A. Fayi Farozi, C. Putra Mahendra, and R. Kembang Hapsari, “Klasifikasi Penderita Penyakit Diabetes Berdasarkan Decision Tree Menggunakan Algoritma C4.5,” *INTEGER: Journal of Information Technology*, vol. 8, no. 1, 2023.

**BIOGRAPHIES OF AUTHORS**

	<p><b>Hafiddhudin Al Mubarak</b>, Program Studi D4 Manajemen Informasi dan Kesehatan, STIKes Mitra Husada Karanganyar</p>
	<p><b>Trismianto Asmo Sutrisno, ST, M.Kom</b>, Gelar S1 Teknik Informatika diperoleh di Universitas Gunadarma Jakarta, Magister Sistem Informasi Universitas Diponegoro Semarang. Berprofesi sebagai Dosen di Sekolah Tinggi Ilmu Kesehatan Mitra Husada Karanganyar dengan bidang keahlian Pemrograman Web, Algoritma Pemrograman, Basis Data dan Analisis Perancangan Sistem.</p>



**Sri Sugiarsi, SKM, M.Kes,** Gelar S2 Kesehatan Masyarakat diperoleh di Universitas Airlangga, dan Gelar S2 Kesehatan Masyarakat diperoleh di Universitas Diponegoro. Berprofesi sebagai Dosen di Sekolah Tinggi Ilmu Kesehatan Mitra Husada Karanganyar dengan bidang keahlian Metodologi Penelitian, Epidemiologi, dan Biostatistik.