ALGORITMA C4.5 UNTUK DIAGNOSA PENYAKIT TUBERKULOSIS

Amrin[1]; Irawan Satriadi[2]; Oki Rosanto[3]
Program Studi Teknologi Komputer, Fakultas Teknologi Informasi
Universitas Bina Sarana Informatika[1]
Program Studi Sistem Informasi Akuntansi, Fakultas Teknologi Informasi
Universitas Bina Sarana Informatika[2]
Program Studi Sistem Informasi, STIMIK Nusa Mandiri[3]
amrin.ain@bsi.ac.id[1] ; irawan.irs@bsi.ac.id[2]; oki.okr@nusamandiri.ac.id[3]

Abstract—*Tuberculosis is a contagious and deadly disease in the world, even the World Health Organization (WHO) has declared it a world emergency. Many symptoms can occur in someone who contracted pulmonary tuberculosis, and to analyze the symptoms is not an easy thing, it needs to do sputum tests to patients. In addition, it is also necessary a method that can facilitate when performing analysis and extract patient information from medical record data available. In this study, the author will apply the data mining classification method, namely the C4.5 Algorithm to diagnose tuberculosis. Based on the results of the performance measurement of the model using Cross Validation, Confusion Matrix and ROC Curve testing methods, it is known that C4.5 algorithms have an accuracy rate of 84,56% and area under the curva (AUC) value of 0,938. This shows that the resulting model including the classification category is very good because it has an AUC value between 0.90-1.00.*

Keywords: *Algorithm C4.5, confusion matrix, ROC curva, tuberculosis*

Intisari—Penyakit tuberkulosis merupakan penyakit menular dan mematikan di dunia, bahkan *World Health Organization* (WHO) mencanangkan sebagai penyakit kedaruratan dunia (*global emergency*). Banyak gejala yang bisa terjadi pada seseorang yang terjangkit tuberkulosis, dan untuk menganalisa gejala tersebut bukan hal yang mudah, perlu dilakukan tes dahak pada penderita. Selain itu, dibutuhkan juga sebuah metode yang dapat mempermudah saat melakukan analisa dan menggali informasi pasien dari data rekam medik yang tersedia. Pada penelitian ini, penulis akan menerapkan metode klasifikasi data mining, yaitu Algoritma C4.5, Tujuan dari penelitian ini adalah untuk mendiagnosa penyakit tuberculosis dengan menggunakan metode algoritma C4.5. Berdasarkan hasil pengukuran performa dari model tersebut dengan menggunakan metode pengujian *Cross Validation, Confusion Matrix* dan Kurva ROC, diketahui bahwa algoritma C4.5 memiliki tingkat akurasi sebesar 84,56% dan nilai area *under the curva* (AUC) sebesar 0,938. Hal ini menunjukkan bahwa model yang dihasilkan termasuk kategori klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

Kata Kunci: Algoritma C4.5, kurva ROC, matriks konfusi, tuberkulosis.

PENDAHULUAN

Tuberculosis yang disingkat TBC atau TB adalah penyakit menular yang disebabkan oleh bakteri mycobacterium tuberculosis yang ditularkan melalui udara (dropet nuclei) saat seorang pasien TBC batuk dan percikan ludah yang mengandung bakteri tersebut terhirup oleh orang lain saat bernapas (Widoyono, 2011). Penyakit TB (Orhan, Temurtas, & Tanrıkulu, 2010) adalah penyakit menular yang disebabkan oleh bakteri yang disebut mycobacterium tuberculosis dan merupakan penyebab kematian paling tinggi yang terjadi pada usia produktif 15-50 tahun, kelompok ekonomi lemah, dan berpendidikan rendah. Menurut kementerian kesehatan dalam (Amrin, 2018) penyakit tuberculosis dapat menular sehingga perlu penanganan yang intensif, setidaknya diperlukan pengobatan minimal 6 bulan secara rutin dan terus menerus. Sedangkan Indonesia menempati peringkat ke-2 di dunia setelah India dengan pasien TBC terbanyak dan diperkirakan ada 1.020.000 kasus TB di Indonesia.

Penularan tuberkulosis (TBC) sangat cepat melalui udara. Bagi penderita diharapkan selalu melakukan pemeriksaan dan pengobatan sampai tuntas. TBC ditularkan melalui udara. Percikan ludah atau dahak yang dikeluarkan menjadi media penularan yang sangat cepat di dunia ini. Penularan TBC melalui udara akan sangat rentan terjadi di ruang publik. Dari berbagai penelitian akan ada puluhan ribu kuman yang keluar dari batuk dan bersin. Oleh karenanya diharapkan masyarakat untuk menggunakan masker di tempat-tempat umum dan senantiasa

berperilaku hidup bersih dan sehat menurut Kemenkes dalam (Amrin, 2018).

