

Klasifikasi Berbasis Algoritma C4.5 untuk Deteksi Kenaikan *Case Fatality Rate* Demam Berdarah

Anik Andriani

Manajemen Informatika, AMIK BSI, Yogyakarta, Indonesia

anik.aai@bsi.ac.id

Abstract— Extraordinary Events Dengue often occurs in Indonesia which causes patients died that affects the percentage deaths due to Dengue increased. It's called Case Fatality Rate (CFR) Dengue. The government's target on CFR Dengue is less than 1%. But CFR Dengue in some provinces in Indonesia is over 1%. Application of classification technique with C4.5 algorithm in this cases to know data pattern from CFR Dengue dataset in all province in Indonesia. Analysis process from historical data used for classification process applies stages of Knowledge Discovery in Databases consist of nine stages. The Classification result is Classification rule. It can describe data pattern extracted from CFR Dengue dataset. One of utilization to knowing the most influence factors to increase of CFR Dengue in a province. Classification result are evaluated by Confusion Matrix and ROC (Receiver Operating Characteristic) curve to know classification result performance's. Evaluation result is known performance of classification rule generated into Excellent category.

Keywords: *CFR Dengue, Classification, C4.5 Algorithm*

I. PENDAHULUAN

Penyakit Demam Berdarah merupakan salah satu penyebab kesakitan dan kematian pada anak di Asia Tenggara [1]. Wilayah Indonesia dilihat dari letak geografisnya termasuk daerah tropis dengan curah hujan ya cukup tinggi. Karena tipe seperti itulah yang menyebabkan Indonesia sebagai salah satu negara di Asia Tenggara sebagai wilayah yang mudah untuk perkembangbiakan nyamuk penyebab Demam Berdarah. Hal ini mengakibatkan Demam Berdarah menjadi salah satu masalah pada kesehatan di Indonesia. Bahkan Kejadian Luar Biasa (KLB) sering terjadi [2]. Berbagai upaya telah dilakukan untuk mencegah angka kematian yang diakibatkan penyakit ini, namun masih saja ditemukan pasien yang meninggal akibat terkena Demam Berdarah yang berdampak pada persentase nilai kematian akibat Demam Berdarah atau biasa disebut dengan *Case Fatality Rate* (CFR) masih cukup tinggi.

Program pemberantasan dan pencegahan penyakit Demam Berdarah telah berlangsung lebih dari empat puluh tahun dalam menurunkan angka CFR penyakit ini. Pada Renstra Pembangunan di Indonesia untuk CFR Demam Berdarah ditargetkan kurang dari 1%. Berdasarkan data statistik dari Kementerian Kesehatan masih banyak provinsi di Indonesia yang

persentase CFR Demam Berdarah diatas 1% [3]. Sehingga dapat diambil kesimpulan bahwa pelaksanaan program pemberantasan dan pencegahan penyakit Demam Berdarah sudah mampu menurunkan angka CFR penyakit tersebut hanya di beberapa provinsi saja, sedangkan beberapa daerah masih banyak menunjukkan presentase CFR akibat Demam Berdarah masih tinggi [4].

Penelitian ini bertujuan membuat klasifikasi data pada dataset CFR penyakit Demam Berdarah berdasarkan provinsi di Indonesia untuk mendeteksi faktor yang paling berpengaruh terhadap kenaikan CFR penyakit Demam Berdarah. Klasifikasi sendiri merupakan salah satu teknik dari Data Mining yang merupakan proses penambangan data pada sebuah dataset yang dilakukan dengan cara otomatis maupun semi otomatis dimana tujuannya adalah untuk menemukan pola data dari dataset tersebut. Pola data yang dihasilkan dapat divalidasi maupun diidentifikasi dan dapat dimanfaatkan dalam hal pembuatan suatu prediksi [5]. Teknik klasifikasi sendiri merupakan teknik dalam data mining yang dapat digunakan untuk membuat sebuah prediksi [6]. Teknik klasifikasi bekerja dengan cara menempatkan *object* dataset ke dalam sebuah *class*. Proses penempatan tersebut menghasilkan pola data yang dapat dimanfaatkan untuk prediksi, deteksi, maupun sebagai rekomendasi dalam pengambilan keputusan [7]. Dalam teknik klasifikasi dikenal banyak algoritma yang dapat digunakan untuk membuat klasifikasi data. Salah satu algoritma yang terkenal dan mempunyai kehandalan yang baik dalam membuat klasifikasi data adalah Algoritma C4.5. Algoritma ini bekerja dengan cara memetakan *object* data berupa atribut-atribut ke dalam sebuah *class*. Hasil dari proses tersebut diperoleh *classification rule* [8].

