e-ISSN 2540-7902 dan p-ISSN 2541-366X

Implementasi Metode C4.5 dalam Mendiagnosa Penyakit Pernapasan

1st *Naimah Gairil Massora
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
naimahgairilmassora98@gmail.com

2nd Dirgahayu Lantara
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
dirgahayu.lantara@umi.ac.id

3rd Wistiani Astuti*
Universitas Muslim Indonesia
Fakultas Ilmu Komputer
Makassar, Indonesia
wistiani.astuti@umi.ac.id

***Abstrak*—Salah satu organ terpenting pada manusia adalah saluran pernapasan. Jika organ ini mengalami gangguan maka akan menyebabkan manusia susah dalam melakukan kegiatannya. Ada anggapan bahwa penyakit yang diawali dengan gejala batuk, nyeri dada, sesak nafas tidak membahayakan. Ini disebabkan karena tidak mengetahui apakah gejala tersebut bisa menjadi gejala awal dari suatu penyakit yang parah. Oleh karena itu, dari permasalahan tersebut, dilakukan penelitian dengan memanfaatkan perkembangan teknologi yang bertujuan untuk membantu seseorang dalam mendiagnosa suatu penyakit yang diawali dari gejala utama penyakit pernapasan tanpa harus melakukan pemeriksaan kesehatan. Penelitian ini mengimplementasikan metode C4.5. Dengan memanfaatkan data sebanyak 23 dan 19 gejala. Metode C4.5 bisa menghasilkan nilai akurasi yang sangat tinggi tergantung seberapa kompleks data yang digunakan. Hasil yang diperoleh dari penelitian ini berupa pohon keputusan yang kemudian diuji dan memperoleh tingkat akurasi sebesar 14.29%.**

Kata Kunci*—*metode C4.5;diagnosa; pohon keputusan; penyakit pernapasan; sesak nafas

I. PENDAHULUAN

Diagnosa merupakan langkah awal yang harus dilakukan untuk menangani suatu penyakit. Diagnosa memiliki prinsip bahwa suatu penyakit dapat dikenali dengan memperhatikan gejala-gejala apa yang dirasakan oleh pasien yang ditimbulkan oleh penyakit tersebut. Dalam proses diagnosa dibutuhkan data-data seperti hasil pemeriksaan laboratorium. Diagnosa penyakit pernapasan dapat dilakukan dengan memperhatikan gejala utamanya. dari gejala utama tersebut, seorang pakar dapat mengajukan pertanyaan-pertanyaan yang berkaitan dengan gejala utamanya untuk mengetahui hasil akhir dari diagnosa [1].

Salah satu penyakit yang umum terjadi pada semua kategori umur yaitu penyakit pada saluran pernapasan. Saluran pernapasan merupakan seperangkat organ tubuh yang dimulai dari hidung kemudian alveoli bersama dengan organ adneksa [2]. Penyakit pernafasan merupakan penyakit yang menyerang pada sistem respirasi. Gejala utama dari penyakit pernafasan adalah batuk, sesak nafas, nyeri dada.

Sesak nafas merupakan suatu keadaan dimana seseorang merasa seperti kekurangan udara atau tidak leluasa menghirup udara sehingga frekuensi nafasnya menjadi cepat sehingga muncul rasa sesak di dada.

Manusia akan sulit untuk melakukan kegiatan sehari-harinya apablia sistem pernapasannya terganggu[3]. Jenis penyakit yang bisa menyerang sistem pernapasan sangat bermacam-macam tergantung dari gejala-gejala yang dialami oleh penderita [4].

Saat ini teknologi sudah berkembang dengan pesat diseluruh bidang, salah satunya di bidang kesehatan. Informasi yang cepat dan akurat bisa diperoleh melalui teknologi. Setiap tahunnya berbagai bidang dalam lingkungan kesehatan dapat menghasilkan data dalam jumlah yang besar. Namun, data yang diperoleh kebanyakan tidak bisa memberikan informasi secara cepat sehingga diperlukan cara yang efektif untuk mengelolah data yang banyak agar dapat memberikan pengetahun yang baru.[3] *Data mining* merupakan cara yang bisa digunakan dalam proses menemukan pengetahuan.