Salah satu teknik yang dapat digunakan untuk melakukan prediksi penyakit adalah klasifikasi. Klasifikasi adalah salah satu teknik yang ada dalam data mining, membutuhkan pohon keputusan untuk membuat data dalam grup atau kelas (Adhatrao, Gaykar, Dhawan, Jha, & Honrao, 2013). Klasifikasi data penyakit TB pada medis merupakan tugas penting dalam memprediksi penyakit, bahkan dapat membantu dokter dalam mengambil keputusan diagnosis penyakit tersebut (Fine, 2012), dengan demikian sangat penting melakukan diagnosa secara dini agar dapat mengurangi penularan TB kepada masyarakat luas.

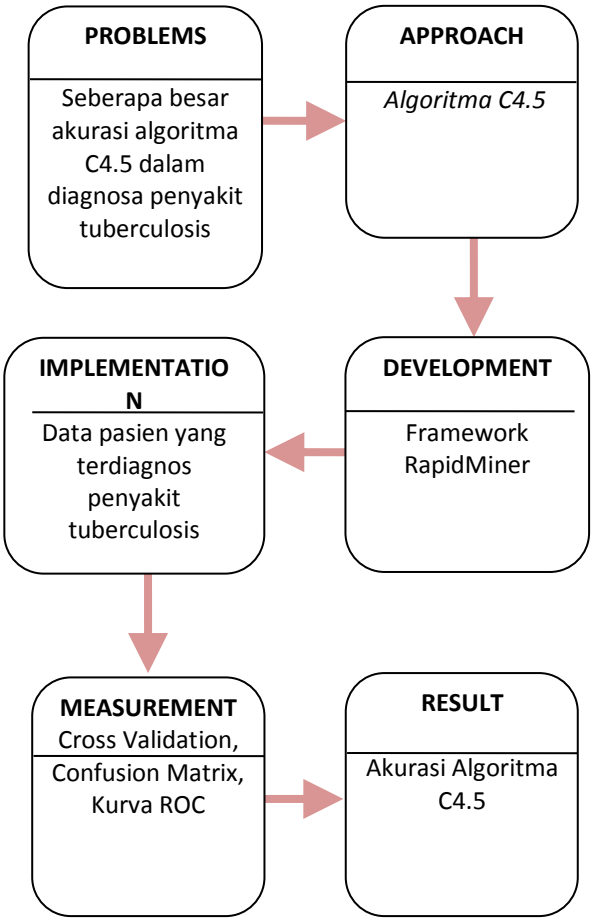
Beberapa penelitian yang berkaitan dengan diagnosa penyakit dengan teknik klasifikasi, diantaranya dilakukan oleh (Kumar & Umatejaswi, 2017) dengan melakukan komparasi algoritma Naive Bayes, Random Tree, C4.5, dan Simple Logistics untuk diagnosa penyakit diabetes. Hasil dari penelitian tersebut menyatakan bahwa algoritma C4.5 memberikan tingkat akurasi yang paling baik dibandingkan metode yang lainnya. Selain itu, ada juga penelitian yag dilakukan oleh (Purushottam, Saxena, & Sharma, 2016) dalam prediksi penyakit jantung menggunakan algoritma C4.5.

Pada penelitian ini, penulis akan menerapkan algoritma C4.5 untuk mendiagnosa penyakit tuberculosis. Data yang penulis gunakan adalah data pasien puskesmas Bojonggede yang terdiagnosa tuberculosis. Pengolahan data menggunakan algoritma C4.5 menghasilkan pohon keputusan, kemudian akan diinterpretasikan ke dalam aturan-aturan keputusan (rules) yang dapat digunakan sebagai acuan untuk melakukan diagnosa penyakit TBC.