Berdasarkan latar belakang di atas, maka dapat dirumuskan permasalahan penelitian yang diuraikan dalam beberapa *research question* berikut:

1. Bagaimanakah model yang dibangun Algoritma C4.5 dalam mengklasifikasi dataset CFR Demam Berdarah?

2. Bagaimanakah fungsionalitas dari Algoritma C4.5 dalam membuat klasifikasi dataset CFR Demam Berdarah?
3. Bagaimanakah tingkat performa dari *classification rule* yang dihasilkan?

II. KLASIFIKASI DATASET CFR DEMAM BERDARAH

A. CFR Demam Berdarah

Demam berdarah merupakan penyakit yang disebabkan karena gigitan nyamuk *Aedes Aegypti* [1]. Di wilayah Indonesia berpotensi mengalami Kejadian Luar Biasa (KLB) penyakit Demam Berdarah karena wilayah geografisnya masuk daerah tropis yang mudah untuk perkembangbiakan nyamuk dan penyebaran penyakit Demam Berdarah. Dampak terjadinya KLB Demam Berdarah adalah banyaknya korban meninggal yang disebut dengan angka kematian. Sedangkan persentase angka kematian yang disebabkan oleh Demam Berdarah yang dihitung per jumlah angka kematian disebut dengan *Case Fatality Rate* (CFR) Demam Berdarah [3].

B. Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma dari teknik Klasifikasi [5]. Klasifikasi sendiri merupakan salah satu teknik dalam Data Mining. Selain untuk mengetahui pola data dalam dataset, pemanfaatan data mining dapat digunakan untuk mengeksplorasi kemungkinan pengetahuan yang masih tersembunyi dalam sebuah dataset [9]. Dalam kasus ini pemanfaatan data mining untuk mengetahui pola data yang mempengaruhi kenaikan CFR Demam Berdarah dengan menggunakan teknik klasifikasi. Pada dasarnya penerapan teknik Klasifikasi dengan membagi objek-objek sehingga setiap objek ditetapkan dalam *class* yang merupakan sebuah kategori [10]. Algoritma C4.5 membuat klasifikasi data untuk mendapatkan *classification rule* yang hasilnya dapat diterapkan pada kasus lain [11]. *Classification rule* tersebut lebih mudah dibaca bila berbentuk *decision tree* [5]. Tahapan pembuatan *decision tree* antara lain [12]:

- 1) *Persiapan data training*: diambil dari dataset yang sudah disiapkan. Dataset dapat berasal dari data histori CFR Demam Berdarah.
- 2) *Perhitungan akar pohon*: perhitungan dilakukan dengan menghitung nilai entropy yang dilanjutkan perhitungan nilai gain masing-masing atribut. Atribut dengan nilai gain tertinggi akan menjadi akar pohon. Rumus perhitungan entropy [11] dapat dilihat pada persamaan (1).

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

Keterangan:

S=Himpunan kasus

n=Jumlah partisi S

P_i=Proporsi S_i terhadap S

Persamaan diatas menunjukkan rumus untuk menghitung entropy dari suatu himpunan S dimana S merupakan nilai-nilai dalam atribut yang digunakan pada dataset. Sedangkan n merupakan jumlah partisi dari himpunan S dalam hal ini ada partisi ada dua yaitu “naik” dan “tidak”.

