PENERAPAN DATA MINING DALAM MENDIAGNOSIS PENYAKIT DIABETES

Iin Ernawati
Program Studi Sistem Informasi, Fakultas Ilmu Komputer UPN "Veteran" Jakarta
Jl. RS. Fatmawati Pondok Labu Jakarta Selatan - 12450
Telp. 021 7656971 E-mail: iin_ernawati@yahoo.com

Abstract

The hospital's database generally contains a large amount of data with different varieties, but it hasn't been used optimally. It needs a data mining system to process a large amount of data into a strategic valuable information. This article also discusses the use of the data mining to diagnose the diseases, especially diabetes. There are 12 variables used to determine the classification of model formation in this field, those are: age, sex, and some laboratory checking results, including blood glucose and urine glucose. The data mining system can be used to predict whether or not a patient has diabetes and so it can help medical system to prevent it.

Key Words: *data mining, diabetes, classification based association*

PENDAHULUAN

Diabetes adalah suatu penyakit, dimana tubuh penderitanya tidak secara otomatis mengendalikan tingkat gula (glukosa) dalam darahnya. Pada tubuh yang sehat, pancreas melepas hormon insulin yang bertugas mengangkut gula melalui darah ke otot-otot dan jaringan lain untuk memasok energi. Penderita diabetes tidak bisa memproduksi insulin dalam jumlah yang cukup, atau tubuh tidak mampu menggunakan insulin secara efektif, sehingga terjadilah kelebihan gula di dalam darah. Kelebihan gula yang kronis di dalam darah (hiperglikema) ini menjadi racun dalam tubuh.

Diabetes melitus jika tidak dikelola dengan baik akan dapat mengakibatkan terjadinya berbagai penyakit menahun, seperti penyakit jantung koroner, penyulit pada mata, ginjal dan syaraf. Jika kadar glukosa darah dapat selalu dikendalikan dengan baik, diharapkan semua penyakit menahun tersebut dapat dicegah, paling sedikit dihambat.

Dengan bertambahnya angka harapan hidup, perhatian masalah kesehatan beralih dari penyakit infeksi ke penyakit degeneratif. Selain penyakit jantung koroner dan hipertensi, diabetes merupakan salah satu penyakit degeneratif yang saat ini makin bertambah jumlahnya di Indonesia.

Pola prevalensi diabetes telah mengalami pergesaran. Pada awal tahun 1990an umumnya masih tertanam keyakinan bahwa diabetes hanya menyerang mereka yang berusia lanjut, dan merupakan "penyakit orang kaya". Kenyataannya sekarang ini diabetes sudah tidak mengenal perbedaan kelas, diabetes dapat menyerang siapa saja, baik di "gedongan", daerah kumuh, golongan tua maupun muda. Berbagai faktor genetic, lingkungan dan cara hidup berperan dalam perjalanan penyakit diabetes.

Pada dasarnya diabetes dapat dikelompokkan menjadi dua, yaitu diabetes tipe 1 yang terjadi sejak kecil karena cacat sejak lahir, dan diabetes tipe 2 yang berkembang setelah dewasa sebagai akibat gaya hidup yang salah. Diabetes tipe 2 adalah jenis yang paling banyak ditemukan.

Tanpa intervensi yang efektif, jumlah penderita diabetes tipe 2 akan meningkat disebabkan oleh berbagai hal antara lain bertambahnya usia harapan hidup, berkurangnya kematian akibat infeksi dan meningkatnya faktor resiko yang disebabkan karena cara hidup yang salah seperti kegemukan, kurang gerak dan pola makan tidak sehat.

