Prediksi Penyakit Diabetes Menggunakan Algoritma ID3 dengan Pemilihan Atribut Terbaik *(Diabetes Prediction using ID3 Algorithm with Best Attribute Selection)*

Muhamad Subhan Efendi[1], Helmie Arif Wibawa[2]

Departemen Ilmu Komputer/ Informatika, Fakultas Sains Matematika, Universitas Diponegoro
Jalan Prof. H. Soedarto, SH.Tembalang, Semarang 50275

[1]fendysubhan@gmail.com
[2]Helmie.arif@gmail.com

Abstrak – Penyakit diabetes atau sering disebut dengan penyakit kencing manis adalah suatu penyakit gangguan metabolik menahun yang ditandai oleh kadar glukosa dalam darah yang melebihi nilia normal. Penyakit diabetes sering disebut sebagai *silent killer* dengan mengacu pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes sampai diketahui sudah kronis. Hal ini memicu peningkatan jumlah penderita diabetes dari tahun ke tahun. Maka dari itu penelitian ini mencoba menerapkan suatu metode klasifikasi *Data Mining* untuk memprediksi apakah seseorang terkena penyakit diabetes atau tidak. Algoritma yang digunakan adalah algoritma *Decision Tree* ID3 dengan bantuan seleksi atribut dalam pemilihan atribut yang digunakan. Algoritma seleksi atribut yang dimaksud adalah *Correlation based Feature Selection* (CFS) dan *Information Gain*. Berdasarkan hasil penelitian ini diperoleh bahwa performa tertinggi dicapai ketika algoritma ID3 menggunakan 5 atribut yaitu *gpost*, *glun*, *upost*, *urn*, dan *actn*. Dimana kelima atribut tersebut diperoleh menggunakan algoritma *Correlation based Feature Selection* (CFS) dengan nilai rata-rata akurasi sebesar 84.77, nilai rata-rata *sensitifity* sebesar 87.18, dan nilai rata-rata *specificity* sebesar 82.37.

Kata Kunci – Penyakit Diabetes, Data Mining, ID3, Seleksi Atribut

Abstract – Diabetes is a chronic metabolic disease disorder characterized by levels of glucose in the blood that exceeds normal value. Diabetes is often called as a silent killer with reference to many who do not realize that he was exposed to diabetes until it is said to be chronic. This cause an increase of number of diabetics from year to year. Therefore, this research tried to apply a classification method of Data Mining to predict whether a person is exposed to diabetes or not.. The algorithm used in this research is ID3 Decision Tree algorithm with attribute selection to select attributes. That attribute selection algorithm is Correlation based Feature Selection (CFS) and Information Gain. Based on results of this research, it is found that the highest performance is obtained when the ID3 algorithm uses 5 attributes namely gpost, glun, upost, urn, and actn. That attributes are obtained using Correlation based Feature Selection (CFS) algorithm with an average accuracy is 84.77, average sensitivity is 87.18, and average of specificity is 82.37.

Keyword – Diabetes, Data Mining, ID3, Attribute Selection

I. PENDAHULUAN

International Diabetes Federation (IDF) pada tahun 2013 membuat estimasi bahwa jumlah pengidap diabetes di dunia mencapai 382 juta orang. Diperkirakan dari 382 juta orang tersebut, sekitar 175 juta dia antaranya belum terdiagnosa, sehingga terancam berkembang tanpa disadari dan tanpa pencegahan. Jumlah tersebut diperkirakan akan naik menjadi 592 juta orang pada tahun 2035 [1].

Di Indonesia sendiri jumlah penderita diabetes cukup tinggi, yaitu sekitar 12 juta orang pada tahun 2013. Jumlah tersebut ternyata meningkat daripada tahun-tahun sebelumnya. Pada tahun 2007-2013, Riskesdas (Riset Kesehatan Dasar) melakukan survei untuk menghitung proporsi penderita diabetes untuk usia 15 tahun ke atas. Survei diambil dari data orang yang pernah didiagnosa menderita penyakit diabetes oleh dokter dan yang belum pernah didiagnosa oleh dokter tetapi dalam 1 bulan terakhir mengalami gejala-gejala awal diabetes. Hasil survei tersebut mendapatkan jumlah penderita diabetes pada tahun 2013 meningkat dua kali lipat dibandingkan tahun 2007 [1].