Sedangkan perhitungan nilai gain [11] dilakukan dengan rumus berikut:

$$Gain(S, A) = entropy(S) - \sum_{i=1}^n \frac{|S_i|}{S} * Entropy(S_i) \quad (2)$$

Keterangan:

S=Himpunan Kasus

A=Fitur

n=Jumlah partisi atribut A

|S_i|=Proporsi S_i terhadap S

|S|=Jumlah kasus dalam S

Hasil perhitungan entropy digunakan untuk perhitungan gain berdasarkan persamaan (2). Atribut dengan nilai gain tertinggi selanjutnya menjadi akar dari *decision tree* atau *root*, sedangkan atribut lain yang bukan merupakan atribut dengan nilai gain tertinggi akan dihitung kembali. Dari hasil perhitungan ulang tersebut akan diperoleh nilai gain tertinggi kembali yang selanjutnya dijadikan *sub root* di bawah *root*. Selanjutnya proses perhitungan seperti sebelumnya diulang kembali sampai semua atribut terpilih menjadi *sub root* sehingga semua atribut dalam dataset terpartisi membentuk sebuah pohon keputusan.

- 3) Ulangi perhitungan entropy dan gain pada atribut-atribut yang belum terpartisi
- 4) Proses perhitungan diatas diulangi terus sampai semua atribut terpartisi. Proses ini akan berhenti bilamana semua record dalam simpul N mendapat kelas yang sama, tidak ada atribut yang belum terpartisi, tidak ada record dalam cabang yang kosong.

Setelah proses klasifikasi selesai, selanjutnya diperlukan proses evaluasi. Evaluasi dilakukan untuk mengetahui tingkat performa dari hasil klasifikasi yang berupa *classification rule*. Kriteria yang dinilai paling tepat dalam mengukur tingkat performa dari *classification rule* adalah prediksi akurasi yang secara rinci menjelaskan kinerja *classifier*. Dalam prediksi akurasi kinerja *classifier* adalah seberapa tepat kasus X diklasifikasikan ke dalam kelas X secara benar [10]. Nilai

akurasi juga digunakan untuk menunjukkan kemampuan algoritma dalam mengklasifikasi *record* data secara benar [13]. Penentuan prediksi akurasi dapat dilakukan dengan menggunakan Confusion Matrix dan Kurva ROC (*Receiver Operating Characteristic*). Perbedaan Confusion Matrix dan Kurva ROC yaitu Confusion Matrix menampilkan nilai akurasi berupa angka persentase ketepatan *classification rule* dalam mengklasifikasi data yang dilengkapi dengan nilai *precision* atau *confidence* dan nilai *recall* atau *sensitivity* [14]. Sedangkan Kurva ROC menunjukkan nilai akurasi dengan menggunakan grafik dua dimensi. Kurva ROC disebut juga dengan AUC yaitu *The area under curve* [15]. Nilai yang ditunjukkan pada kurva ROC dikategorikan dalam beberapa kategori yang ditunjukkan pada Tabel 1.

Tabel 1. KATEGORI NILAI KURVA ROC [7]

Nilai akurasi	Kategori
0,90-1,00	<i>Excellent Classification</i>
0,80-0,90	<i>Good Classification</i>
0,70-0,80	<i>Fair Classification</i>
0,60-0,70	<i>Poor Classification</i>
0,50-0,60	<i>Failure</i>

Tabel 1 menunjukkan nilai evaluasi yang ditunjukkan kurva ROC bila di atas 0,6 maka klasifikasi dianggap berhasil dan dibagi ke dalam kategori Excellent, Good, Fair, dan Poor. Sedangkan nilai evaluasi dibawah 0,6 dianggap proses klasifikasi gagal dan rule hasil klasifikasi tidak bisa diterapkan.

III. METODOLOGI

Tahapan-tahapan dalam membuat klasifikasi pada dataset CFR Demam Berdarah terdiri dari sembilan tahapan yang mengacu pada tahap-tahap dalam *Knowledge Discovery in Databases* (KDD). Penggunaan KDD mempermudah dalam proses analisis data. Proses analisis data dengan tahapan-tahapan KDD membahas secara detail dari tahap membuat dataset, pembersihan data, transformasi, pembagian data dan sebagainya [16]. Tahapan-tahapan KDD tersebut antara lain:

A. Domain understanding and KDD goals

Pada tahap ini ditentukan tujuan dari KDD yaitu membuat klasifikasi pada dataset CFR Demam Berdarah untuk dapat mendeteksi potensi kenaikan CFR Demam Berdarah dan mengetahui faktor yang paling berpengaruh terhadap kenaikan tersebut.

