ARTIKEL TUGAS AKHIR

PENENTUAN BESAR AKURASI METODE KLASIFIKASI MENGGUNAKAN ALGORITMA C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION PADA PREDIKSI PENYAKIT DIABETES

Di Susun Oleh :

Nama : Farid Nurhidayat
NIM : A11.2009.05013
Fakultas : Ilmu Komputer
Program Studi : Teknik Informatika-S1

FAKULTAS ILMU KOMPUTER
UNIVERSITAS DIAN NUSWANTORO
SEMARANG
2013

PENENTUAN BESAR AKURASI METODE KLASIFIKASI MENGGUNAKAN ALGORITMA C4.5 BERBASIS PARTICLE SWARM OPTIMIZATION PADA PREDIKSI PENYAKIT DIABETES

Farid Nurhidayat

Program Studi Teknik Informatika - S1, Fakultas Ilmu Komputer,
Universitas Dian Nuswantoro Semarang
URL : http://dinus.ac.id/
Email :
faridnhidayat7@gmail.com

ABSTRAK

Penyakit diabetes adalah salah satu penyakit yang dapat menyebabkan komplikasi bahkan dapat menyebabkan kematian. Saat ini penyakit diabetes semakin lama semakin meningkat jumlah penderitanya. Banyak penelitian yang menggunakan metode support vector machines dalam memprediksi penyakit diabetes tetapi nilai akurasi yang dihasilkan masih kurang akurat. Dalam penelitian ini dibuatkan model algoritma C4.5 dan model algoritma C4.5 berbasis Particle Swarm Optimization untuk mendapatkan rule dalam memprediksi penyakit diabetes dan memberikan nilai akurasi yang lebih akurat. Setelah dilakukan pengujian dengan dua model yaitu Algoritma C4.5 dan C4.5 berbasis Particle Swarm Optimization maka hasil yang didapat adalah algoritma sehingga didapat pengujian dengan menggunakan C4.5 dimana didapat nilai accuracy adalah 73.56 % dan nilai AUC adalah 0.773, sedangkan pengujian dengan menggunakan C4.5 berbasis Particle Swarm Optimization didapatkan nilai accuracy 76.84% dan nilai AUC adalah 0.785 dengan tingkat diagnosa good classification. Sehingga kedua metode tersebut memiliki perbedaan tingkat akurasi yaitu sebesar 3,28% dan perbedaan nilai AUC sebesar 0,012.

Kata Kunci : Diabetes, Algoritma C4.5, Seleksi Atribut, Particle Swarm Optimization

1. PENDAHULUAN

Perkiraan terakhir populasi penderita penyakit diabetes menunjukkan 171 juta orang di dunia pada tahun 2000 dan diperkirakan akan meningkat menjadi 366 juta pada 2030 (Report WHO, 2006). Penyakit diabetes disebabkan oleh peningkatan kadar glukosa dalam darah, apabila kadar glukosa darah meningkat dalam jangka waktu yang lama maka akan menyebabkan komplikasi seperti gagal ginjal, kebutaan dan serangan jantung (Jayalskshmi & Santhakumaran, 2010). Kontrol glukosa darah merupakan hal terpenting dalam praktek medis penyakit diabetes dan penyakit kritis lainnya (Iancu, Iancu, & Sfredel, 2010). Kelainan darah diabetes dan gula lain disebabkan oleh apa yang kita makan dan bagaimana cara kita hidup (Mason, 2005).

Penyakit diabetes merupakan salah satu penyakit yang mematikan, faktor resiko tinggi dalam keluarga yang menyebabkan penyakit diabetes antara lain dikarenakan orang gemuk yang tidak melakukan latihan fisik, dan orang-orang yang memiliki gaya hidup yang tidak sehat dan makanan yang belebih dari apa yang dibutuhkan oleh tubuh (Nuwangi, Oruthotaarachchi, Tilakaratna, & Caldera, 2010). Sehingga untuk menghindari penyakit diabetes diupayakan kita memiliki gaya hidup yang sehat serta tidak makan berlebihan dari apa yang diperlukan oleh tubuh.