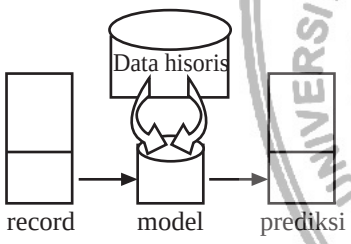
Salah satu alternatif sebagai solusi dari masalah tersebut adalah membuat suatu sistem *data mining* yang bisa melakukan penelusuran pada data historis untuk mengidentifikasi pola dan prediksi trend didasarkan pada sifat-sifat yang teridentifikasi sebelumnya, kemudian memberikan alternatif pengobatan atau pencegahan bila ditemukan indikasi yang mengarah pada timbulnya penyakit diabetes. Informasi yang dihasilkan untuk selanjutnya bisa

digunakan oleh edukator diabetes maupun dokter sebagai dasar untuk melakukan tindakan-tindakan yang diperlukan.

Metodologi Data Mining

Ada dua konsep yang penting dalam *data mining*, yaitu konsep pertama berkaitan dengan mencari pola di dalam data yang biasanya berupa kumpulan data yang sering muncul, tetapi secara umum berupa suatu daftar atau pola data yang muncul lebih sering dari yang diharapkan saat dilakukan secara acak. Konsep yang kedua, adalah *sampling*, yang bertujuan untuk memperoleh keterangan mengenai dengan mengamati hanya sebahagian saja dari populasi itu.

Hal lain yang juga penting yang berhubungan dengan data mining adalah validasi model prediksi yang muncul dari algoritma *data mining*. Model adalah deskripsi dari data historis dimana model tersebut dibangun untuk bisa diterapkan ke data baru dengan tujuan membuat prediksi tentang nilai-nilai yang terputus atau untuk membuat pernyataan tentang nilai yang diharapkan. Pola adalah suatu kejadian atau kombinasi kejadian dalam suatu basis data yang terjadi atau muncul lebih sering dari yang diharapkan.



Gambar 1. Model Proses Pembuatan Data Mining
Sumber: Berson *et al*, 2001

PEMBAHASAN

Association rule merupakan salah satu teknik *data mining* yang paling banyak digunakan dalam penelusuran pola pada sistem pembelajaran *unsupervised*. Association rule menjelaskan kejadian-kejadian yang sering muncul dalam suatu kelompok.

Satu itemset adalah himpunan bagian A dari semua kemungkinan item i. satu itemset yang mengandung i item disebut i-itemset. Prosentase transaksi yang mengandung itemset disebut support. Untuk suatu itemset yang akan diamati, support-nya harus lebih besar atau sama dengan nilai yang dinyatakan oleh user, sehingga itemset tersebut dikatakan sering muncul (frequent).

Bentuk umum aturan asosiasi adalah A1, A2,….,An→B1, B2,…,Bm, yang berarti jika item Ai muncul, item Bj juga muncul dengan peluang tertentu. Misalkan X adalah itemset, transaksi T dikatakan mengandung X jika dan hanya jika X Õ T.

Mining association rule dilakukan dalam dua tahap, yaitu: (1) mencari semua association rule yang mempunyai minimum support Smin dan minimum confidence C_{min}. Itemset dikatakan sering muncul (*frequent)* jika Support(A)≥S_{min}, dan (2) menggunakan itemset yang besar untuk menentukan *association rule* untuk basis data yang mempunyai tingkat kepercayaan C di atas nilai minimum yang telah ditentukan (C_{min}).

Algoritme apriori menghitung seringnya *itemset* muncul dalam basis data melalui beberapa iterasi. Setiap iterasi mempunyai dua tahapan; menentukan kandidat dan memilih serta menghitung kandidat. Pada tahap pertama, himpunan yang dihasilkan dari kandidat *itemset* berisi seluruh i-itemset, yaitu seluruh item dalam basis data. Pada tahap kedua, *algoritme* ini menghitung support-nya mencari melalui keseluruhan basis data yang pada akhirnya hanya i-itemset dengan batas minimum tertentu saja yang dianggap sering muncul *(frequent)*. Sehingga setelah iterasi pertama, seluruh *i-itemset* yang sering muncul akan diketahui. Pada iterasi kedua, *algoritme apriori* mengurangi sekelompok kandidat itemset yang dihasilkan dari iterasi pertama dengan menghapus kandidat *itemset* yang tidak sering muncul. Penghapusan ini berdasarkan pengamatan yaitu apakah *itemset* tersebut sering muncul atau tidak.

