PENGEMBANGAN *DECISION TREE J48* UNTUK DIAGNOSIS PENYAKIT DIABETES MELLITUS

I Putu Dody Lesmana
Jurusan Teknologi Informasi, Politeknik Negeri Jember, PO Box 164, Jember 68101
E-mail : dody@polije.ac.id

ABSTRAK

Penelusuran informasi dari data kesehatan merupakan salah satu cara efektif untuk membentuk sistem pendukung keputusan untuk diagnosis penyakit. Salah satunya menggunakan teknik data mining yang bertujuan mengekstraksi dan menemukan pola dari kumpulan informasi yang berharga. Dalam penelitian ini, data mining menggunakan decision tree J48 digunakan untuk memprediksi diagnosis penyakit diabetes mellitus. Kumpulan infomasi yang digunakan berasal dari dataset diabetes Pima Indians yang terdiri dari kumpulan data klinis laboratorium dari pasien yang dinyatakan positif atau negatif menderita diabetes mellitus. Pengolahan data mining dibagi menjadi dua tahap, yaitu tahap pre-processing data yang meliputi identifikasi dan seleksi atribut, penanganan nilai data yang tidak lengkap, dan diskritisasi nilai dan kemudian tahap klasifikasi data menggunakan metode decision tree J48. Efektifitas metode ini diuji menggunakan 10-fold cross validation, dimana dari hasil pengukuran matrik confusion didapatkan akurasi sebesar 74.72%. Hal ini berarti metode decision tree J48 efektif dan dapat digunakan untuk prediksi diagnosis penyakit diabetes mellitus.

Kata kunci : Pima, decision tree, J48, diabetes, atribut, data mining, diagnosis

1. Pendahuluan

Diabetes Mellitus (DM) adalah suatu penyakit yang terjadi akibat kadar glukosa di dalam darah tinggi karena tubuh tidak dapat melepaskan atau menggunakan insulin secara normal. Kadar glukosa darah sepanjang hari bervariasi, meningkat setelah makan dan kembali normal dalam waktu dua jam. Glukosa darah normal pada pagi hari setelah malam sebelumnya berpuasa adalah 70-110 mg/dL. Glukosa darah biasanya kurang dari 120-140 mg/dL pada dua jam setelah makan atau minum cairan yang mengandung gula maupun karbohidrat lainnya. Glukosa darah normal cenderung meningkat secara ringan tetapi progresif setelah usia 50 tahun, terutama pada orang – orang yang tidak aktif beraktifitas. Insulin adalah hormon yang dilepaskan oleh pankreas, merupakan zat utama yang bertanggungjawab dalam mempertahankan kadar glukosa darah yang tepat. Insulin menyebabkan glukosa berpindah ke dalam sel sehingga bisa menghasilkan energi. DM terjadi jika tubuh tidak menghasilkan insulin yang cukup untuk mempertahankan glukosa darah normal atau jika sel tidak memberikan respon yang tepat terhadap insulin.

Terdapat dua tipe DM, yaitu DM tipe 1 yang merupakan diabetes yang tergantung pada insulin, dimana pankreas menghasilkan sedikit insulin atau sama sekali tidak menghasilkan insulin. Sedangkan pada DM tipe 2, pankreas tetap menghasilkan insulin tetapi kadang kadarnya lebih tinggi dari normal dimana kejadian ini akan menyebabkan tubuh membentuk kekebalan terhadap efeknya, sehingga kekurangan insulin relatif. Gejala awal dari DM ini biasanya diawali oleh tiga kondisi, yaitu poliuri (meningkatnya pengeluaran kemih), polidipsi (rasa haus yang berlebihan), dan polifagi (meningkatnya rasa lapar).

Banyak penyandang DM yang terdiagnosis setelah mengalami komplikasi. Padahal, apabila dilakukan diagnosis secara dini, maka penanganan bisa dilakukan lebih cepat dan komplikasi yang membahayakan dapat dihindari. Dalam perkembangan di dunia kedokteran saat ini, para peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi kondisi DM dan mencegah atau menghambat berkembangnya komplikasi. Untuk mendukung hal ini dapat digunakan teknik data mining untuk menggali informasi yang berharga dari kumpulan informasi diabetes. Dalam penelitian ini dilakukan *data mining* dari dataset DM kelompok suku Pima Indians, Amerika Serikat, dimana berdasarkan penelitian yang dilakukan oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) sejak tahun 1965 lebih dari 50% populasinya menderita diabetes tipe 2 dan rata – rata angka kematian akibat DM ini 10 kali lebih besar dibandingkan populasi lainnya di Amerika Serikat [1, 2]. Dataset Pima meliputi sembilan atribut pengukuran dari pasien dengan DM tipe 2 positif dan pasien dengan diagnosis DM negatif. Dalam penelitian ini akan dikembangkan sistem pendukung keputusan untuk diagnosis DM menggunakan *decision tree J48* dengan melakukan ekstraksi informasi dalam bentuk pohon telusur (tree) dari dataset Pima. Untuk mengetahui efektifitas dan akurasi dari pengklasifikasi *decision tree J48* ini maka dilakukan analisa dari matrik *confusion* [3].

