Klasifikasi Data Rekam Medis Berdasarkan Kode Penyakit Internasional Menggunakan Algoritma C4.5

Wenefrida Tulit Ina

Jurusan Teknik Elektro, Fakultas Sains dan Teknik, Universitas Nusa Cendana
Jl. Adi-Sucipto, Penfui, Kupang, Indonesia. 85000

Email: wenefrida150477@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengetahui model klasifikasi penyakit berdasarkan tumpukan data rekam medis, menggunakan salah satu metode dalam data mining, yaitu algoritma C4.5. Untuk mencapai tujuan penelitian, maka dipilih 4 atribut sesuai data rekam medis. Data rekam medis yang dimaksud terdiri dari atribut diagnosa penyakit berdasarkan *International Classification of Diseases* -10th (ICD-10), jenis kelamin, umur pasien, bulan masuk pasien ke rumah sakit. Hasil penelitian menunjukkan bahwa ada 5 jenis klasifikasi penyakit, yaitu A00-B99, I00-I99, J00-J99, O00-O99 dan Z00-Z99. Penyakit A00-B99 umumnya diderita oleh laki-laki dengan kategori umur muda dan dewasa, perempuan dengan kategori umur tua, kategori bayi dan anak hanya terjadi pada bulan Januari, Maret, April, Mei. Penyakit I00-I99 umumnya diderita oleh laki-laki dengan kategori umur tua. Penyakit J00-J99 umumnya diderita oleh laki-laki dengan kategori umur bayi dan anak pada bulan Nopember. Penyakit O00-O99 umumnya diderita oleh perempuan dengan kategori umur muda dan dewasa. Penyakit Z00-Z99 umumnya diderita oleh bayi dan anak pada bulan Pebruari, Juni, Juli, Agustus, September, Oktober, Desember, sedangkan pada bulan Nopember diderita oleh bayi dan anak dengan jenis kelamin perempuan. Algoritma C4.5 kurang maksimal dalam menghasilkan klasifikasi data rekam medis karena jumlah kelas tujuan atau label kelas sangat banyak dan persentasi data yang terbaca kurang dari 50%. Klasifikasi penyakit yang dihasilkan hanya 5 kelas dari 21 kelas keseluruhan sesuai kode penyakit internasional.

Abstract

This study aims to determine the classification of disease models based on the data stack of medical records, using one of the methods in the data mining algorithm C4.5. To achieve the research goal is then selected four attributes appropriate medical records consisting of attributes Diagnosis of disease based on the International Classification of Diseases-10th (ICD-10), Gender, patient age, Month patient admission to the hospital. The results show that there are 5 types of disease classification are A00-B99, I00-I99, J00-J99, O00-O99 and Z00-Z99. A00-B99 disease generally affects men with young and adult age categories, women with older age category, the category of infants and children only occurred in January, March, April May. I0-I99 disease generally affects men with older age category. J00-J99 diseases commonly suffered by men with age categories of infants and children in November. O00-O99 illnesses commonly suffered by women with young and adult age categories. Z00-Z99 disease usually affects infants and children in February, June, July, August, September, October, December, whereas in November suffered by infants and children with the female gender. C4.5 algorithm generates a classification less than the maximum in the medical records for the number of classes or class label purposes very much and the percentage of data that is read is less than 50%. Classification of diseases produced only 5 classes of 21 overall class by international disease code.

Keywords: Medical Records, ICD-10, C4.5

1. Latar Belakang

Berkembangnya **Ilmu Data Mining** memberikan inovasi baru dalam hal pendayagunaan kumpulan data yang banyak sehingga dapat bermanfaat bagi pengembangan pengetahuan, baik secara khusus pada bidang yang berkaitan dengan data tersebut maupun secara global. Banyak fungsi yang dapat diterapkan dari ilmu data mining antara lain, estimasi, prediksi, klasterisasi, klasifikasi dan asosiasi. Untuk mencapai fungsi-fungsi tersebut dilakukan dengan berbagai metode (algoritma) seperti regresi untuk estimasi, *Support Vector Macsine* (SVM) untuk prediksi, *K-Means* untuk klasterisasi, C4.5 untuk klasifikasi, *apriori* untuk asosiasi. [1].