Klasifikasi dan prediksi banyak digunakan dalam *data mining* untuk menganalisis suatu data atau memprediksi data. Klasifikasi merupakan metode yang termasuk kepada pendekatan deskriptif. Dalam *data mining* proses klasifikasi terdiri dari dua tahap, tahap pertama yaitu menganalisis kumpulan data *training* dengan menggunakan algoritma klasifikasi atau dikenal dengan tahap pembelajaran. Tahap kedua yaitu penggunaan model untuk klasifikasi dan kumpulan data testing digunakan untuk memperkirakan keakuratan dalam aturan klasifikasi. Model pengklasifikasian disajikan dengan aturan klasifikasi atau menemukan pola.

Pada penelitiaan ini akan mengimplementasikan salah satu metode dalam *data mining* yaitu metode C4.5 untuk mendiagnosa suatu penyakit yang diawali dari gejala utama penyakit pernapasan.

II. METODOLOGI

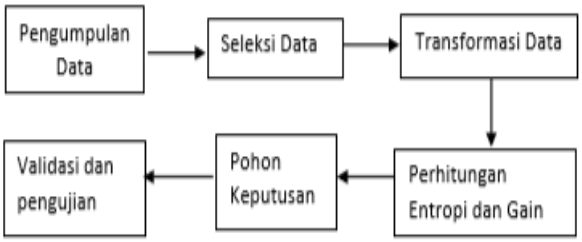
A. *Metode C4.5*

Ekstensi algoritma ID3 mengacu pada metode C4.5 [5]. Untuk pengolahan data numerik dan diskret algoritma C4.5 merupakan salah satu alternatif algoritma yang dapat digunakan[6]. Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk pohon keputusan[3]. Ide dasar dari algoritma ini adalah pembuatan pohon keputusan

*Corresponding Author

berdasarkan pemilihan atribut yang memiliki prioritas tertinggi atau memiliki nilai gain tertinggi berdasarkan nilai entropi atribut. Pada tahapannya algoritma C4.5 memiliki 2 prinsip kerja, yaitu: Membuat pohon keputusan, dan membuat aturan-aturan (rule model) [7]. Aturan aturan yang terbentuk dari pohon keputusan akan membentuk suatu kondisi dalam bentuk if then. Algoritma C4.5 secara rekursif mengunjungi setiap simpul keputusan, memilih pembagian yang optimal, sampai tidak bisa dibagi lagi.

Metode penelitian yang digunakan dalam penerapan algoritma C4.5 untuk diagnosa penyakit pernapasan, seperti pada Gambar 1.



Gambar. 1.Flowchart rancangan proses penelitian

1) Pengumpulan Data

Pengumpulan data merupakan proses untuk mengumpulkan data yang akan digunakan dalam proses diagnosa menggunakan metode C4.5.

2) Seleksi Data

Seleksi data adalah memilah data yang akan digunakan yang bertujuan untuk menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada sampel data dimana penemuan akan dilakukan [9].

3) Transformasi Data

Transformasi data adalah proses mengubah data ke dalam bentuk yang sesuai agar dapat di proses dengan perhitungan algoritma C4.5 [7].

4) Perhitungan Entropi dan Gain

Entropi adalah ukuran dari teori informasi yang dapat mengetahui karakteristik yang bermacam-macam dari kumpulan data. Dari nilai Entropi tersebut kemudian dihitung nilai gain masing-masing atribut. Gain adalah informasi yang didapatkan dari perubahan entropi pada suatu kumpulan data, baik melalui observasi atau bisa juga dengan cara melakukan partisipasi terhadap suatu set data. Nilai gain tertinggi yang akan dijadikan sebagai simpul akar pada pembuatan pohon keputusan.

