IMPLEMENTASI TEKNIK *DATA MINING* DIDALAM ANALISIS PENYAKIT DIABETES MELLITUS TIPE II MENGGUNAKAN *DECISION TREE*

Rodiyatul FS[1], Bayu Adhi Tama[2]

[1,2]Fakultas Ilmu Komputer, Universitas Sriwijaya
[1] *abecedeh@yahoo.co.id,* [2]*bayu@unsri.ac.id*

ABSTRACT
Diabetes Mellitus (DM) is a major cause of morbidity and mortality in the modern society. The detection of DM from various factors or symptoms is an issue which is not free from false presumptions accompanied by unpredictable effects. Data mining could be used as an alternative way, especially in knowledge discovery from data. This paper uses C4.5 algorithm in the data mining process. It tests use an open source tool WEKA to acquire information from the analysis of historical data of patient medical records. Finally, it offers a decision-making support on the early detection of DM for physician and other medical decision-makers.

Keywords. *Data Mining, Decision Tree, Diabetes Mellitus (DM).*

I. PENDAHULUAN

Data mining merupakan disiplin ilmu yang baru dan yang sedang berkembang didalam beberapa tahun terakhir ini. Seiring dengan perkembangan teknologi informasi dan komunikasi, teknologi data mining yang digunakan untuk menganalisa volume data yang besar menjadi populer saat ini. Data mining merupakan bidang ilmu yang multidisiplin, termasuk didalamnya adalah sistem basis data, statistik, *machine learning*, visualisasi, and ilmu informasi. Selain itu, berdasarkan jenis datanya sistem *data mining* merupakan integrasi dari teknik-teknik lain seperti analisis data spasial, temu kembali informasi, *pattern recognition*, pemrosesan sinyal, grafika komputer, teknologi Web, ekonomi, bisnis, bioinformatika, atau psikologi [3]

Berdasarkan survei yang telah dilakukan, diperkirakan pada tahun 2020 akan ada 178 juta penduduk berusia diatas 20 tahun memiliki prevelansi terkena DM, suatu jumlah yang besar untuk dapat ditangani sendiri oleh para ahli DM [4]. Tingginya angka-angka statistik diatas, tentunya patut diantisipasi oleh pihak penyedia layanan kesehatan seperti rumah sakit untuk mencegah timbulnya ledakan pasien DM.

Paper ini mencoba untuk menemukan informasi yang berharga dari data menggunakan teknik *data mining* untuk membantu pihak pengambil keputusan di bidang kesehatan dalam memahami *rules* yang mungkin terjadi didalam diagnosa penyakit DM tipe II. Analisis data terhadap *record* pasien di salah satu rumah sakit ini memungkinkan pula diagnosa penyakit yang benar dan tepat.

II. METODE *DECISION TREE*

Konsep klasifikasi dengan pengawasan (*supervised classification*) adalah untuk membangun sebuah model dari data yang telah diketahui, atau sering disebut sebagai *classifier*. Model atau fungsi ini kemudian dapat digunakan untuk memetakan data didalam suatu basis data kepada suatu atribut target, selanjutnya dapat memperkirakan suatu kelas dari data yang baru.

Algoritma *decision tree* merupakan salah satu algoritma klasifikasi didalam *data mining* yang bekerja berdasarkan teori informasi (*information theory*). *Decision tree* memiliki beberapa keunggulan yaitu mudah dalam pengembangan sebuah model, mudah dipahami oleh pengguna, dan mampu menangani *noisy data* dan *unknown data* [3].

Decision tree terdiri dari beberapa bagian yaitu simpul dalam (*inside nodes*), cabang (*branches*), dan simpul daun *(leaf nodes*). Simpul teratas disebut juga simpul akar (*root nodes*); simpul dalam mereprentasikan nilai dari suatu atribut.