B. Selection and addition

Pada tahap ini dibangun dataset yang akan digunakan. Dataset dibangun berdasarkan data statistik Kementerian Kesehatan Republik Indonesia tahun 2008 sampai dengan 2014. Dataset yang dibangun berisi 170 *record* data dengan jumlah 6 atribut dan 1 *class*. Atribut merupakan faktor-faktor

yang berpengaruh pada kenaikan CFR Demam Berdarah. Atribut yang digunakan yaitu presentase penduduk miskin di suatu provinsi, kepadatan penduduk provinsi, presentase tempat tinggal dan sanitasi layak masyarakat di suatu provinsi, curah hujan di suatu provinsi, kelembaban udara di suatu provinsi, dan presentase jumlah masyarakat pada suatu provinsi yang menerapkan STBM (Sanitasi Total Berbasis Masyarakat). Sedangkan *class* digunakan untuk nilai prediksi dimana dalam hal ini hanya bernilai “naik” atau “tidak” yang menunjukkan terjadinya kenaikan presentase CFR Demam Berdarah pada suatu provinsi atau tidak.

C. Preprocessing: data cleaning

Dataset yang telah dibangun dilakukan pembersihan dari data noise (data yang isinya tidak tepat atau salah) dan missing value (data yang isinya tidak lengkap). Dalam proses ini diperoleh data sebanyak 156 *record* data.

D. Transformation data

Tahap transformasi data merupakan tahap untuk meningkatkan kehandalan data dengan cara mentransformasi dataset dalam bentuk kategori. Selain melakukan transformasi data dalam bentuk kategori, tahap ini juga membagi data dalam dua kelompok yaitu data training dan data testing dengan komposisi 80:20. Dalam proses pembagian data menggunakan teknik systematic random sampling. Hasilnya diperoleh data training sebanyak 125 *record* data dan data testing sebanyak 31 *record* data.

E. Data Mining: choosing the appropriate data mining task

Tahap ini merupakan tahap penentuan teknik data mining yang digunakan dalam mencapai KDD goals. Teknik data mining yang dipilih adalah teknik klasifikasi karena tujuannya adalah membuat prediksi dengan mencari pola data dari data histori.

F. Data Mining: choosing the data mining algorithm

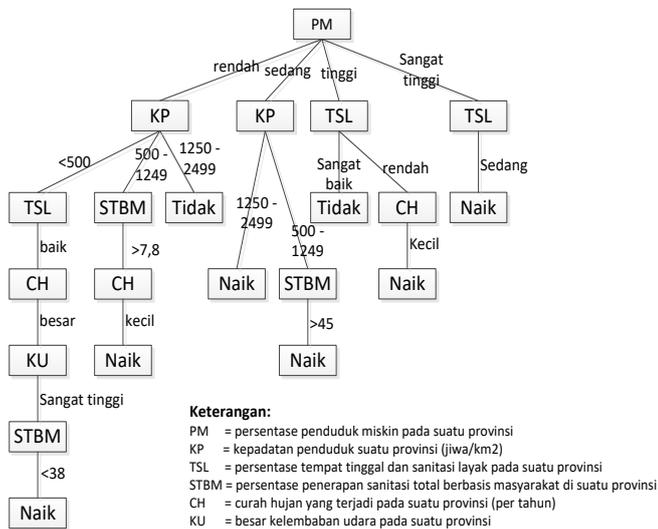
Tahap ini menentukan algoritma yang akan digunakan dalam penelitian. Algoritma yang dipilih untuk membuat klasifikasi data pada data histori CFR Demam Berdarah adalah Algoritma C4.5.

G. Employing the data mining algorithm

Tahap ini merupakan proses membuat klasifikasi data dengan algoritma yang telah ditentukan.