Penyakit diabetes perlu diprediksi dengan akurat karena penyakit diabetes merupakan penyakit sosial yang serius dan bisa terkena orang dalam jumlah besar, serta menyebabkan komplikasi dan melibatkan biaya yang tinggi serta dapat meningkatkan keadaan sakit melaui penyakit diabetes terutama pada anak-anak dan anak muda (Iancu, Mota, & Iancu, 2008).

Dewasa ini pendekatan data mining berkembang untuk mengatasi berbagai permasalahan menyangkut tentang pengolahan data. Beberapa peneliti menggunakan teknik *data mining* untuk menyelesaikan permasalahan prediksi (Suhartina & Ernastuti, 2010).

Data mining adalah suatu cara yang bertujuan dalam penemuan pola secara otomatis atau semi otomatis

dari data yang sudah ada di dalam database atau sumber data lain yang dimanfaatkan untuk menyelesaikan suatu masalah melalui berbagai aturan proses(Witten, I.H, 2011). *Data mining* memiliki beberapa teknik, diantaranya klasifikasi dan *clustering*. Teknik klasifikasi adalah teknik pembelajaran yang digunakan untuk memprediksi nilai dari atribut kategori target (Vercellis, 2009). Klasifikasi bertujuan untuk membagi objek yang ditugaskan hanya ke salah satu nomor kategori yang disebut kelas (Max Bramer, 2007). *Clustering* merupakan pengelompokkan objek atau data berdasarkan kemiripan antar data, sehingga anggota dalam satu kelompok memiliki banyak kemiripan dibandingkan dengan kelompok lain (Gorunescu, 2011). Untuk memprediksikan kelulusan mahasiswa, maka hasil pengolahan data akan diklasifikasikan menjadi dua kelas, yaitu tepat dan terlambat. Sehingga teknik klasifikasi paling tepat untuk digunakan dalam *data mining* ini. Metode yang paling populer digunakan untuk teknik klasifikasi adalah *Decision Trees*, *Naïve Bayes Classifiers* (NBC), *Statistical analysis,* dan lain lain (Gorunescu, 2011).

Beberapa penelitian mengenai analisis prediksi penyakit diabetes dengan metode klasifikasi *data mining* telah banyak dilakukan diantaranya adalah yang dilakukan oleh Frisma Handayana pada tahun 2012 yaitu penerapan *particle swarm optimization* untuk seleksi atribut pada metode *support vector machine* untuk prediksi penyakit diabetes. Dalam penelitian tersebut dibuatkan model algoritma *support vector machine* dan model algoritma *support vector machine* berbasis *Particle Swarm Optimization* untuk memberi nilai akurasi yang lebih akurat. Hasilnya model algoritma *support vector machine* berbasis *Particle Swarm Optimization* lebih akurat. Dalam penelitian yang dilakukan oleh Frisma hanya menggunakan satu model algoritma *data mining* yaitu *support vector machine,* jadi belum diketahui nilai keakuratan apabila menggunakan model algoritma lain. *Decision tree* memang populer dan sering digunakan dalam klasifikasi karena

memiliki hasil yang cukup baik jika dibanding algoritma lainnya. C4.5 juga dalam membentuk suatu model pembelajaran dari data tergolong cepat, selain itu karena model digambarkan dalam bentuk diagram pohon maka mudah dipahami. Namun, jika ada data yang tidak relevan dapat menurunkan akurasi C4.5 (Tsai & Chen, 2009, pp. 1-3). Di C4.5 seluruh atribut diseleksi untuk kemudian dibagi menjadi himpunan bagian yang lebih kecil (wu, 2009). Dengan jumlah data yang terlalu banyak, model yang terbentuk menjadi sulit dibaca seperti terbentuknya node yang *redundan*t. Data yang akan diolah sebaiknya dilakukan proses *pre-prosesing* data.