BAHAN DAN METODE

A. Kerangka Pemikiran Pemecahan Masalah

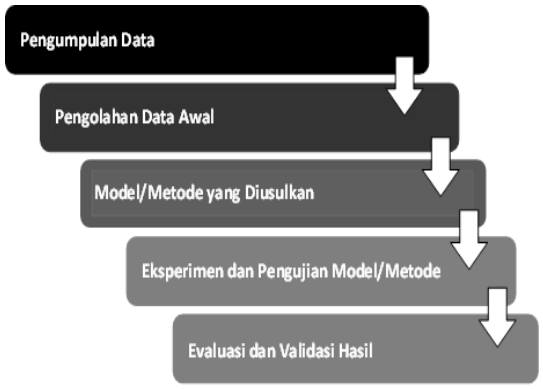
Penelitian ini terdiri dari beberapa tahap seperti terlihat pada kerangka pemikiran Gambar 1 Permasalahan (*problem*) pada penelitian ini adalah Belum diketahui algoritma yang akurat untuk diagnosa penyakit tuberculosis. Oleh karena itu dibuat *approach* (model) *Algoritma C4.5* untuk memecahkan permasalahan kemudian dilakukan pengujian terhadap kinerja dari metode tersebut. Pengujian menggunakan metode *Cross Validation*, *Confusion Matrix* dan kurva ROC. Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner. Desain ekperimennya digunakan CRISP-DM (Sumathi & Sivanandam, 2006).



Gambar 1. Kerangka Pemikiran Pemecahan Masalah

B. Tahap Penelitian

Menurut (Dawson, 2009) terdapat empat metode penelitian yang umum digunakan, diantaranya: *Action Research, Experiment, Case Study,* dan *Survey*. Penelitian ini adalah penelitian eksperimen dengan menjalankan beberapa langkah proses penelitian seperti terlihat pada gambar berikut :



Sumber: Dawson (2009)

Gambar 2. Tahapan Penelitian

C. Evaluasi dan Validasi Model

Untuk mengukur akurasi model maka dilakukan evaluasi dan validasi menggunakan teknik:

1. *Confusion matrix*
   Confusion Matrix adalah alat (*tools*) visualisasi yang biasa digunakan pada supervised learning. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya (Gorunescu, 2011). Confusion matrix berisi informasi aktual (actual) dan prediksi predicted) pada sisitem klasifikasi.
2. Kurva ROC (*Reciever Operating Characteristic*)
   Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vercellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC dihitung menggunakan rumus: (Liao, 2007)

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\,(\mathrm{x}i^r, \mathrm{x}j^r)$$

Dimana

$$\boldsymbol{\psi}(\mathrm{X,Y}) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

Performance keakurasian AUC dapat diklasifikasikan menjadi lima kelompok yaitu (Gorunescu, 2011):
0.90 – 1.00 = *Exellent Clasification*
0.80 – 0.90 = *Good Clasification*
0.70 – 0.80 = *Fair Clasification*
0.60 – 0.70 = *Poor Clasification*
0.50– 0.60 = *Failure*

HASIL DAN PEMBAHASAN

A. Analisa Data

Pada penelitian ini data yang digunakan sebanyak 136 data pasien tuberculosis (TBC) baik yang positif maupun yang negatif. Variabel input pada penelitian ini terdiri dari enam variabel, yaitu: 1. Keringat pada malam hari tanpa aktivitas fisik, 2. Berat badan turun, 3. Nafsu makan berkurang 4. Mudah lelah dan lemah, 5. Demam, 6. Batuk berdahak lebih dari tiga minggu disertai batuk darah, Sedangkan variabel output adalah variabel penyakit TBC. Perangkat lunak yang digunakan untuk menganalisa adalah RapidMiner *versi 5.3*.

B. Algoritma C4.5

Berikut akan dibahas prediksi apakah pasien terdiagnosa tuberculosis atau tidak, menggunakan metode klasifikasi.
Langkah untuk membuat pohon keputusan, yaitu :

1. Hitung nilai *entropy*. Dari data training diketahui jumlah kasus ada 136, pasien yang termasuk kelas TBC Ya 60 *record* dan Tidak 76 *record* sehingga didapat *entropy*:

$$Entropy(S) = \sum_{i=1}^{n} - pi \cdot \log_2 pi$$

= (-60/136· $\log_2$ (60/136))+(-76/136· $\log_2$ (76/136))
= 0.9900

2. Hitung nilai *gain* untuk tiap atribut, lalu tentukan nilai *gain* tertinggi. Yang mempunyai nilai *gain* tertinggi itulah yang akan dijadikan akar dari pohon. Misalkan untuk atribut keringat malam, didapat nilai *gain*:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\,(Si)$$

Gain (*S,A*) = 0.9900 – (75/136(0.9311)+61/136(0.6808)
= 0.1712

Perhitungan *entropy* dan *gain* untuk semua atribut dilakukan, untuk mendapatkan nilai gain tertinggi. Hasil perhitungan seluruh atribut terlihat pada Tabel 1.