Berikut adalah *algoritme apriori* :

```
K=1
    C1=I(semua item)
While Ck>0
(a) Sk=Ck
(b) Ck+1= semua himpunan dg K=1
    elemen yg terbentuk dg
    mengabungkan dua itemset dalam
    sk
(c) Ck+1=Ck+1
(d) S=S+Sk
(e) K++
Return S
```

Metode *classification-based association* yang digunakan adalah CPAR (*Classification based on Predictive Association Rule*). Algoritme ini mengambil ide dari FOIL(*First Order Inductive Lenear*) dalam menghasilkan aturan dan mengintegrasikannya dengan *association classification.*

Ide dasar CPAR berasal dari FOIL yang menggunakan *algoritme greedy* untuk mempelajari aturan yang membedakan contoh positif dengan contoh *negatif*. FOIL secara berulang mencari aturan terbaik dan memindahkan seluruh contoh positif yang

dicakup oleh aturan sampai seluruh contoh positif dalam *data set* tercakup. *Algoritme* FOIL diperlihatkan berikut ini,

```
Masukan : training set D=P » N. (P dan N adlh
himpunan contoh positif & contoh negatif)
Keluaran : himpunan aturan untuk memprediksi label
kelas dari contoh

Procedure FOIL
   Rule set ← fl _
   While |P| > 0
        N' ← N, P'← P
        Rule r fl empty_rule
       While |N| > 0 and r.length < max_rule_length
           Find the literal p that brings most
Gain according to P' and N' append p to r
 Remove from P' all examples not satisfying r
 Remove from N' all examples not satisfying r
   End
   R ← R » {r}
   Remove from p all examples satisfying r's body
    End
Return R
```

Data yang digunakan diambil dari tanggal 1 oktober 2004 sampai dengan 31 desember 2005, meliputi (1) data pasien yang diduga menderita penyakit diabetes berjumlah 9.919 pasien. Pasien yang dipilih adalah pasien dengan minimal satu kali kunjungannya di diagnosa diabetes, pasien yang melahirkan bayi di atas 4.000 gr, pasien dengan tekanan darah tinggi, pasien dengan berat badan yang termasuk ke dalam kategori obesitas, dan (2) data terapi obat serta hasil pemeriksaan laboratorium yang dilakukan pada pasien selama periode 1 oktober 2004 sampai dengan 31 desember 2005 baik yang berasal dari rawat inap.

Persyaratan catatan medis yang dijadikan sampel mengacu pada *international classification of deseases tenth revision* (ICD 10) dimana penyakit diabetes melitus diberi kode E.10 *Insulin-dependent diabetes melitus*, E.11 *Non-insulin-dependent diabetes melitus*, E.12 *Malnutrition-related diabetes melitus*, E.13 *Other specified diabetes melitus* dan *E.14 Unspecified diabetes melitus.*

Dari 9.919 pasien yang diduga menderita penyakit diabetes didapat sebanyak 159.476 record terapi obat dan 211.694 hasil laboratorium. Untuk membentuk *data training* dan *testing*, diambil 10 jenis data pemeriksaan hasil laboratorium yang paling banyak dilakukan yaitu Kolesterol Total (CHOL), Trigliserida (TG), Glukosa Urin Puasa (URN), Aseton Urin Puasa (ACTN), Glukosa Darah Puasa (GLUN), Kolesterol HDL (HDL), Koleslterol LDL (LDL), Glukosa Urin 2 jam PP (UPOST), Aseton Urin 2 jam PP (ACTPP). Terdapat 41.958 *record* hasil pemeriksaan laboratorium yang memenuhi kriteria tersebut.

Banyaknya pasien yang mempunyai catatan lengkap adalah 1.386 orang. Rata-rata umur ± standar deviasi (SD) dari data tersebut adalah 59.29 ± 10.25 tahun. Dari data tersebut diperoleh rasio laki-laki : perempuan adalah 4 : 6.