5) Pohon Keputusan

Pohon keputusan merupakan hasil dari proses perhitungan entropi dan gain yang dilakukan secara berulang-ulang sampai semua atribut pohon memiliki kelas dan tidak bisa dilakukan perhitungan lagi.

6) Validasi dan pengujian

Dilakukan untuk mengetahui semua fungsi apakah bekerja dengan baik atau tidak. Validasi dan pengujian dilakukan untuk mengetahui tingkat akurasi, presisi, dan *recall* dari hasil prediksi klasifikasi. Akurasi adalah presentase dari catatatan yang diklasifikasikan dengan benar dalam pengujian dataset. Presisi adalah persentase data yang diklasifikasikan sebagai model baik yang sebenarnya juga baik. Recall adalah pengukuran tingkat pengenalan positif sebenarnya [10]. Untuk validasi dan pengujian menggunakan *tools rapidminer. Rapid miner* merupakan salah satu software untuk pengolahan data mining yang bersifat open source.

B. Pohon Keputusan

Pohon keputusan adalah sebuah struktur data yang terdiri dari *node* dan *edge* . Proses dari pohon keputusan ini dimulai dari *node* akar hingga ke *node* daun yang dilakukan secara rekursif dimana setiap percabangan menyatakan kondisi dan setiap ujung pohon akan menyatakan keputusan. Simpul pada pohon keputusan dibedakan menjadi tiga yaitu akar simpul, simpul percabangan, dan simpul akhir[8]. Pohon keputusan berfungsi untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variable input dengan sebuah variable target. Proses pada pohon keputusan adalah mengubah bentuk data(tabel) menjadi model pohon,mengubah bentuk pohon menjadi rule, dan menyederhanakan rule.

Pohon keputusan adalah sebuah struktur yang dapat digunakan untuk mengubah data menjadi pohon keputusan yang akan menghasilkan aturan-aturan keputusan besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Manfaat utama dalam penggunaan pohon keputusan adalah kemampuannya untuk memecah proses pengambilan keputusan yang kompleks menjadi lebih sederhana sehingga pengambilan keputusan akan lebih mudah. Adapun kelebihan dari pohon keputusan adalah :

- Pengambilan keputusan dari kasus kompleks dan global bisa diubah menjadi lebih sederhana dan spesifik
- Bisa menghilangkan perhitungan-perhitungan yang tidak diperlukan
- Pohon keputusan bersifat fleksibel sehingga bisa meningkatkan kualitas keputusan yang dihasilkan

Selain memiliki kelebihan, pohon keputusan juga memiliki kekurangan, diantaranya yaitu :

- Kesulitan dalam mendesain pohon keputusan yang optimal
- Terjadi overlap terutama pada saat kelas dan kriteria yang digunakan sangat banyak sehingga bisa memperlambat waktu pengambilan keputusan dan meningkatkan penggunaan memori
- Hasil kualitas keputusan yang diperoleh sangat tergantung pada desain pohon keputusan.

Adapun tahapan dari metode C4.5 [3] adalah :

1) Merancang data training

2) Menentukan akar dari pohon keputusan

Proses pencarian akar dilakukan dengan menghitung entropi masing-masing kategori. Rumus untuk menghitung entropi :

$$Entropi\ (S) = \sum_{i=1}^{n} pi\ log_2\ pi \quad (1)$$

Keterangan :

S : Himpunan (dataset)

n : banyaknya *record*

pi : probabilitas yang di dapat dari jumlah ya atau tidak dibagi keseluruhan total kasus

Setelah memperoleh nilai entropi masing-masing kategori, langkah selanjutnya adalah menghitung nilai gain. Nilai gain tertinggi yang dijadikan sebagai akar. Rumus untuk menghitung nilai gain :

$$Gain\ (S,A) = Entrophy\ (s) - \sum_{i=1}^{n} \frac{|si|}{|s|} * Entrophy\ (si) \quad (2)$$

Keterangan :

S : himpunan (dataset)

A : atribut yang akan di pakai

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

3) Pembentukan cabang

4) Ulangi langkah ke-2 hingga semua record terpartisi.