Salah satu penerapan ilmu data mining, yaitu pada permasalahan penumpukan data rekam medis di Rumah Sakit. Rekam medis adalah berkas yang berisikan catatan dan dokumen tentang identitas pasien, pemeriksaan, pengobatan, tindakan dan pelayanan lain yang diberikan kepada pasien. Rekam medis harus dibuat secara tertulis, lengkap, dan jelas atau secara elektronik. Penyelenggaraan rekam medis dengan menggunakan teknologi informasi elektronik diatur oleh peraturan tersendiri. Informasi dalam rekam medis dijaga kerahasiaannya oleh dokter, tenaga kesehatan dan petugas pengelola serta pimpinan sarana pelayanan kesehatan. Data rekam medis terus terakumulasi setiap hari seiring dengan aktivitas rumah sakit. Pemanfaatan rekam medis dapat dipakai sebagai: (1) pemeliharaan kesehatan dan pengobatan pasien; (2) alat bukti dalam proses penegakkan hukum, disiplin kedokteran dan kedokteran gigi, dan penegakkan etika kedokteran dan kedokteran gigi; (3) keperluan pendidikan dan penelitian; (4) dasar pembayaran biaya pelayanan kesehatan; (5) data statistik kesehatan. [2]

Berdasarkan permasalahan yang telah dijelaskan terdahulu, maka dibuat penelitian untuk klasifikasi data rekam medis berdasarkan kode penyakit internasional (ICD-10). Dalam hal ini menggunakan salah satu metode data mining, yaitu algoritma C4.5 dalam fungsi klasifikasi.

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Algoritma ini merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami dan dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* (SQL) yang berguna untuk mencari *record* pada kategori tertentu. Pohon keputusan berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan variabel target. [3]. Karena pohon keputusan memadukan antara eksplorasi data dan pemodelan, maka sangat baik untuk langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain. Sebuah pohon keputusan adalah sebuah struktur yang dapat diigunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan demikian, masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain. [4]

Sebuah model pohon keputusan terdiri dari sekumpulan aturan untuk membagi sejumlah populasi yang heterogen menjadi lebih kecil, lebih homogen dengan memperhatikan variabel tujuannya. Variabel tujuan biasanya dikelompokkan dengan pasti dan model pohon keputusan lebih mengarah pada perhitungan probabilitas dari tiap-tiap *record* terhadap kategori-kategori tersebut atau untuk mengklasifikasi *record* dengan mengelompokkannya dalam satu kelas. Selain itu, pohon keputusan juga dapat digunakan untuk mengestimasi nilai dari variabel kontinu meskipun ada beberapa teknik yang lebih sesuai untuk kasus ini. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Salah satu atribut merupakan atribut yang menyatakan data solusi per item data yang disebut target atribut atau atribut kelas tujuan. Atribut memiliki nilai-nilai yang dinamakan *instance* [5].

Ada 2 variabel yang dipakai dalam menentukan akar dari pohon keputusan, yaitu nilai *entropy* dan nilai gain. Nilai *entropy* diperoleh dari rumus:

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

Keterangan :

S = Himpunan Kasus
n = Jumlah partisi S
p = Proporsi dari S_i terhadap S

Nilai *gain* diperoleh dari rumus :

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

Keterangan:

S = Himpunan Kasus
A = Atribut
n = Jumlah partisi atribut A
$|S_i|$ = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

Atribut dengan nilai gain tertinggi akan dipilih menjadi akar dari pohon keputusan. Secara umum dalam membangun pohon keputusan dengan algoritma C4.5 akan

melalui proses sebagai barikut: (1) pilih atribut dengan *gain* tertinggi sebagai akar pohon; (2) buat cabang untuk tiap-tiap nilai; (3) bagi kasus dalam cabang; (4) ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama. Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon yang bisa direpresentasikan ke dalam aturan (rule) [6]. Pada penelitian ini, data rekam medis diubah menjadi pohon keputusan untuk menghasilkan klasifikasi penyakit berdasarkan kode penyakit internasional (ICD-10).