Tabel 1. Karakteristik Umum Data Pasien

Data	Sex	Jumlah Baris	Prosentase Setiap Kelas	Umur Mean SD
Diabetes	Laki-laki	121	7,37	59,88 ± 8,31
	Perempuan	283	17,24	60,49 ± 8,71
Bukan	Laki-laki	469	28,56	59,20 ± 10,69
Diabetes	Perempuan	769	46,83	58,81 ± 10,72

Tabel 2. Rata-rata Variabel Pemeriksaan Darah

Variabel	Umur Mean ± SD
Glukosa Darah Puasa (mg/dl)	125,33 ± 55,59
Glukosa Darah 2 Jam PP (mg/dl)	171,19 ± 86,50
Glukosa Urin 2 Jam PP	0,74 ± 1,21
Aseton Urin Puasa	0,03 ± 0,23
Glukosa Urin Puasa	0,27 ± 0,80
Aseton Urin 2 Jam PP	0,01 ± 0,16
Kolesterol LDL (mg/dl)	129,37 ± 38,66
Kolesterol HDL (mg/dl)	47,26 ± 12,61
Kolesterol Total (mg/dl)	207,82 ± 45,21
Trigliserida (mg/dl)	159,40 ± 100,58

Penentuan kelas positif diabetes atau negatif diabetes ditentukan oleh diagnosa yang terdapat pada catatan medis. Kelas ditetapkan sebagai positif diabetes jika kode ICD pada diagnosa adalah E.10 Insulin-dependent diabetes mellitus atau E.11 Non-insulin-dependent diabetes mellitus atau E.12 malnutrition-related diabetes mellitus atau E.13 Other specified diabetes mellitus atau E.14 Unspecified diabetes mellitus. Pembentukan nilai kategori selengkapnya sebagai berikut:

Tabel 3. Kategori untuk tabel Sampel Data

Atribut	Nilai Kontinyu	Kategori
Umur	Umur _ 20	1
Umur	20 ≤ umur ≤ 40	2
Umur	Umur _ 40	3
Sex	Sex = laki-laki	4
Sex	Sex = perempuan	5
Glun	Glun _ 70	6
Glun	70 ≤ Glun _ 110	7
Glun	Glun ≥ 110	8
Gpost	Gpost _ 100	9
Gpost	110 ≤ Gpost _ 140	10
Gpost	Gpost ≥ 140	11
Upost	Upost ≤ 0	12
Upost	Upost _ 0	13

Actn	Actn ≤ 0	14
Actn	Actn _ 0	15
Urin	Urin ≤ 0	16
Urin	Urin _ 0	17
Actpp	Actpp ≤ 0	18
Actpp	Actpp _ 0	19
LDL	LDL _ 130	20
LDL	LDL ≥ 130	21
HDL	HDL _ 40	22
HDL	40 ≤ HDL _ 60	23
HDL	HDL ≥ 60	24
Chol	Chol _ 200	25
Chol	Chol ≥ 200	26
Tg	Tg _ 50	27
Tg	50 ≤ Tg _ 150	28
Tg	Tg ≥ 150	29

Data yang sudah dalam bentuk kategori selanjutnya dilakukan proses *cleaning* dengan menghapus baris-baris yang tidak lengkap. Sampel hasil transformasi, integrasi dan cleaning data adalah sebagai berikut:

Tabel 4. Sampel Data Positif Diabetes dan Negatif Diabetes

1	2	3	4	5	6	7	8	9	10	11	12	13
3	5	7	9	12	14	16	18	21	22	26	29	31
3	5	7	11	12	14	16	18	21	23	26	29	31
3	4	8	11	13	14	16	18	21	23	26	29	30
3	5	8	10	12	14	16	18	21	23	26	29	31
3	5	8	11	13	14	17	18	21	24	26	29	30
3	5	7	11	12	14	16	18	20	24	26	29	31
3	5	8	11	13	14	16	18	21	23	26	29	30
3	5	7	10	12	14	16	18	21	23	26	29	31
3	4	8	11	13	14	16	18	20	23	26	29	30
3	5	7	9	12	14	16	18	20	23	26	29	31