2. Metode Penelitian

Penelitian ini terbagi menjadi dua tahap yaitu pertama, tahap *pre-processing* data dan kedua, tahap penyusunan *decision tree J48*. Tahap *pre-processing* data meliputi identifikasi dan pemilihan atribut (*attribute identification and selection*), penanganan nilai atribut yang tidak lengkap (*handling missing values*), dan proses diskritisasi nilai. Sedangkan proses penyusunan decision tree J48 meliputi penyusunan informasi dalam bentuk *tree* menggunakan aplikasi *data mining* Weka [4].

2.1 Tahap *Pre-Processing* Data

2.1.1 Identifikasi dan Pemilihan Atribut

Dataset dalam penelitian ini diambil dari repositori database Pima Indians, UCI [5]. Table 1 menjelaskan atribut dataset diabetes Pima Indians. Dataset Pima ini terdiri dari 768 data klinis yang semuanya berasal dari jenis kelamin wanita dengan umur sekurang – kurangnya 21 tahun. Penggunaan setiap atribut pada dataset Pima ini akan memberikan hasil yang berbeda – beda pada akurasi diagnosis DM [6] dan hal ini berkaitan ada atau tidaknya kelengkapan nilai dari setiap atribut.

Tabel 1: Atribut dataset diabetes Pima Indians

Atribut	*Singkatan*	*Deskripsi*	*Satuan*	*Tipe Data*
Pregnant	Pregnant	Banyaknya kehamilan	-	Numerik
Plasma-Glucose	Glucose	Kadar glukosa dua jam setelah makan	Mg/dL	Numerik
Diastolic Blood-Pressure	DBP	Tekanan darah	Mm Hg	Numerik
Tricepts Skin Fold Thickness	TSFT	Ketebalan kulit	mm	Numerik
Insulin	INS	Insulin	mu U/ml	Numerik
Body Mass Index	BMI	Berat Tubuh	Kg/m^2	Numerik
Diabetes pedigree function	DPF	Riwayat diabetes dalam keluarga	-	Numerik
Age	Age	Umur	Years	Numerik
Class variable	Class	Positif diabetes (1) dan negatif diabetes (0)	-	Nominal

2.1.2 Penanganan Nilai Yang Tidak Lengkap

Dari hasil analisa dataset Pima Indians dapat diketahui bahwa tidak semua atribut memiliki nilai yang lengkap, dimana kelengkapan atribut ini akan menentukan seberapa baik hasil dari pengklasifikasi. Jumlah data tidak lengkap pada masing – masing atribut yaitu atribut *pregnant* sebanyak 110, atribut *glucose* sebanyak 5, atribut DBP sebanyak 35, atribut TSFT sebanyak 227, atribut INS sebanyak 374, atribut BMI sebanyak 11, sedangkan atribut *age* dan *class* memiliki nilai yang lengkap.

Untuk mengatasi nilai yang tidak lengkap pada masing – masing data atribut dapat dilakukan menggunakan empat cara. Pertama, cara termudah dengan menghapus data yang tidak memiliki nilai, tetapi hal ini menyebabkan hilangnya informasi penting pada beberapa atribut yang lain. Kedua, mengganti nilai yang hilang dengan menggunakan nilai rata-rata (*mean*), tetapi cara ini tidak sesuai jika jumlah nilai yang hilang sangat banyak karena akan menyebkan dataset tidak sesuai dengan kondisi sebenarnya. Ketiga, mengganti semua nilai yang tidak ada dengan nilai nol, tetapi hal ini akan menyebabkan hasil klasifikasi yang tidak baik. Cara terakhir adalah dengan mengganti nilai yang tidak ada dengan nilai yang dari tetangga sekelilingnya yang memiliki jarak atau kemiripan terdekat (*eucledian distance*). Salah satu metode yang digunakan adalah *K-nearest neighbor*. Cara keempat ini lebih baik dari cara pertama, kedua, dan ketiga, tetapi masih juga menyebabkan data tidak mencerminkan kondisi sebenarnya jika nilai yang hilang terlalu banyak [7].