Proses partisi akan berhenti saat:

- Semua record dalam node N mendapat kelas yang sama
- Tidak ada atribut didalam record yang dipartisi lagi sampai tidak ada record di dalam cabang yang kosong.

III. HASIL DAN PEMBAHASAN

A. *Pengumpulan Data*

Data yang digunakan dalam penelitian ini diperoleh dari buku catatan status pasien rawat inap Rumah Sakit Sardjito. Jumlah data yang akan digunakan adalah 23 data pasien dengan gejala utama sesak nafas, dan 19 gejala-gejala yang berkaitan dengan sesak nafas. Dari 23 data yang ada diperoleh 6 hasil diagnosa sementara yaitu: anemia berjumlah 4 diagnosa, asma berjumlah 3 diagnosa, bronkitis akut berjumlah 4 diagnosa, bronkitis kronik berjumlah 3 diagnosa dan TBC dengan 9 diagnosa, Tabel I.

TABEL I. DATA PASIEN DENGAN GEJALA UTAMA SESAK NAFAS

NO	E1	E2	E3	E4	…	E19	Diagnosa
1	ya	tidak	tidak	Ya	…	tidak	anemia
2	ya	tidak	tidak	tidak	…	tidak	anemia
3	ya	tidak	ya	Ya	…	tidak	anemia
4	ya	tidak	ya	Ya	…	tidak	anemia
5	Ya	ya	tidak	tidak	…	tidak	asma
6	Ya	Tidak	tidak	tidak	…	tidak	asma
7	tidak	Tidak	tidak	tidak	…	tidak	asma
8	Ya	Tidak	tidak	tidak	…	tidak	bronkitis akut
9	Ya	Tidak	tidak	ya	…	tidak	bronkitis akut
10	tidak	Tidak	tidak	tidak	…	tidak	bronkitis akut
11	tidak	Tidak	tidak	tidak	…	tidak	bronkitis akut
…	…	…	…	…	…	…	…
23	tidak	Ya	tidak	tidak		tidak	tbc

Berdasarkan Tabel I, akan dijadikan sebagai data uji untuk dijadikan sebagai acuan dalam melakukan perhitungan nilai entropi dan nilai gain.

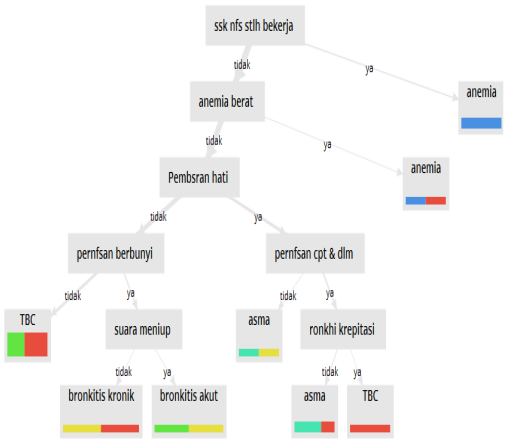
B. *Perhitungan Nilai Entropy dan Gain*

Dalam kasus yang terdapat pada Tabel I, akan dibuat pohon keputusan untuk diagnosa penyakit pernapasan. Proses pembentukan pohon keputusan diawali dengan melakukan perhitungan nilai entropi dan gain, Tabel II.

Berdasarkan kasus ini, dari hasil perhitungan, yang memperoleh nilai gain tertinggi adalah atribut E1 yaitu sebesar 2,174 sehingga bisa dijadikan sebagai *node*.

Hitung seluruh nilai entropi dan gain dari semua nilai atribut yang ada. Setelah semua nilai entropi dan gain dihitung ambil nilai gain tertinggi untuk dijadikan sebagai simpul akar. Hilangkan atribut yang dipilih sebelumnya dan ulangi perhitungan nilai Entropi, Gain, dengan memilih Gain terbesar dan dijadikan simpul internal pohon. Ulangi perhitungan tersebut hingga semua atribut pohon memiliki kelas.