Keterangan: 1 = umur; 2 = sex; 3 = Glun; 4 = Gpost; 5 = Upost; 6 = Actn; 7 = Urin; 8 = Actpp; 9 = LDL; 10 = HDL; 11 = Chol; 12 = Tg; 13 = Kelas

Kolom terakhir pada tabel 4 menunjukkan kelas. Angka 30 menunjukkan kelas positif diabetes dan 31 menunjukkan kelas negatif diabetes. Salah satu faktor resiko penyebab adalah orang yang obesitas. Obesitas adalah suatu keadaan dimana ditemukan adanya kelebihan lemak dalam tubuh sehingga bertambahnya berat badan. Ukuran untuk menentukan seseorang mempunyai berat badan lebih digunakan indeks masa tubuh (IMT). Nilai IMT dihitung berdasarkan formula berikut :

$$IMT = \frac{\text{Berat Badan (kg)}}{\text{Tinggi Badan (m)}}$$

IMT normal wanita = 18.5 – 23.5 kg/m2
IMT normal pria = 22.5 – 25 kg/m2

Data training dan *data testing* memuat informasi tentang data input berupa umur, sex, hasil tes laboratorium, data output berupa diagnosa penyakit (positif diabetes, negatif diabetes atau beresiko diabetes). Model yang sudah terbentuk dibandingkan dengan hasil pengujian pada *data testing*. Semakin sama hasil perbandingan output kedua model berarti model semakin akurat.

Pengujian dilakukan terhadap data training dengan mengambil data sampel contoh positif dan contoh negatif. Pada kedua kelompok data tersebut algoritma CPAR digunakan untuk mencari pola-pola dari nilai yang diprediksi. Selanjutnya model diperbaiki dengan menggunakan sampel data lain agar tidak hanya bisa bekerja dengan data training.

Proses pelatihan data mining dilakukan dengan mengambil sebanyak 700 sampel, dengan perbandingan sampel positif diabetes dan negatif diabetes 4 : 6. 700 sampel data yang mempunyai catatan medis lengkap dikumpulkan. Rata-rata umur 57 ± 17 tahun. 51.14% pada kelompok usia di atas 60 tahun, 34.37% pada kelompok usia 51-60 tahun, 11.39% pada kelompok usia 41-50 tahun dan 3.11% pada kelompok usia di bawah 40 tahun. 17.6% diantaranya mempunyai hubungan keluarga.

Tabel 5. Karakteristik Umum Data Training

Data	Sex	Jumlah Baris	Prosentase Setiap Kelas	Umur Mean ± SD
Positif Diabetes	Laki-laki	93	13,29	58 ± 14
	Perempuan	207	29,57	56 ± 16
Negatif Diabetes	Laki-laki	138	19,71	54 ± 23
	Perempuan	262	37,43	59 ± 14

Proses optimalisasi dilakukan dengan merubah kategori Glun dan Gpost yang menjadi penentu utama positif diabetes atau negatif diabetes. Berikut ini adalah hasil proses optimalisasi.

Tabel 6. Kategori untuk Sampel Data Setelah Proses Optimalisasi

Atribut	Nilai Kontinyu	Kategori
Umur	Umur _ 20	1
Umur	20 ≤ umur ≤ 40	2
Umur	Umur _ 40	3
Sex	Sex = laki-laki	4
Sex	Sex = perempuan	5
Glun	Glun _ 70	6

Glun	70 ≤ Glun _ 110	7
Glun	Glun ≥ 110	8
Glun	Glun ≥ 140	9
Gpost	Gpost _ 100	10
Gpost	100 ≤ Gpost _ 140	11
Gpost	140 ≤ Gpost _ 200	12
Gpost	Gpost ≥ 200	13
Upost	Upost ≤ 0	14
Upost	Upost _ 0	15
Actn	Actn ≤ 0	16
Actn	Actn _ 0	17
Urin	Urin ≤ 0	18
Urin	Urin _ 0	19
Actpp	Actpp ≤ 0	20
Actpp	Actpp _ 0	21
LDL	LDL _ 130	22
LDL	LDL ≥ 130	23
HDL	HDL _ 40	24
HDL	40 ≤ HDL _ 60	25
HDL	HDL ≥ 60	26
Chol	Chol _ 200	27
Chol	Chol ≥ 200	28
Tg	Tg _ 50	29
Tg	50 ≤ Tg _ 150	30
Tg	Tg ≥ 150	31

Beberapa aturan yang dihasilkan setelah proses optimalisasi dengan Gain similarity ratio 99%, 80%, 50%, 20% dan 10%.