Dalam penelitian ini digunakan beberapa kombinasi dari keempat cara diatas untuk mengisi nilai yang tidak lengkap pada masing – masing atribut dengan aturan sebagai berikut:

1. Nilai nol pada atribut *pregnant* dapat diasumsikan bahwa nilai tersebut menyatakan pasien belum pernah melahirkan, sehingga hal ini dimungkinkan sesuai kondisi sebenarnya.
2. Data dengan nilai nol pada atribut *glucose*, DBP, dan BMI dapat dihilangkan karena jumlahnya tidak terlalu banyak sehingga tidak begitu mempengaruhi hasil klasifikasi.

3. Karena atribut TSFT dan INS memiliki jumlah nilai yang tidak ada sangat besar, maka kedua atribut ini tidak mungkin dihilangkan dan tidak mungkin dipakai dalam pengklasifikasian. Oleh karena itu, dalam penelitian ini atribut TSFT dan INS tidak digunakan.

Setelah penerapan proses penanganan nilai yang hilang dilakukan sesuai ketiga aturan yang ditetapkan, maka didapatkan 625 data dari 768 data yang siap diolah lebih lanjut dengan pilihan atribut yaitu *pregnant*, *glucose*, DBP, BMI, DPF, *age*, dan *class*.

2.1.3 Diskritisasi Atribut

Diskritisasi atribut bertujuan untuk mempermudah pengelompokan nilai berdasarkan kriteria yang telah ditetapkan. Hal ini juga bertujuan untuk menyederhanakan permasalahan dan meningkatkan akurasi dalam proses pembelajaran. Atribut *pregnant* dibagi menjadi tiga kelompok, yaitu *low*, *medium*, dan *high* [8]. Atribut *glucose* dibagi menjadi tiga, yaitu *low*, *medium*, dan *high* [6]. Atribut DBP dibagi menjadi tiga, yaitu normal, *normal-to-high*, dan *high* [9]. Sedangkan atribut BMI dikelompokkan menjadi empat, yaitu *low*, normal, *obese*, dan *severely-obese* [9, 10]. Atribut DPF terbagi menjadi dua kelompok, yaitu *low* dan *high*. Pada atribut *age* dibagi menjadi tiga macam, yaitu *young*, *medium*, dan *old.* Atribut *class* dibagi menjadi dua kelompok, yaitu positif DM dan negatif DM. Parameter diskritisasi ditunjukkan secara lengkap pada Tabel 2.

Tabel 2: Diskritisasi atribut dataset diabetes Pima Indians

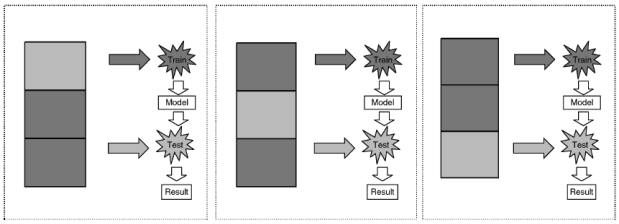
Atribut	*Diskritisasi*
Pregnant	low (0,1), medium (2, 3, 4, 5), high (> 6)
Glucose	low (< 95), medium (95-140), high (> 140)
DBP	normal (< 80), normal-to-high (80-90), high (> 90)
BMI	low (< 24.9), normal (25-29.9), obese (30-34.9), severely-obese (> 35)
DPF	low (< 0.5275), high (> 0.5275)
Age	young (< 40), medium (40-59), old (> 60)
Class	positive (1), negative (0)

2.2 Tahap Penyusunan *Decision Tree J48*

Decision tree J48 merupakan implementasi dari algoritma C4.5 yang memproduksi *decision tree.* Ini merupakan standar algoritma yang digunakan dalam *machine learning*. *Decision tree* merupakan salah satu algoritma klasifikasi dalam *data mining.* Algoritma klasifikasi merupakan algoritma yang secara induktif dalam pembelajaran dalam mengkonstruksikan sebuah model dari dataset yang belum diklasifikasikan (*pre classified dataset*) . Setiap data dari item berdasarkan dari nilai dari setiap atribut. Klasifikasi dapat dilihat sebagai mapping dari sekelompok set dari atribut dari kelas tertentu. *Decision tree* mengklasifikasikan data yang diberikan menggunakan nilai dari atribut [11]. Dataset dengan atribut pilihan pada Tabel 2 kemudian diklasifikasikan menggunakan *decision tree J48.*