III. ALGORITMA C4.5 DI *DECISION TREE*

3.1 *Entropy*

Entropy adalah ukuran ketidakpastian (*uncertainty*) atau kekacuan (*confusion*) dari sebuah sistem, dan kuantitas informasi dari sebuah sistem adalah ukuran tingkat sistemisasinya (*degree of systemization*). *Record* data yang telah jelas diketahui akan

memiliki nilai *entropy* nol; dan ketika hasilnya adalah sebuah variabel, maka nilai *entropy* akan meningkat.

Rumusan *entropy* adalah sebagai berikut:

$$S = -\sum_I (p_i * \log(p_i))$$

3.2 Prinsip Algoritma C4.5

C4.5 merupakan salah satu algoritma yang telah secara luas digunakan, khususnya di area *machine learning* yang memiliki beberapa perbaikan dari algoritma sebelumnya, ID3, yaitu dalam hal metode pemangkasannya (*prunning*) [5]. Adapun perbaikannya adalah sebagai berikut:

- Algoritma C4.5 menghitung *gain ratio* untuk masing-masing atribut, dan atribut yang memiliki nilai yang tertinggi akan dipilih sebagai simpul. Penggunaan *gain ratio* ini memperbaiki kelemahan dari ID3 yang menggunakan *information gain*.
- Pemangkasan dapat dilakukan pada saat pembangunan pohon (*tree*) ataupun pada saat proses pembangunan phon selesai.
- Mampu menangani *continues attribute*.
- Mampu menangani *missing data*
- Mampu membangkitkan *rule* dari sebuah pohon.

Sebagai contoh kita memiliki 2 kelas yaitu *P* dan *N*, sedangkan *x* dan *y* adalah banyaknya *record* dari kelas *P* dan *N* dari suatu atribut *S*, maka *entropy* dari *S* adalah [2]:

$$Info\ (S) = Info(S_p, S_n) = -\left(\frac{x}{x+y} * \log\frac{x}{x+y} + \frac{y}{x+y} * \log\frac{y}{x+y}\right)$$

Kita kemudian mengambil variabel *D* sebagai *root node* dari *decision tree*, dan membagi *S* kedalam beberapa simpul anak (*child nodes*) *{S_1, S_2, ….S_k}* dan masing-masing S_i (*i=1,2,…k*) memasukkan x_i (jumlah kelas *P*) dan y_i (jumlah kelas *N*), maka nilai *entropy* untuk simpul anak adalah:

$$Info(D,S) = \sum_i^k \frac{x_i + y_i}{x+y} * Info(S_{ni}, S_{pi})$$

Information gain didefenisikan sebagai:

$$Gain(D) = Info(S) - Info(A,S)$$

Sehingga dapat kita simpulkan fungsi dari *information gain* adalah :

$$Gain(D,S) = Info(S) - Info\ (D,S)$$

$$Info(S) = I(P) = I(P_1, P_2, \ldots..P_k) = I\left(\frac{|C_1|}{|S|}, \frac{|C_2|}{|S|}, \ldots\frac{|C_k|}{|S|}\right) = -(p_1 * \log p_1 + p_2 * \log p_2 + \cdots + p_k * \log p_k)$$

$$Info(D,S) = \sum_i^k \left(\frac{|S_i|}{|S|}\right) * Info(S_i)$$

IV. ANALISIS

4.1 *Data Preparation*

Data utama yang digunakan pada penelitian ini berupa sekumpulan data rekam medis pasien rawat inap RSMH Palembang untuk penyakit *Diabetes Mellitus* tahun 2008 yang berjumlah 185 *instances.* Sebelum proses *data mining* dimulai, dilakukan *preprocessing* data rekam medis dengan memisahkan *tuples* yang redundan dan atribut yang tidak diperlukan. Setelah melalui proses pengumpulan (*collection*), pembersihan (*cleaning*), dan integrasi (*integration*) dan melalui proses *preprocessing*, maka didapatkan *dataset* seperti pada Tabel 1.