Dibawah ini merupakan beberapa kelebihan dari pohon keputusan (Gorunescu, 2011):

a. Hasil analisa berupa diagram pohon yang sangat mudah dimengerti.
b. Mudah untuk dibangun, serta membutuhkan data percobaan yang lebih sedikit dibandingkan algoritma klasifikasi lainnya.
c. Mampu mengolah data nominal dan kontinyu.
d. Model yang dihasilkan dapat dengan mudah dimengeri, berbeda dengan teknik klasifikasi yang lain seperti neural network yang menyajikan model dengan informasi logis yang tersirat.
e. Menggunakan teknik statistik sehingga dapat divalidasikan.
f. Waktu komputasi relative lebih cepat dibandingkan teknik klasifikasi yang lain.
g. Akurasi yang dihasilkan mampu menandingi teknik klasifikasi yang lainnya.

Salah satu algoritma optimasi yang cukup populer adalah *PSO (Particle Swarm Optimization). PSO* banyak digunakan untuk memecahkan masalah optimasi, serta sebagai masalah seleksi fitur (Liu, Wang, Chen, Dong, Zhu, & Wang, 2011). Algoritma *PSO* terinspirasi dari sekelompok burung yang bergerak secara dinamis kemudian dapat bersinergi serta dapat terorganisir. Ketika diterapkan dalam beberapa kasus untuk mengoptimalisasi algoritma

klasifikasi, mampu meningkatkan akurasi lebih baik daripada *Genetic Algorithm* adalah *PSO*(Sousa, Silva, & Neves, 2004, p. 768).

2. METODE PENELITIAN

Desain penelitian

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

1. Pengumpulan data

   Pada tahap ini ditentukan data yang di proses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan,

   mengintregasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.
2. Pengolahan data awal

   Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan ke bentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.
3. Metode yang diusulkan

   Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.
4. Eksperimen dan pengujian metode

   Pada tahap ini model yang diusulkan akan diuji utuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.
5. Evaluasi dan validasi

   Pada tahap ini dilakukan evaluasi terhadap model yang ditetapkan untuk mengetahui tingkat keakurasian model.

Pengumpulan data

Teknik pengumpulan data ialah teknik atau cara-cara yang dapat digunakan untuk menggunakan data (Riduwan, 2008). Dalam pengumpulan data terdapat sumber data, sumber data yang terhimpun langsung oleh peneliti disebut denga sumber primer, sedangkan apabila melalui tangan kedua disebut

sumber sekunder (Riduan, 2008). Data yang diperoleh adalah data sekunder karena diperoleh dari Pima Indian diabetes database dalam UCI (singkatan dari Pima Diabetes). Masalah yang harus dipecahkan di sini adalah prediksi terjadinya diabetes melitus dalam waktu 5 tahun dengan menggunakan Pima yang berisi 786 orang yang diperiksa dan sebanyak 500 pasien tidak terdeteksi terkena penyakit diabetes, sehingga 268 pasien terdeteksi penyakit diabetes. Dengan atribut dari penyakit diabetes adalah berapa kali hamil, konsentrasi glukosa, tekanan darah, ketebalan lipatan kulit, serum insulin, indeks massa tubuh, diabetes silsilah fungsi dan umur dan kelas sebagai label yang terdiri atas ya dan tidak.

Pengolahan data awal

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 768 data, namun tidak semua data dapat digunakan dan tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*). Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan sebagai berikut (vecelis, 2009):

1. Data validation, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
2. Data *integration and transformationi*, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam software *Rapidminer.*
3. Data size reduction and discritisation, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informative.

Metode yang diusulkan

Pada tahap modeling ini dilakukan pemprosesan data traning sehingga akan membahas metode algoritma yang diuji dengan memasukan data penyakit diabetes kemudian di analisa dan dikomparasi.

Eksperimen dan pengujian metode

Tahap modeling untuk menyelesaikan prediksi penyakit diabetes dengan menggunakan dua metode yaitu algoritma C4.5 dan algoritma optimasi PSO.

1. Algoritma C4.5

Disebut juga dengan *Desicion Tree* adalah pengklasifikasian statistik yang didasarkan pada *Desicion Tree* yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas.