2.3 Evaluasi Pengklasifikasi *Decision Tree J48* Menggunakan *K-Fold Cross-Validation*

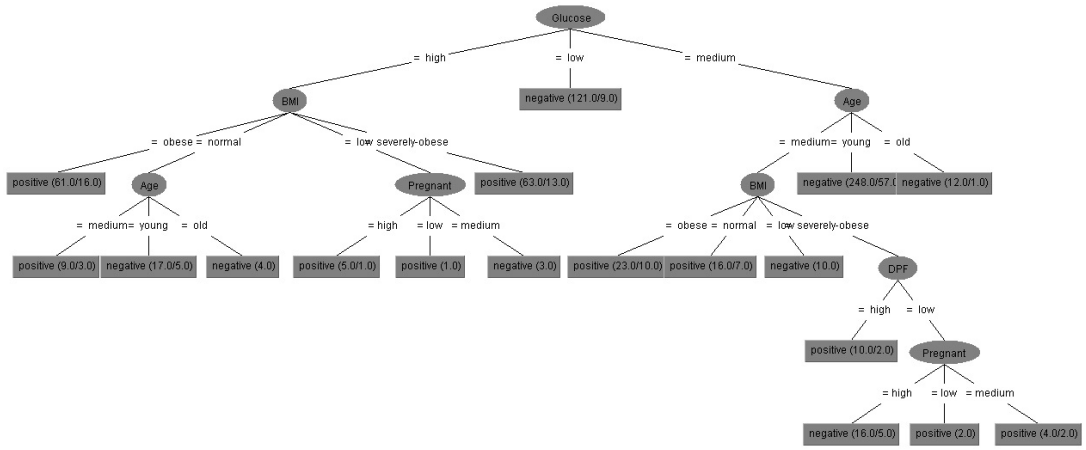
Dalam *k-fold cross-validation*, data pengujian dipisah secara acak ke dalam *k* himpunan bagian yang *mutually exclusive* atau "*folds* (lipatan)", D_1, D_2,..., D_k, yang masing – masing kurang lebih berukuran sama. Pelatihan dan pengujian dilakukan sebanyak *k* kali. Pada iterasi ke-*i*, partisi D_i digunakan sebagai data tes, dan partisi sisanya digunakan bersama untuk melatih model. Dalam iterasi pertama, yaitu himpunan bagian D_2, ..., D_k secara bersama bertindak sebagai data pelatihan untuk memperoleh model pertama, yang diuji pada D_1; iterasi kedua dilatih pada himpunan bagian D_1, D_3, ..., D_k dan diuji pada D_2; dan seterusnya seperti dicontohkan pada Gambar 1. Dalam penelitian ini digunakan *10-fold cross-validation.*



Gambar 1: Ilustrasi *3-fold cross validation*

3. Analisa dan Pembahasan

Dari hasil pengolahan dan uji coba menggunakan *decision tree J48* pada dataset dihasilkan penyusunan informasi dalam bentuk *tree* seperti yang ditunjukkan pada Gambar 2. Dari Gambar 2 dapat diketahui bahwa *glucose* merupakan *root* dari *tree*. Jika *glucose* bernilai *low*, maka hasil klasifikasi menunjukkan negatif yang berarti pasien tersebut tidak mengalami DM. Jika *glucose* bernilai *high* dan pasien mengalami obesitas (BMI *obese*) maka pasien tersebut akan terdeteksi mengalami gangguan DM. Hal ini sama dialami oleh pasien dengan *glucose high* tetapi dengan BMI normal pada usia parobaya (medium) cenderung untuk terkena gangguan DM. Penelusuran informasi dari *tree* ini terus dilakukan untuk cabang – cabang yang lainnya. Informasi nilai yang terletak dalam kurung pada setiap akhir node menunjukkan banyaknya data yang dilatih dan jumlah data yang salah dikenali. Apabila hanya terdapat satu nilai saja menunjukkan bahwa semua data di akhir node tersebut dapat diklasifikasikan dengan baik semuanya. Dari 625 data dengan atribut yang dipilih, 467 data (74.72%) dapat diklasifikasikan dengan benar, sedangkan 158 data (25.28%) salah diklasifikasikan. Hal ini ditunjukkan pada Tabel 3.