Tabel 1 *Dataset* yang dihasilkan setelah melalui *preprocessing* terhadap *noisy data* dan *disorderly data*.

usia	jk	bb	td	….
Dwasa	LK	Overweight	>=140/90	….
Tua	PR	Kurus	>=140/90	….
Tua	LK	Sedang	>=140/90	….
Tua	PR	Overweight	<140/90	….
Tua	PR	Overweight	>=140/90	….
Tua	LK	Sedang	<140/90	….
Tua	LK	Kurus	>=140/90	….
Tua	LK	Overweight	>=140/90	….
Tua	PR	Overweight	>=140/90	….
….	….	….	….	….

4.2 Aplikasi WEKA

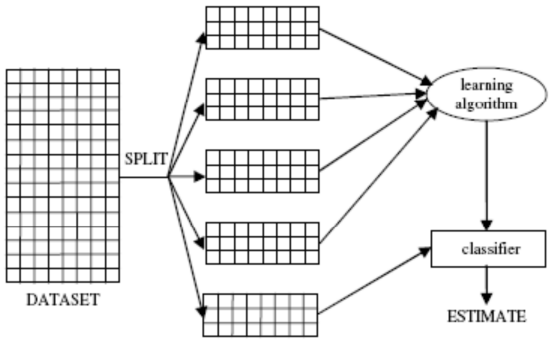
Dalam penelitian ini kami menggunakan WEKA [4], baik dalam tahapan *preprocessing,* pembangunan *decision tree*, dan validasi *classifier* yang dihasilkan. Tampilan seluruh

atribut setelah melewati *prepocessing* dapat dilihat pada Gambar 2 berikut ini.

4.3 Proses *Tranining* Algoritma C4.5

Kami menggunakan algoritma C4.5 untuk membangun sebuah *decision tree* (Gambar 3). Kami juga menggunakan metode *cross validation* untuk menghitung estimasi kesalahan (*error*) dari pohon yang telah dihasilkan [1]. Dengan kata lain, kami memecahkan data secara acak kedalam 10 bagian (*folds*) dan secara berulang, masing-masing *folds* tersebut diperuntukkan sebagai *training data* dan sisanya sebagai *test data* (Gambar 1).

Pada bagaian terakhir, kami bandingkan hasilnya, dan hasil akhir yang memiliki tingkat akurasi yang paling baik yang akan dipilih menjadi *decision tree*. Kesalahan yang dihasilkan selama berlangsungnya proses *training* adalah sebesar 8,1081% dengan tingkat akurasi sebesar 91,8819% (Tabel 2).



Gambar 1. Metode *k-folds cross validation*, dengan *k=10*.

Tabel 2. Hasil untuk 10-*folds cross validation*

Instances	*Correctly classified*	*Incorrectly classified*
185	170 (91,8919%)	15 (8,1081%)

4.4 Hasil dan Pembahasan

Berdasarkan hasil penelitian yang kami peroleh, maka dapat dihasilkan beberapa *rule* yang dapat digunakan dalam mendiagnosa penyakit diabetus mellitus sebagai berikut:

- Plasmainsulin memiliki *gain ratio* yang paling tinggi, oleh karena itu atribut ini paling berpengaruh terhadap penyakit diabetes mellitus tipe II
- Pasien yang memiliki plasmainsulin yang tinggi dan gdpuasa lebih besar atau sama dengan 126, maka risiko akan mengidap penyakit diabetes melitus tipe II lebih besar.
- Pasien yang memiliki plasmainsulin tinggi belum tentu mengidap penyakit diabetes mellitus tipe II selam memiliki gdpuasa kurang dari 126, dan tidak memiliki riwayat keturunan terhadap penyakit ini.
- Berdasarkan penelitian ini, pasien yang memiliki berat badan berlebih dan yang memiliki berat badan sedang walaupun memiliki plasmainsulin yang rendah, relatif mengidap penyakit diabetes mellitus selama memiliki gdpuasa lebih besar atau sama dengan 126 dan memiliki gdsewaktu lebih besar atau sama dengan 200.
- Pasien yang memiliki plasmainsulin rendah dan memiliki gdpuasa yang kurang dari 126, tidak berisiko mengidap penyakit diabetes mellitus tipe II.