2. *Particle Swarm Optimization*

Yaitu metode optimasi yang melakukan pencarian menggunakan populasi (*swarm*) dari individu (partikel) yang diperbaharui dari iterasi dengan menyeleksi atribut yang ada.

Pada penelitian kali ini yang digunakan adalah penelitian *experiment*. Penelitian eksperimen melibatkan penyelidikan hubungan kausal menggunakan tes dikendalikan oleh si peneliti itu sendiri.

Alat penelitian

Dalam penelitian ini penulis menggunakan spesifikasi software dan hardware sebagai alat bantu dalam penelitian yang tercantum pada tabel 3.10 dibawah ini.

Software	Hardware
Sistem operasi : Windows 7 Ultimate	Prosesor : Intel(R) Core(TM)2Duo CPU P7570 @2.26Ghz 2.26Ghz
Data mining : RapidMiner versi 5.3.008	RAM : 2.00 GB

Validasi dan evaluasi

Dalam tahap ini dilakukan validasi dan pengukuran keakuratan hasil yang dicapai oleh model menggunakan beberapa teknik yang terdapat dalam framework RapidMiner versi 5.3 yaitu confusion matrix dan kurva ROC untuk pengukuran akurasi model, dan cross-validation untuk validasi.

3. HASIL DAN PEMBAHASAN

Tujuan utama penelitian ini adalah untuk mengetahui nilai akurasi dari algoritma C4.5 dan C4.5 berbasis PSO pada pembobotan atribut yang digunakan akan berpengaruh pada hasil pohon keputusan yang terbentuk. Kemudian, berdasar tingkat akurasi dan kurva AUC digunakan untuk membandingkan kedua algoritma tersebut sehingga

dapat diperoleh salah satu algoritma yang terbaik.

Hasil pengujian C4.5

Percobaaan	C4.5		lama waktu eksekusi
	Akurasi	performa AUC	
1	73.56 %	0.773	3 s
2	73.56 %	0.773	3 s

Hasil di atas menunjukan algoritma C4.5 yang diterapkan pada *data set* prediksi penyakit diabetes data menghasilkan nilai akurasi *confusion matrix* sebesar 73.56% dan akurasi AUC 0,773 dalam selang waktu 3 detik.

Hasil pengujian C4.5 berbasis PSO

Percobaaan	C4.5 berbasis PSO		lama waktu eksekusi
	Akurasi	performa AUC	
1	76.84 %	0.785	01:04:50
2	76.84 %	0.785	01:04:50

Hasil di atas menunjukan algoritma C4.5 berbasis PSO yang diterapkan pada *data set* prediksi penyakit diabetes data menghasilkan nilai akurasi *confusion matrix* sebesar 76.84% dan akurasi AUC 0,785 dalam selang waktu 1 jam 4 menit 50 detik.

Setelah melakukan pemodelan dan perhitungan berdasar kedua algoritma diatas, kemudian dilakukan perbandingan hasil yang berupa nilai akurasi dan peforma AUC. Maka diperoleh data perbandingan sebagai berikut :

Perbandingan	C4.5	C4.5-PSO
Akurasi (%)	73.56%	76.84%
performa AUC	0.773	0.785
Waktu eksekusi	3 s	1 jam 4 m 50 s

4. KESIMPULAN

Dalam kesimpulan ini dilakukan pengujian model dengan menggunakan algoritma C4.5 dan C4.5 berbasis *Particle Swarm Optimization* dengan menggunakan data penyakit diabetes yang terkena penyakit atau tidak. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy*, dan AUC dari setiap algoritma sehingga didapat pengujian dengan menggunakan C4.5 nilai *accuracy* adalah 73.56% dengan nilai AUC 0,773 dan C4.5 berbasis *Particle*

Swarm Optimization didapat nilai *accuracy* adalah 76.84% dengan nilai AUC adalah 0,785. Maka dapat disimpulkan pengujian data set diabetes UCI menggunakan algoritma C4.5 berbasis Particle Swarm Optimization akurasi dan nilai AUC lebih tinggi daripada algoritma C4.5 dengan selisih nilai *accuracy* 3.28% dan nilai AUC 0,012.