Dari hasil perhitungan entropi dan gain yang di dapat kemudian diubah ke dalam bentuk pohon keputusan, pada Gambar 2.



Gambar. 2.pohon keputusan diagnosa penyakit pernapasan

TABEL II. NILAI ENTROPI DAN GAIN SELURUH ATRIBUT

Total		jum.kasus	anemia	…	tbc	entropi	gain
		23	4		9	2,174	
E1							2,174
	ya	12	4	…	4	0	
	tidak	11	0	…	5	0	
E2							0,404726
	ya	5	0	…	4	0	
	tidak	18	4	…	5	2,260785	
E3							0,472188
	ya	5	2	…	2	0	
	tidak	18	2	…	7	2,174584	
E4							0,627483
	ya	6	3	…	2	0	
	tidak	17	1	…	7	2,092395	
E5							0,411891
	ya	5	1	…	3	0	
	tidak	18	3	…	6	2,251629	
E6							0,346228
	ya	4	1	…	2	0	
	tidak	19	3	…	7	2,21261	
E7		0					0,116927
	ya	1	0	…	0	0	
	tidak	22	4	…	9	2,150614	
E8							0,111336
	ya	8	1	…	2	2,25	
	tidak	15	3	…	7	1,962807	
E9							0,595611
	ya	5	0	…	1	0	
	tidak	18	4	…	8	2,016876	
E10							0
	ya	0	0	…	0	0	
	tidak	23	4	…	9	2,174036	
E11							0,328881
	ya	3	0	…	1	0	
	tidak	20	4	…	8	2,121928	
E12				…			0,415516
	ya	4	1		1	0	
	tidak	19	3	…	8	2,128735	
E13				…			0,415516
	ya	4	1		1	0	
	tidak	19	3	…	8	2,128735	
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
E19							
	ya	0	0	…		0	0
	tidak	23	4	…		9	2,174036

Atribut E1 sampai E19 mewakili gejala-gejala yang berkaitan dengan sesak nafas, dimana:

E1 : sesak nafas setelah bekerja akibat anemia

E2 : Kenaikan desakan venosa

E3 : ronki krepitas

E4 : Suara pernapasan bronchial

E5 : Demam

E6 : Sesak nafas berat

E7 : Nafas cuping hidung

E8 : Pernapasan bunyi

E9 : Suara meniup timbulnya akut

E10 : tanda-tanda shok setelah disuntik vaksin / penisilin

E11 : ada riwayat serangan nafas berbunyi

E12 : Suara pernapasan bronchial

E13 : perkusi redup

E14 : setengah sadar

E15 : dehidrasi

E16 : pernafasan cepat dan dalam

E17 : anemia berat

E18 : Pembesaran hati atau edema

E19 : shok ketakutan atas apa yang terjadi

Perhitungan entropi:

Entropi (Total) = $(-4/23) * \log_2 (4/23) + (-3/23) * \log_2 (3/23) + (-4/23) * \log_2 (4/23) + (3/23) * \log_2 (3/23) + (-9/23) * \log_2 (9/23)$

= 2,174036

Kunci pencarian entropi:

- Jika di antara kolom "Ya" atau "Tidak" ada yang bernilai 0 (nol) maka entropi-nya dipastikan juga bernilai 0 (nol)
- Jika kolom "Ya" dan "Tidak" mempunyai nilai yang sama maka entropi-nya dipastikan juga bernilai 1 (satu)

Hitung keseluruhan nilai entropi untuk masing-masing atribut kemudian hitung nilai gainnya. Perhitungan nilai gain :

Gain (Total, E1) = entropi total – ((jumlah data E1 bernilai ya / total kasus * entropi E1 bernilai Ya) + ((jumlah data E1 bernilai tidak / total kasus * entropi E1 bernilai tidak)

= 2,174036-((12/23*1,5775) + (11/23*0))

= 1,350987

Gunakan cara yang sama untuk menghitung nilai gain atribut lain. Apabila semua nilai gain telah diperoleh, maka ambil data dengan nilai gain tertinggi untuk dijadikan sebagai *node* akar dan data acuan perhitungan selanjutnya.