V. KESIMPULAN DAN SARAN

Melalui teknik *data mining* yang digunakan, paper ini telah berhasil mengumpulkan dan menganalisa data rekam medis pasien diabetes mellitus tipe II, dan menghasilkan beberapa *rule*s yang dapat digunakan pihak rumah sakit dalam pengambilan keputusan di bidang kesehatan, khususnya dalam mendiagnosa penyakit diabetes mellitus tipe II.

Penelitian lanjutan hendaknya dilakukan dengan menggabungkan metode *decision tree* dengan metode lain seperti *association rules*, *Bayesian*, *Neural Network (NN)* dan *Support Vector Machine* (SVM). Kuantitas data yang dilibatkan juga perlu ditambah, sehingga mampu memberikan hasil yang lebih signifikan.

VI. DAFTAR PUSTAKA

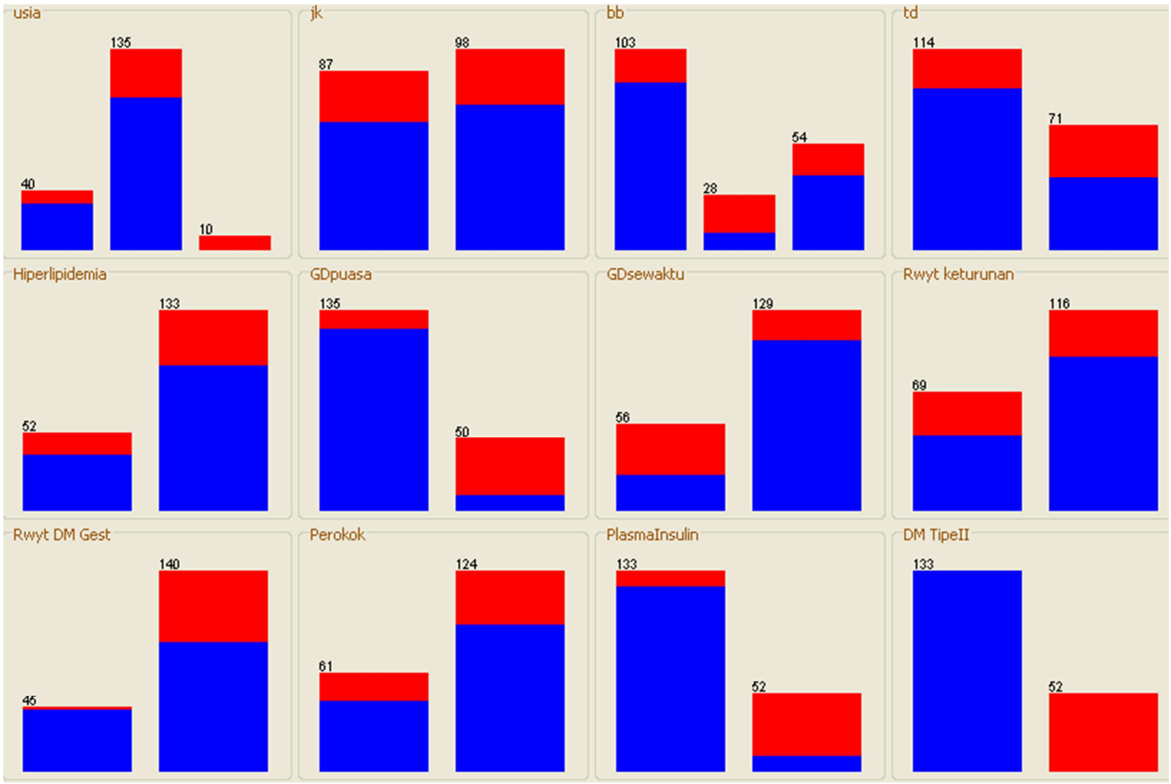
[1] Bramer, Max. Principles of Data Mining, Springer-Verlag London Limited, 2007.