Tabel 1. Nilai *entropy* dan *gain* untuk menentukan simpul akar

Simpul	Kasus	Ya	Tidak	*entropy*	*gain*
(KM)	136	60	76	0,9900	0,1712
Ya	75	49	26	0,9311	
Tidak	61	11	50	0,6808	
(BBT)	136	60	76	0,9900	0,1593
Ya	87	53	34	0,9653	
Tidak	49	7	42	0,5917	
(NMB)	136	60	76	0,9900	0,1451
Ya	78	49	29	0,9520	
Tidak	58	11	47	0,7007	

(BBB)	136	60	76	0,9900	0,1360
Ya	54	38	16	0,8767	
Tidak	82	22	60	0,8390	
(MLL)	136	60	76	0,9900	0,1416
Ya	81	50	31	0,9599	
Tidak	55	10	45	0,6840	
(Demam)	136	60	76	0,9900	**0,1760**
Ya	65	45	20	0,8905	
Tidak	71	15	56	0,7439	

Dari hasil perhitungan *entropy* dan *gain* yang didapat pada Tabel 1, terlihat bahwa atribut demam mempunyai nilai gain tertinggi yaitu 0.1760. Oleh karena itu maka demam merupakan simpul akar pada pohon keputusan. Untuk menentukan simpul berikutnya, dilakukan lagi perhitungan *entropy* dan *gain* berdasarkan atribut demam yang bernilai Ya dan Tidak, seperti terlihat pada tabel 2 dan 3 di bawah ini:

Tabel 2. Nilai *entropy* dan *gain*atribut demam=Ya

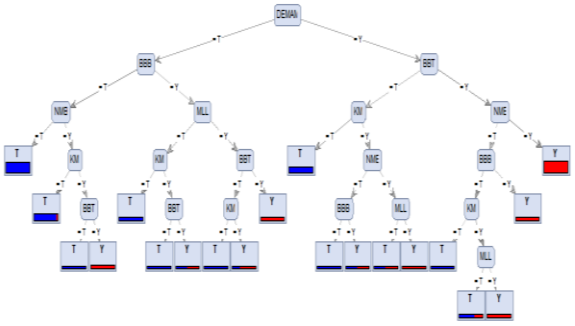
simpul	kasus	Ya	Tidak	entropy	gain
(KM)	65	45	20	0,8905	0,1847
Ya	43	37	6	0,5830	
Tidak	22	8	14	0,9457	
(BBT)	65	45	20	0,8905	**0,2925**
Ya	45	40	5	0,5033	
Tidak	20	5	15	0,8113	
(NMB)	65	45	20	0,8905	0,1847
Ya	43	37	6	0,5830	
Tidak	22	8	14	0,9457	
(BBB)	65	45	20	0,8905	0,0988
Ya	34	29	5	0,6024	
Tidak	31	16	15	0,9992	
(MLL)	65	45	20	0,8905	0,1847
Ya	43	37	6	0,5830	
Tidak	22	8	14	0,9457	

Tabel 2 merupakan perhitungan untuk atribut demam dengan nilai Ya. Dari hasil perhitungan pada tabel 2, terlihat bahwa atribut berat badan turun memiliki nilai *gain* tertinggi, yaitu 0.2925. Ini berarti atribut yang menjadi simpul di bawah atribut demam = Ya (simpul 1.1) adalah atribut berat badan turun.

Tabel 3 Nilai *entropy* dan *gain*atribut demam = Tidak

simpul	kasus	Ya	Tidak	entropy	gain
(KM)	71	15	56	0,7439	0,0988
Ya	32	12	20	0,9544	
Tidak	39	3	36	0,3912	
(BBT)	71	15	56	0,7439	0,0680
Ya	42	13	29	0,8926	
Tidak	29	2	27	0,3621	
(NMB)	71	15	56	0,7439	0,0768
Ya	35	12	23	0,9275	
Tidak	36	3	33	0,4138	
(BBB)	71	15	56	0,7439	**0,0989**
Ya	20	9	11	0,9928	
Tidak	51	6	45	0,5226	
(MLL)	71	15	56	0,7439	0,0945
Ya	38	13	25	0,9268	
Tidak	33	2	31	0,3298	