C. Analisis Hasil pengujian

Analisis hasil dilakukan dengan melakukan perhitungan menggunakan *tools rapidminer* dan diperoleh tingkat akurasi data secara keseluruhan sebanyak 14,29%, seperti pada Gambar 3.

accuracy: 14.29%

	true anemia	true asma	true bronkitis akut	true bronkitis kro...	true TBC	class precision
pred. anemia	1	0	1	0	0	50.00%
pred. asma	0	0	0	0	0	0.00%
pred. bronkitis akut	0	0	0	1	3	0.00%
pred. bronkitis kr...	0	0	0	0	0	0.00%
pred. TBC	0	1	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	0.00%	0.00%	

Gambar. 3.Nilai akurasi, presisi, dan recall

Nilai akurasi yang diperoleh sangat rendah karena jumlah *data training* yang digunakan bukan dalam jumlah yang besar (data yang sedikit).

IV. KESIMPULAN

Dari hasil penelitian ini dapat disimpulkan bahwa algoritma C4.5 dapat digunakan untuk memudahkan dalam pengambilan keputusan dengan memproyeksikan data-data yang ada ke dalam bentuk pohon keputusan, berdasarkan nilai entropi dan *gain* yang dimiliki masing-masing atribut data. Penerapan metode C4.5 dapat diimplementasikan dalam proses mendiagnosa penyakit pernapasan dengan nilai akurasi sebesar 14.29%. Untuk hasil prediksi yang lebih akurat dibutuhkan data dalam jumlah besar, artinya semakin besar jumlah data yang digunakan maka semakin akurat hasil prediksi yang dihasilkan.

DAFTAR PUSTAKA

[1] Y. Diawali, D. Gejala, and U. Nyeri, "Proses representasi pengetahuan dalam rancangan sistem pakar pendiagnosa penyakit pernafasan yang diawali dari gejala utama nyeri dada," pp. 1–13.

[2] U. I. Indragiri, "Aplikasi Sistem Pakar Diagnosa Penyakit Pernapasan Menggunakan Metode Case-Based Reasoning," vol. 3, 2017.

[3] C. Algoritma, "Perancangan Data Mining untuk Klasifikasi Prediksi Penyakit ISPA dengan," vol. 3, no. 1, pp. 179–182, 2017.

[4] "No Title," no. Saad 2005, pp. 2005–2007, 2011.

[5] F. F. Harryanto, S. Hansun, U. M. Nusantara, G. Serpong, and C. Pegawai, "Penerapan Algoritma C4 . 5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT WISE," vol. 3, no. 2, pp. 95–103, 2017.

[6] "Algoritma c4.5 untuk simulasi prediksi kemenangan dalam pertandingan sepakbola," pp. 53–58.

[7] M. F. Arifin and D. Fitrianah, "Rekomendasi Penerimaan Mitra Penjualan Studi Kasus : PT Atria Artha Persada," no. January 2018.

[8] F. I. Komputer and U. D. Nuswantoro, "Penyakit Stroke Dengan Klasifikasi Data Mining Pada," 2011.

[9] C. Algoritma *et al.*, "Kredit (Studi Kasus Di Koperasi Pegawai Republik," vol. 1, no. 2, pp. 6–10, 2014.

[10] Y. Altujjar, W. Altamimi, I. Al-turaiki, and M. Al-razgan, "Predicting Critical Courses Affecting Students Performance : A Case Study," *Procedia - Procedia Comput. Sci.*, vol. 82, no. March, pp. 65–71, 2016.