H. Evaluation

Tahap ini dilakukan untuk mengukur tingkat performa dari algoritma yang digunakan dalam mengklasifikasi data. Evaluasi dilakukan pada *classification rule* yang dihasilkan dari proses klasifikasi. Evaluasi dilakukan dengan Confusion Matrix dan Kurva ROC.



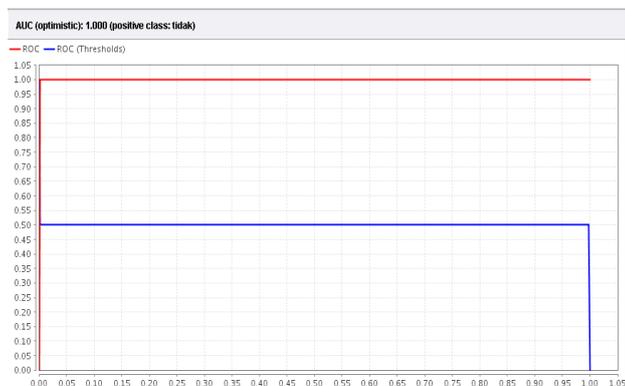
Gambar 2. Pohon Keputusan

Rule hasil klasifikasi selanjutnya dievaluasi untuk mengetahui tingkat performanya. Proses evaluasi menggunakan Confussion Matrix dan kurva ROC dilakukan dengan mengevaluasi *classification rule* terhadap data training yang digunakan untuk klasifikasi dan terhadap data testing yang merupakan data baru yang tidak digunakan untuk proses klasifikasi. Hasil evaluasi *classification rule* dengan Confussion Matrix terhadap data training ditunjukkan pada Tabel 2.

Tabel 2. NILAI AKURASI DARI RULE KLASIFIKASI TERHADAP DATA TRAINING MENGGUNAKAN CONFUSION MATRIX

accuracy: 99,20%			
	true naik	true tidak	class precision
pred. naik	50	0	100,00%
pred. tidak	1	74	98,67%
class recall	98,04%	100,00%	

Pada Tabel 2 diketahui nilai akurasi yang menunjukkan ketepatan rule klasifikasi dalam memprediksi “naik” dan “tidak” pada data training sebesar 99,20%. Sedangkan hasil evaluasi dengan kurva ROC ditunjukkan Gambar 3.



Gambar 3. Nilai Akurasi dari Rule Klasifikasi Terhadap Data Training menggunakan kurva ROC

Nilai akurasi yang ditunjukkan Gambar 3, menunjukkan ketepatan rule klasifikasi dalam memprediksi “naik” dan “tidak” CFR Demam Berdarah pada data training sebesar 1.00. Selain dievaluasi terhadap data training, rule klasifikasi juga dievaluasi dengan menggunakan data baru yaitu data testing. Hasil evaluasi dengan Confussion matrix ditunjukkan Tabel 3.

Tabel 3. NILAI AKURASI DARI RULE KLASIFIKASI TERHADAP DATA TESTING MENGGUNAKAN CONFUSION MATRIX

accuracy: 61,29%			
	true tidak	true naik	class precision
pred. tidak	9	5	64,29%
pred. naik	7	10	58,82%
class recall	56,25%	66,67%	

Pada Tabel 3 diketahui nilai akurasi yang menunjukkan ketepatan rule klasifikasi dalam memprediksi “naik” dan “tidak” pada data testing sebesar 61,29%. Sedangkan hasil evaluasi dengan kurva ROC ditunjukkan Gambar 4.



Gambar 4. Nilai Akurasi dari Rule Klasifikasi Terhadap Data Testing menggunakan kurva ROC

Nilai akurasi yang ditunjukkan Gambar 4, menunjukkan ketepatan rule klasifikasi dalam memprediksi “naik” dan “tidak” CFR Demam Berdarah pada data training sebesar 0,825.

Berdasarkan hasil evaluasi terhadap data training dan data testing menggunakan Confussion matrix dan kurva ROC selanjutnya dilakukan perhitungan rata-rata yang ditunjukkan Tabel 4.