Tabel 7. Aturan setelah Proses Optimalisasi dengan Gain Similarity Ratio 99%

No.	Aturan	Laplace Accuracy
1	If Gpost ≥ 200 then positif diabetes	0,74
2	If Glun ≥ 140 then positif diabetes	0,73
3	If Upost _ 0 then positif diabetes	0,70
4	If 70 ≤ Glun _ 110 then positif diabetes	0,82
5	If Upost ≤ 0 then negatif diabetes	0,72
6	If 100 ≤ Gpost _ 140 then neagtif diabetes	0,84
7	If Gpost _ 100 then negatif diabetes	0,95

Data pada tabel 7 memperlihatkan bahwa aturan If Gpost >= 200 then positif diabetes mempunyai Laplace Accuracy tertinggi pada kelas positif diabetes yaitu 74%. Ini berarti bahwa pasien dengan hasil pemeriksaan glukosa darah 2 jam pp (Gpost) lebih besar atau sama dengan 200 mempunyai peluang terkena penyakit diabetes sebesar 74%. Aturan If Gpost < 100 then negatif diabetes mempunyai Laplace Accuracy paling tinggi pada kelas negatif diabetes yaitu 95%. Ini berarti pasien dengan hasil pemeriksaan glukosa darah 2 jam pp (Gpost) kurang dari 100 mg/dl mempunyai peluang tidak terkena penyakit diabetes sebesar 95%.

Tabel 8. Aturan setelah Proses Optimalisasi dengan *Gain Similarity Ratio 80%*

No.	Aturan	Laplace Accuracy
1	If Glun ≥ 140 AND If Gpost then positif diabetes	0,74
2	If Upost _ 0 then positif diabetes	0,70
3	If 70 ≤ Glun _ 110 then negatif diabetes	0,82
4	If Upost ≤ 0 then negatif diabetes	0,72
5	If 100 ≤ Gpost _ 140 AND If Gpost _ 100 then negatif diabetes	0,95

Tabel 9. Aturan setelah Proses Optimalisasi dengan Gain Similarity Ratio 50%

No.	Aturan	Laplace Accuracy
1	If Glun ≥ 140 AND If Gpost then positif diabetes	0,74
2	If Upost _ 0 then positif diabetes	0,70
3	If 70 ≤ Glun _ 110 then negatif diabetes	0,82
4	If Upost ≤ 0 then negatif diabetes	0,72
5	If 100 ≤ Gpost _ 140 AND If Gpost _ 100 then negatif diabetes	0,95

Tabel 10. Aturan setelah Proses Optimalisasi dengan Gain Similarity Ratio 20%

No.	Aturan	Laplace Accuracy
1	If Glun ≥ 200 AND If Glun ≥ 140 then positif diabetes	0,74
2	If Upost _ 0 then positif diabetes	0,70
3	If Urin _ 0 then positif diabetes	0,69
4	If Tg ≥ 150 then positif diabetes	0,51
5	If 110 ≤ Glun -140 then positif diabetes	0,60
6	If Upost ≤ 0 AND If 70 ≤ Glun _ 110 then negatif diabetes	0,77
7	If 100 ≤ Gpost _ 140 then negatif diabetes	0,84
8	If Gpost < 100 then negatif diabetes	0,95
9	If Urin ≤ 0 then negatif diabetes	0,62
10	If 50 ≤ Tg _ 150 then negatif diabetes	0,63

Tabel 10 menunjukkan bahwa dengan menurunkan nilai Gain similarity ratio menjadi 20% menghasilkan aturan yang sebelumnya tidak muncul, yaitu If Urn > 0 then positif diabetes dengan akurasi 0.69, If Tg>=150 then positif diabetes dengan akurasi 0.51, If Urn <=0 then negatif diabetes dengan akurasi 0.62, dan If 50<=Tg<150 then negatif diabetes dengan akurasi 0.63.