2. Metode Penelitian

Penelitian ini difokuskan pada proses menganalisis data rekam medis dengan algoritma C4.5 menggunakan program WEKA (*Tools Data Mining)* untuk memperoleh hasil klasifikasi. Ada 4 atribut yang dipakai dalam penelitian, yaitu: (1) jenis kelamin yang terdiri dari perempuan (P) dan laki-laki (L); (2) umur yang dikelompokkan dalam kategori bayi & anak (umur <15 tahun), muda & dewasa (umur 15-50 tahun), tua (umur >50 tahun); (3) bulan yang terdiri dari Januari, Pebruari, Maret, April, Mei, Juni, Juli, Agustus, September, Oktober, Nopember, Desember; (4) diagnosa (ICD-10) merupakan atribut tujuan yang terdiri dari kelompok penyakit sesuai kode penyakit internasional (ICD-10), 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 [7].

2.1 Tahapan Praproses

Tahapan praproses dilakukan untuk menyeleksi data-data yang memiliki atribut sesuai dengan kebutuhan penelitian. Berdasarkan total data rekam medis RSUD Malinau tahun 2010 sebanyak 2986 data, diperoleh data yang lengkap atributnya sebanyak 2774 data. Pada tahapan praproses ini juga dilakukan pengelompokan umur pasien serta diagnose penyakit sesuai kategori yang telah ditentukan pada rancangan penelitian. Data hasil praproses disimpan dalam format excel CSV sesuai dengan format yang dibutuhkan pada aplikasi Weka.

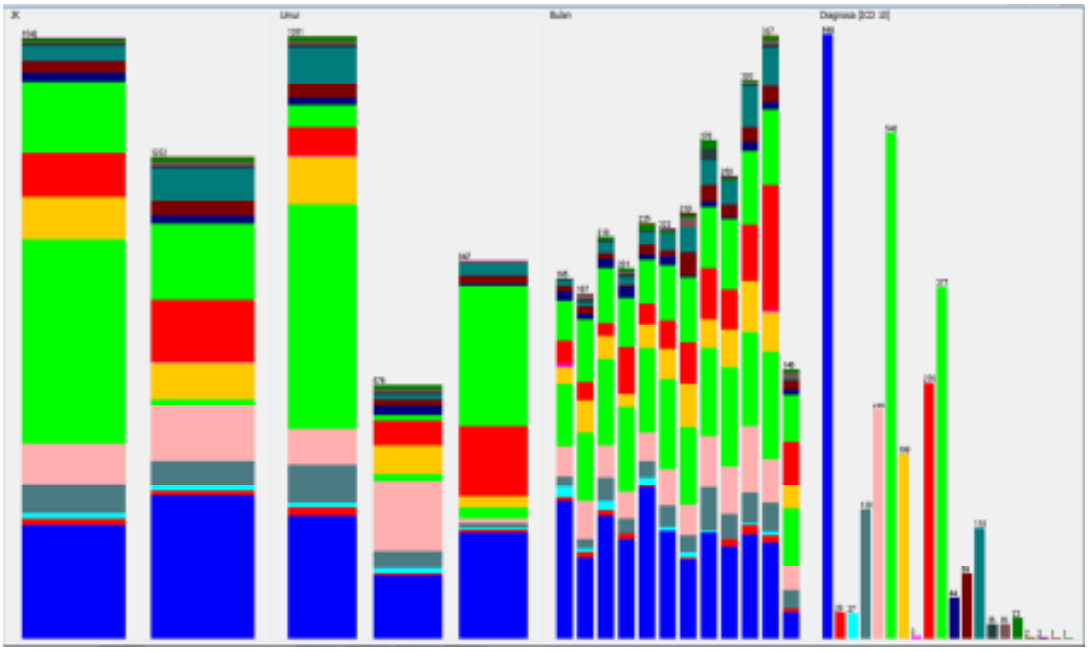
2.2 Tahapan Proses

Tahapan proses merupakan tahapan pelaksanaan penelitian. Pada tahapan ini dilakukan klasifikasi data rekam medis berdasarkan kode penyakit internasional (ICD-10) menggunakan algoritma C4.5 (J48 yang ada pada aplikasi WEKA 3.7.9). Tahapan proses sebagai berikut: (1) memanggil file data rekam medis CSV melalui WEKA untuk melihat visualisasi data secara keseluruhan; (2) menentukan atribut kelas yang menjadi atribut tujuan klasifikasi; (3) melakukan proses klasifikasi dengan memilih algoritma yang digunakan; (4) menampilkan hasil klasifikasi serta rekapitulasi kinerja algoritma.