[2] Dong-Peng, Y., et al., *Application of Data Mining Methods in the Evaluation of Client Credibility*, Application of Data Mining in E-Business and Finance, IOS Press, 2008. *pp.35-43.*

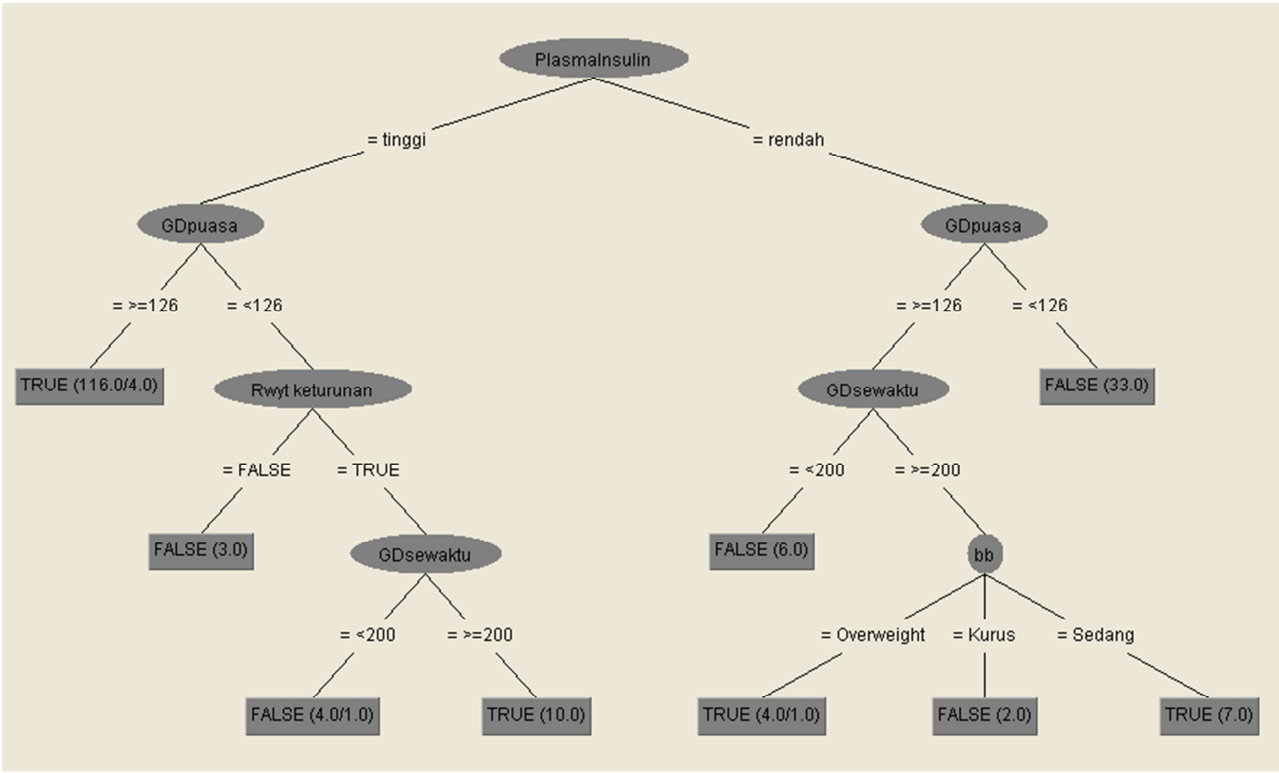
[3] Han, J., et al. Data Mining: Concepts and Techniques 2nd Edition, Morgan Kaufmann Publisher, 2006.

[4] Indah. 2009. [Online] Tersedia: www.indahmuhariani.com/index.php/2009/02/01. [diakses terakhir tanggal 10 Februari 2009]

[5] Witten, Ian H. And Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques 2nd Edition, Morgan Kaufmann Publisher, 2005.



Gambar 2. Visualisasi semua atribut setelah tahapan *preprocessing*.



Gambar 3. *Decision tree* yang dihasilkan dengan algoritma C4.5