Saran

Agar penelitian ini bisa ditingkatkan, berikut adalah saran-saran yang diusulkan:

1. Penelitian ini diharapkan dapat digunakan pihak medis sebagai bahan pertimbangan memprediksi penyakit diabetes, sehingga dapat meningkatkan akurasi dalam prediksi penyakit diabetes.
2. Penelitian ini dapat dikembangkan dengan metode optimasi lainnya seperti *Ant Colony Optimization* (AOC), *Genetic Algorhitm* (GA), dan lainnya.
3. Penelitian ini dapat dikembangkan dengan metode klasifikasi data mining lainnya seperti *Naive Bayes*, *KNN* dan lainnya untuk melakukan perbandingan.

5. DAFTAR PUSTAKA

Abraham, A., Grosan, C., & Ramos, V. (2006). *Swarm Intelligence In Data Mining*. Verlag Berlin Heidelberg: Springer.

Larose, D. T. (2007). *Data Mining Methods And Models*. New Jersey: A John Wiley & Sons.

Kusrini, dan Lutfhfy, E T. 2009. "*Algoritma Data Mining*". Yogyakarta: Andi Publishing.

Santosa, B. 2007. "*Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*". Yogyakarta: Graha Ilmu.

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

Mason, R. (2005). *The Natural Diabetes Cure.* Usa: 4th Printing Spring 2012.

Nurrahmani, U. (2012). *Stop!Diabetes Mellitus.* Yogyakarta: Familia.

Nugroho, A. S. (2008). Support Vector Machine: Paradigma Baru Dalam Softcomputing. *Konferensi Nasional Sistem Dan Informatika* , 92-99.

Report Who. (2006). *Definition And Diagnosis Of Diabetes Mellitus And Intermediate Hyperglycemia*. Switzerland: Who Document Production Services.

http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Suhartinah, M.S dan Ernastuti .2010. "*Graduation Prediction of Gunadarma University Student Using Algorithm and Naive Bayes C4.5 Algorithm*". Undergraduate Program, Faculty of Industrial Engineering, Gunadarma University.

Alpaydin, E. (2010). *Introduction To Machine Learning*. London: Massachusetts Institute Of Technology.

Bramer, M. (2007). *Principles Of Data Mining*. Verlag London: Springer.

Dong, Y., Xia., Z., Tu, M., & Xing, G. (2007). An Optimization Method For Selecting Parameters In Support Vector Machines. *Sixth Intenational Coference On Machine Learning And Applications*.

Fei, S. W., Miao, Y. B., & Liu, C. L. (2009). Chinese Grain Production Forecasting Method Based On Particle Swarm Optimization-Based Support Vector Machine. *Recent Patents On Engineering* 2009 , 3, 8-12.

Huang, K., Yang, H., King, I., & Lyu, M. (2008). *Machine*

Learning Modeling Data Locally And Globally. Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh.

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Opimization for Feature Selection. *Journal of Bionic Engineering* Vol 8 , 1-10.

Jiang, Y. (2009). Credit Scoring Model Based on Decision Tree and Simulated Annealing Algorithm. *2009 World Congress on Computer Science and Information Engineering* (hal. 18 - 22). Los Angeles: IEEE Computer Society.

Rapid-I. (2010). *Rapid Miner User Manual* . Rapid-I.

Nuwangi, S., Oruthotaarachchi, C. R., Tilakaratna, J., & Caldera, H. A. (2010). Utilization Of Data Mining Techniques In Knowledge Extraction For Diminution Of Diabetes. *2010 Second Vaagdevi International Conference On Information Technology For Real World Problems* , 3-8.

Vercellis, C. (2009). *Business Intelligence Data Mining And Optimization For*

Decision Making . United Kingdom: A John Wiley And Sons, Ltd., Publication.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine*

Learning Tools And Techniques. Burlington, Usa: Morgan Kaufmann Publishers.