Tabel 11. Aturan setelah Proses Optimalisasi dengan Gain Similarity Ratio 10%

No.	Aturan	Laplace Accuracy
1	If Glun ≥ 140 AND If Glun ≥ 200 then positif diabetes	0,74
2	If Upost _ 0 then positif diabetes	0,70
3	If Urin _ 0 then positif diabetes	0,69
4	If Tg ≥ 150 then positif diabetes	0,51
5	If 110 ≤ Glun _ 140 then positif diabetes	0,60
6	If 140 ≤ Gpost _ 200 then positif diabetes	0,50
7	If 70 ≤ Glun _ 110 AND If Upost ≤ 0 then negatif diabetes	0,77
8	If 100 ≤ Gpost _ 140 then negatif diabetes	0,84
9	If Gpost < 100 then negatif diabetes	0,95
10	If Urin ≤ 0 then negatif diabetes	0,62
11	If 50 ≤ Tg _ 150 then negatif diabetes	0,63

Tabel 11 memperlihatkan bahwa dengan menurunkan nilai Gain similarity ratio menjadi 10% ternyata tidak menambah jumlah kategori dalam pembentukan aturan.

SIMPULAN

Data mining menggunakan algoritma CPAR dapat digunakan untuk membantu mendiagnosis penyakit diabetes. Algoritma CPAR hanya memilih kategori yang memiliki nilai Gain terbaik, sehingga ada kemungkinan kategori yang mempunyai kekuatan prediksi yang tinggi tidak muncul dalam aturan. Algoritma CPAR menerima input dalam bentuk kategori, sehingga proses penentuan data kontinyu menjadi data kategori sangat berpengaruh terhadap hasil prediksi.

Pemeriksaan glukosa darah 2 jam pp (Gpost), glukosa urin 2 jam pp (Upost), glukosa darah puasa (Glun) menjadi penentu utama untuk menentukan apakah pasien positif diabetes atau negatif diabetes.

DAFTAR PUSTAKA

Berry MJ & Linoff GS, 2000, *Mastering Data Mining: The art and science of Customer Relationship Management,* New York : John Wiley & Sons, Inc.

Berson A., Smith S & Thearling K, 2001, *Building Data Mining Application for CRM,* mcGraw-Hill

Coenen F, 2004, The LUCS-KDD *Implementations of CPAR (Classification based on Prediction Association Rules), Department* of Computer Science The University of Liverpool.

Han J & Kamber M, 2001, Data Mining Concepts and Techniques, The Morgan Kaufmann Publisher.

Herwanto, 2006, *Pembangunan Sistem Data Mining Untuk Diagnosis Penyakit Diabetes Menggunakan Algoritma Classification Based Association,* SPS-IPB.

Kelling, D.G, J.A, Wentworth, at al, 1997, *Diabetes mellitus. Using a database to implement a systematic management program,* NC Med J 58(5) : 368-371

Lanny. S, Alam .S & Iwan H, 2004, *Diabetes: Informasi Lengkap Untuk Penderita dan Keluarga.*

Michel, C & C. Beguin, 1994, *Using a database to query for diabetes mellitus,* Stud Health Technol Inform 14: 178-182

Smith, J.W, J.E Everhart et al, 1998, *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,* Proceeding of the symposium on Computer Applications on Medical care, IEEE Computer Society Press : 261-265

Soegondo S, Soewondo P & Subekti I, 2002, *Penatalaksanaan Diabetes Mellitus Terpadu,* Jakarta : Balai Penerbit FK-UI.

Yin X & Han J, 2003, CPAR: *Classification based on Predictive Association Rules,* University of Illinois at Urbana-Champaign.