Gambar 2: *Decision tree J48* pada diagnosis diabetes untuk dataset Pima Indians

Tabel 3: Matrik confusion dari pengujian decision tree J48 dengan 10-fold cross validation

Hasil Pengujian J48	Gangguan Diabetes		
	Positif	Negatif	
Positif	128	89	217
Negatif	69	339	408
	197	428	625

Dari Tabel 3 diatas dapat dijelaskan bahwa jumlah data pengujian untuk pasien yang diduga menderita diabetes 217 dimana 128 pasien (*true-positive*/TP) terdeteksi dengan benar menderita DM, sedangkan 89 pasien salah diidentifikasi (*false-positive*/FP) oleh pengklasifikasi J48 dimana kondisi sebenarnya pasien tersebut tidak memiliki gangguan DM. Sedangkan pada pengujian pada pasien yang diduga tidak mengalami gangguan DM menunjukkan 339 pasien (*true-negative*/TN) dikenali dengan benar tidak memiliki gangguan DM, sebaliknya terdapat 69 pasien (*false-negative*/FN) salah dikenali sebagai pasien yang tidak mengalami gangguan DM. Dari matrik *confusion* pada Tabel 3 dapat dihitung akurasi dari pengklasifikasi *decision tree J48* mencapai 74.72%. Besarnya kesalahan yang menyebabkan penurunan akurasi terjadi pada kondisi *false-positive*.

4. Kesimpulan dan Saran

Ekstraksi informasi menggunakan *data mining* dari dataset kesehatan sangat efektif sebagai sistem pendukung kesehatan bagi praktisi kesehatan, dimana tujuan dari *data mining* adalah untuk mendapatkan pola informasi yang tersimpan dalam suatu basis data yang dapat digunakan untuk pengolahan selanjutnya dan sebagai bahan pendukung keputusan dalam diagnosis penyakit.

Dalam penelitian ini dilakukan ekstraksi informasi dari dataset diabetes Pima Indians yang digunakan sebagai sistem pendukung keputusan untuk diagnosis penyakit DM. Untuk meningkatkan kualitas data maka dilakukan tahap *pre-processing* data untuk pemilihan atribut, penanganan nilai yang tidak lengkap dan diskritisasi nilai. Selanjutnya, data hasil

tahap *pre-processing* akan digunakan sebagai inputan pada pengklasifikasi *decision tree J48* dimana didapatkan akurasi 74.72% untuk diagnosis gangguan DM.

Untuk meningkatkan akurasi hasil penelitian berikutnya, diharapkan diterapkannya penanganan nilai yang hilang pada data masing – masing atribut dengan memperhatikan keterkaitan nilai antar atribut.

Daftar Pustaka

[1] National Institute of Diabetes and Digeative and Kidney Diseases, "Conquering Diabetes, A Strategic Plan for the 21st Century," National Institutes of Health, U.S. Department of Health and Human Services, NIH Publication No. 99-4398, 1999.
[2] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proceedings of the Symposium on Computer Applications and Medical Care*, IEEE Computer Society Press, pp. 261-265, 1988.
[3] Zhu, W., Zeng, N., & Wang, N., "Sensitivity, specificity, accuracy confidence interval and ROC analysis with practical SAS implementations," Nesug, Health Care and Life Sciences, pp. 1-9, 2010.
[4] WEKA, Machine Learning Group at University of Waikato, diambil dari http://www.cs.waikato.ac.nz/ml/weka/
[5] Pima Indians Diabetes Dataset, UCI Machine Learning Repository, diambil dari http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
[6] Seibel, J. A., Diabetes Guide, WebMD, diambil dari http://diabetes.webmd.com/guide/oral-glucose-tolerance-test
[7] Jayalskshmi, T., Santhakumaran, A., "Impact of preprocessing for diagnosis of diabetes mellitus using artificial neural networks," *Machine Learning and Computing (ICMLC)*, 2010 Second International Conference on , vol., no., pp.109-112, 9-11 Feb. 2010.
[8] Jianchao, H., Rodriguez, J.C., Beheshti, M., "Diabetes data analysis and prediction model discovery using rapidminer," *Future Generation Communication and Networking*, FGCN '08, vol.3, pp.96-99, 2008.
[9] Patil, B.M., Joshi, R.C., Toshniwal, D., "Association rule for classification of type-2 diabetic patients," *Machine Learning and Computing (ICMLC)*, pp.330-334, 2010.
[10] Zelman, K. M., How Accurate is body mass index, or BMI?, diambil dari WebMD, http://www.webmd.com/diet/features/how-accurate-bodymass-index-bmi
[11] Ian H.W., dan Eibe F., "Data mining practical machine learning tools and techniques", Morgan Kaufmann Publishers is an imprint of Elsevier., San Francisco, 2005.