Tabel 4. NILAI AKURASI RATA-RATA HASIL EVALUASI

	Data training	Data testing	Nilai rata-rata
Confussion Matrix	99,20%	61,29%	80,245%
Kurva ROC	1,000	0,825	0,913

Tabel 4 menunjukkan nilai akurasi rata-rata hasil evaluasi terhadap data training dan data testing dengan Confussion

Matrix sebesar 80,245%. Sedangkan nilai akurasi rata-rata hasil evaluasi terhadap data training dan data testing dengan kurva ROC sebesar 0,913. Berdasarkan tabel kategori nilai pada kurva ROC yang ditunjukkan Tabel 1, maka hasil klasifikasi masuk kategori *Excellent Classification*.

V. KESIMPULAN DAN SARAN

Model klasifikasi data yang dibangun menggunakan algoritma C4.5 pada dataset CFR Demam Berdarah menghasilkan rule klasifikasi yang menggambarkan pola data yang digali dari data histori CFR Demam Berdarah. Fungsionalitas dari klasifikasi data tersebut mampu menunjukkan faktor yang paling berpengaruh terhadap potensi kenaikan angka CFR Demam Berdarah pada suatu provinsi adalah persentase angka kemiskinan pada provinsi tersebut. Performa dari rule klasifikasi yang diperoleh dari hasil klasifikasi data menunjukkan performa yang baik. Berdasarkan hasil evaluasi dengan Confussion Matrix dan kurva ROC dengan menerapkan rule klasifikasi pada data training dan data testing diperoleh nilai akurasi rata-rata sebesar 80,245%, sedangkan nilai akurasi yang ditunjukkan kurva ROC sebesar 0,913. Berdasarkan nilai akurasi yang diperoleh tersebut menunjukkan bahwa penerapan teknik klasifikasi dengan menggunakan algoritma C4.5 pada dataset CFR Demam Berdarah masuk kategori *Excellent Classification*.

Penelitian selanjutnya diharapkan dapat menambah jumlah atribut yang kemungkinan berpotensi terhadap kenaikan CFR Demam Berdarah.

REFERENSI

- [1] M. Rahayu, T. Baskoro, and B. Wahyudi, "Studi Kohort Kejadian Penyakit Demam Berdarah Dengue," *Berita Kedokteran Masyarakat*, pp. 163-170, 2010.
- [2] Gama and F. Betty, "Analisis Faktor Risiko Kejadian Demam Berdarah Dengue di Desa Mojosongo Kabupaten Boyolali," *Eksplanasi*, pp. 1-9, 2010.
- [3] U. S. Sutarjo, et al., "Profil Kesehatan Indonesia 2014," *Kementrian Kesehatan Republik Indonesia*, 2015.
- [4] S. Winarsih, "Hubungan Kondisi Lingkungan Rumah dan Perilaku PSN dengan Kejadian DBD," *Journal of Public Health*, pp. 1-9, 2013.
- [5] H. Witten, E. Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques Third Edition*. Burlington: Elsevier, 2011.
- [6] G. J. Myatt, *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey: John Wiley & Sons, 2007.
- [7] F. Gorunescu, *Data Mining Concepts, Models and Techniques*. Berlin: Springer, 2011.
- [8] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. USA: Elsevier, 2012.
- [9] S. Sumathi and S. N. Sivanandam, *Introduction to Data Mining and Its Applications*. New York, United States of America: Springer, 2006.
- [10] M. Bramer, *Principles of Data Mining*. London: Springer, 2007.
- [11] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. New York: CRC Press, 2009.
- [12] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining, 2nd ed.* New Jersey: Wiley, 2014.
- [13] e. a. Pepe, "Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design," *JNCI*, vol. 100, no. 20, pp. 1432-1438, Oct. 2008.
- [14] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, pp. 37-63, 2011.
- [15] C. Vercellis, *Business Intelligence*. United Kingdom: John Wiley and Sons, 2009.
- [16] Andriani, "Application of C4.5 Algorithm For Detection of Cooperatives Failure in Province Level," in *International Seminar on Scientific Issues and Trends (ISSIT)*, Bekasi, 2014, pp. 168-174.