Tabel 3 merupakan perhitungan untuk atribut demam dengan nilai Tidak. Dari hasil perhitungan pada tabel 3, terlihat bahwa atribut batuk berdahak dan berdarah memiliki nilai *gain* tertinggi, yaitu 0.0989. Ini berarti atribut yang menjadi simpul di bawah atribut demam = Tidak (simpul 1.2) adalah atribut batuk berdahak dan berdarah. Dengan menggunakan RapidMiner 5.3 didapatkan pohon keputusan akhir yang dihasilkan dari perhitungan entropy dan gain untuk seluruh atribut seperti terlihat pada Gambar 2 di bawah ini:



Gambar 3. Pohon Keputusan Hasil Perhitungan Algoritma C4.5

C. Pengujian Model

Model yang telah dibentuk diuji tingkat akurasinya dengan memasukan data uji yang berasal dari data training. Karena data yang didapat dalam penelitian ini setelah proses preprocessing hanya 136 data maka digunakan metode cross validation untuk menguji tingkat akurasi. Untuk nilai akurasi model untuk metode C4.5 sebesar 84.56%.

1. *Confusion Matrix*
   Tabel 4 adalah perhitungan berdasarkan data *training*, diketahui dari 136 data, 66 diklasifikasikan tidak (negatif) sesuai dengan prediksi yang dilakukan dengan metode C4.5, lalu 11 data diprediksi tidak (nagatif) tetapi ternyata ya (positif), 49 data *class* ya (positif) diprediksi sesuai, dan 10 data diprediksi ya (positif) ternyata tidak (negatif).

Tabel 4. Model *Confusion Matrix* Algoritma C4.5

accuracy: 84.56% +/- 4.01% (mikro: 84.56%)

	true T	true Y	class precision
pred. T	66	11	85.71%
pred. Y	10	49	83.05%
class recall	86.84%	81.67%	

2. Kurva ROC
   Hasil perhitungan divisualisasikan dengan kurva ROC. Adapun kurva ROC untuk algoritma C45 terlihat seperti gambar di bawah ini:

Gambar 4. Kurva ROC algoritma C4.5

Kurva ROC pada gambar 3 mengekspresikan *confusion matrix* dari Tabel 4. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.

Dari gambar di atas terlihat bahwa nilai *area undercurve* (AUC) algoritma C4.5 adalah 0,938.

Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011):

a. 0.90-1.00 = klasifikasi sangat baik
b. 0.80-0.90 = klasifikasi baik
c. 0.70-0.80 = klasifikasi cukup
d. 0.60-0.70 = klasifikasi buruk
e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokkan di atas maka dapat disimpukan bahwa algoritma C4.5 termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

KESIMPULAN

Kesimpulan yang dapat diambil berdasarkan penelitian ini adalah bahwa diagnosa penyakit tuberculosis menggunakan algoritma C4.5 ini dapat digunakan sebagai langkah awal dalam mendeteksi kemungkinan terkena tuberculosis dilihat dari gejala klinis utama yang dialami oleh pasien. Berdasasarkan pengujian model didapatkan hasil bahwa performa model algoritma C4.5 memberikan tingkat akurasi kebenaran sebesar 84,56% dengan nilai area under the curve (AUC) sebesar 0,938. Hal ini menunjukkan bahwa model tersebut termasuk katagori klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

REFERENSI

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Students Performance Using ID3 And C4.5 Classification Algorithms. *International Journal of Data Mining & Knowledge Management Process*, *3*(5), 39–52. https://doi.org/10.5121/ijdkp.2013.3504

Amrin, A. (2018). Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes. *Jurikom*, *5*(5), 498–502.

Dawson, C. W. (2009). *Projects in Computing and Information System A Student's Guide*. England: Addison-Wesley.

Fine, J. (2012). *An Overview Of Statistical Methods in Diagnostic Medicine*. Chapel Hill.

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques*. Verlag Berlin Heidelberg: Springer.

Kumar, P. ., & Umatejaswi, V. (2017). Diagnosing Diabetes using Data Mining Techniques. *International Journal of Scientific and Research Publications*, *7*(6), 705–709.

Liao, T. W. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Application.* Singapore: World Scientific Publishing.

Orhan, E., Temurtas, F., & Tanrıkulu, A. Ç. (2010). Tuberculosis Disease Diagnosis Using Artificial Neural Networks. *Springer*, 299–302.

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System using Decision Tree. *International Conference on Computing, Communication and Automation (ICCCA2015)*, 962–969. https://doi.org/10.1016/j.procs.2016.05.288

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. Berlin Heidelberg New York: Springer.

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making*. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.

Widoyono. (2011). *Penyakit Tropis Epidemiologi, Penularan, Pencegahan dan Pemberantasan.* Jakarta: Erlangga.