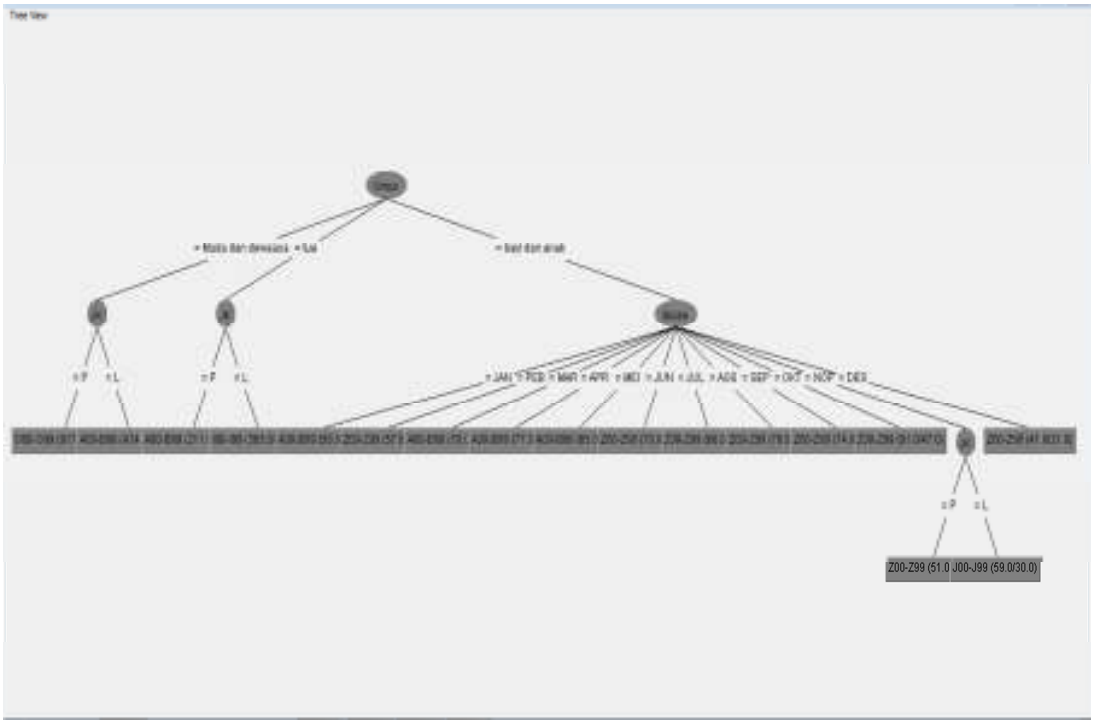
3. Hasil dan Pembahasan

3.1 Hasil Penelitian

Gambar 1 dapat menunjukkan visualisasi data secara keseluruhan yang dijabarkan sesuai atributnya. Selanjutnya data-data tersebut diklasifikasi menggunakan algoritma C4.5 (J48 WEKA) yang menghasilkan pola klasifikasi seperti yang digambarkan dalam Gambar 2.



Gambar 1. Visualisasi Data



Gambar 2. Hasil Pohon Keputusan

Pola klasifikasi yang dihasilkan pada Gambar 2 menunjukkan bahwa ada 5 pola klasifikasi penyakit pada RSUD Malinau, yaitu seperti yang dalam Tabel 1.

Tabel 1 Klasifikasi Penyakit pada RSUD Malinau

Nama Penyakit	Keterangan	
	Kategori	Bulan Kejadian
A00-B99	Umumnya diderita oleh laki-laki dengan kategori umur muda dan dewasa, perempuan dengan kategori umur tua, kategori bayi dan anak	Januari, Maret, April, Mei
I00-I99	Umumnya diderita oleh laki-laki dengan kategori umur tua	
J00-J99	Umumnya diderita oleh laki-laki dengan kategori umur bayi dan anak	November
O00-O99	Umumnya diderita oleh perempuan dengan kategori umur muda dan dewasa	
Z00-Z99	Umumnya diderita oleh bayi dan anak	Pebruari, Juni, Juli, Agustus, September, Oktober, Desember,
	Bayi dan anak dengan jenis kelamin perempuan	Nopember

Rekapitulasi kinerja algoritma pada proses klasifikasi data rekam medis menggunakan aplikasi WEKA 3.7.9. Jumlah data yang dikenali sebesar 41,4924% dari 2774 data, yaitu 1151, sedangkan jumlah data yang tidak dikenali dalam proses sebesar 58,5076% dari 2774 data, yaitu 1623. Kemudian *kappa statistic* diperoleh = 0,2863, *mean absolute error* = 0,0746, *root mean squared error* = 0,194, *relative absolute error* = 86,6521%, *root relative squared error* = 93,5359%.

3.2 Pembahasan Hasil Penelitian

Penelitian ini menunjukkan hasil rekapitulasi algoritma klasifikasi C4.5 dalam aplikasi WEKA kurang maksimal, dimana jumlah data yang dikenali atau terdeteksi hanya 41,4924% dari jumlah keseluruhan data. Hal ini disebabkan karena jumlah kelas tujuan atau label kelas sangat banyak, yaitu 21 label kelas. Hasil penelitian ini didukung oleh beberapa penelitian lain yang dilakukan oleh [8] dan [9] yang menyimpulkan keunggulan algoritma C4.5 dalam mengklasifikasikan data dengan label kelas berjumlah 2 seperti baik, buruk atau juga tinggi, rendah, dan contoh lainnya. Demikian juga label kelas berjumlah 3 seperti baik, cukup, kurang, atau tinggi, sedang, rendah, dan contoh lainnya. Hampir semua penelitian yang menggunakan algoritma C4.5 dengan label kelas sedikit memperoleh akurasi yang baik atau unggul.

Performansi algoritma C4.5 (J48 WEKA) untuk proses klasifikasi data rekam medis berdasarkan kode penyakit internasional dalam penelitian ini juga belum maksimal. Hal ini ditunjukkan pada rekapitulasi kinerja algoritma dimana diperoleh *mean absolute error* = 0,0746, *root mean squared error* = 0,194, *relative absolute error* = 86,6521%, *root relative squared error* = 93,5359%. Menurut penelitian yang dilakukan oleh [3] terhadap algoritma C4.5 menyimpulkan bahwa hasil klasifikasi dari algoritma C4.5 dinyatakan baik jika tingkat akurasi mendekati 1/100 atau 0,001. Pernyataan ini juga mendukung hasil yang diperoleh [10].

4. Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa algoritma C4.5 (J48 dalam aplikasi WEKA) kurang maksimal dalam menghasilkan klasifikasi data rekam medis karena jumlah kelas tujuan atau label kelas sangat banyak (21 label kelas) dan persentasi data yang terbaca kurang dari 50%. Klasifikasi penyakit yang dihasilkan hanya 5 kelas dari 21 kelas keseluruhan sesuai kode penyakit internasional.

DAFTAR PUSTAKA

[1] Sumanthi S. dan Sivanandam S. 2006. ***Introduction to Data Mining and its Applications***. Penerbit Springer.

[2] Anonim. 2008. Permenkes RI. 2008. No. 269 / MENKES/PER/III/2008 tentang **Rekam Medis.**

[3] Kursini dan Luthfi, Emha. 2009. ***Algoritma Data Mining.*** Penerbit ANDI.

[4] Witten, Ian H, dkk. 2011. ***Data Mining Practical Machine Learning Tools and Techniques***. Penerbit Morgan Kaufmann.

[5] Hsu, Hui-Huang . 2006. ***Advanced Data Mining Technologies in Bioinformatics***. Penerbit IDEA Group Publishing.

[6] Liao Shu-Hsien, Chu. Pei-Hui, Hsiao.Pei-Yuan. 2012. ***Data Mining Techniques and Applications – A Decade Review from 2000-2011***. Jurnal ELSEVIER - Expert Systems with Applications 39.

[7] WHO. 2010. ***International Statistical Classification of Diseases and Related Health Problems (ICD-10)***.

[8] Lesmana, I Putu Dody. (2012). ***Perbandingan Kinerja Decision Tree J48 dan ID3 dalam Pengklasifikasian Diagnosis Penyakit Diabetes Millitus***. Jurnal TEKNOMATIKA Vol.2. No.2.

[9] Abidin, Aa Zezen Zaenal. (2011). ***Implementasi Algoritma C4.5 untuk Menentukan Tingkat Bahaya Tsunami.*** SEMNASIF 2011-UPN Veteran. Yogyakarta.

[10] Sakthimurugan T dan Poonkuzhali S. 2012. ***An Effective Retrieval of Medical Record using Data Mining Techniques.*** International Journal of PharmaceuticalScience and Health Care.Vol 2