Peningkatan jumlah penderita diabetes dikarenakan diabetes dikenal sebagai *silent killer*. Hal ini mengacu

pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes. Penderita biasanya diketahui terjangkit penyakit ini ketika sudah terjadi komplikasi tanpa adanya penanganan di awal [1]. Untuk menekan jumlah penderita penyakit diabetes yang semakin bertambah, bisa dilakukan deteksi dini yang dapat dilakukan oleh tenaga ahli.

Untuk melakukan deteksi dini penyakit diabetes, dapat dikembangkan suatu sistem untuk memprediksi penyakit dengan memanfaatkan berbagai metode. Salah satu metode yang dapat digunakan yaitu metode *data mining* dengan prinsip klasifikasi. Metode ini dapat mengolah data dalam jumlah besar yang nantinya digunakan untuk mendapatkan hasil prediksi. Seperti yang telah diterapkan dalam penelitian sebelumnya dengan menggunakan algoritma C4.5 [2].

Salah satu metode yang dapat digunakan dalam klasifikasi data mining adalah *decision tree* (pohon keputusan). Metode ini telah diterapkan untuk memprediksi tingkat kelulusan mahasiswa [3], untuk memprediksi loyalitas pelanggan [4], serta di bidang medis pernah diterapkan untuk memprediksi penyakit kanker payudara [5].

Pada penelitian sebelumnya digunakan *data mining* yaitu algoritma *Naive Bayes* untuk memprediksi ketepatan kelulusan mahasiswa [6] dan algoritma ID3 untuk memprediksi penyakit diabetes [7]. Akurasi dari algoritma ID3 untuk memprediksi diabetes berada pada kisaran 63. Sedang untuk algoritma *Naive Bayes* yang digunakan untuk memprediksi ketepatan kelulusan mahasiswa mendapatkan nilai akurasi yang cukup tinggi setelah pemilihan atribut dengan information gain dengan 3 atribut, yaitu pada kisaran 89.79 yang pada awalnya menggunakan 13 atribut didapat akurasi sebesar 83.07. Disini dapat dilihat bahwa pemilihan atribut yang digunakan dapat mempengaruhi hasil prediksi. Pada kasus lain, dilakukan perbandingan stabilitas penggunaan beberapa algoritma seleksi atribut pada beberapa algoritma *classifier* dengan hasilnya didapatkan algoritma *Correlation based Feature Selection* (CFS) merupakan algoritma yang paling stabil dan mendapatkan nilai akurasi yang lebih tinggi [8]. Dari hal ini dapat diterapkan algoritma seleksi atribut *Information Gagin* dan CFS untuk meningkatkan performa dari algoritma ID3.

Berdasar permasalahan tersebut maka dapat dibuat suatu penerapan algoritma ID3 untuk memprediksi penyakit diabetes menggunakan algoritma seleksi atribut. Algortima seleksi atribut yang dapat diterapkan yaitu *Information Gain* atau *Correlation based Feature Selection*.

A. *Diabetes*

Penyakit diabetes atau sering disebut dengan penyakit kencing manis adalah suatu penyakit gangguan metabolik menahun yang ditandai oleh kadar glukosa dalam darah yang melebihi nilai normal [9].

B. *Data Mining*

Data Mining merupakan proses menambang sebuah data yang berukuran besar untuk menggali nilai tambah dari data tersebut yang nantinya sangat berguna untuk pengembangan. Keluaran dari *data mining* ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan [10].

1) *Data cleaning*
2) *Data Integration*
3) *Data transformation*
4) *Data selection*
5) *Data mining*
6) *Pattern evaluation*
7) *Knowledge presentation*

C. *Imbalance Data*

Imbalance data adalah kasus khusus untuk masalah klasifikasi dimana distribusi kelas tidak sama di antara kelas. Biasanya, ada dua kelas yaitu mayoritas dan minoritas.

Metode yang dapat digunakan dalam penanganan *imbalance data* adalah *undersampling* yaitu dengan mengambil beberapa data mayoritas sehingga jumlah data mayoritas sama besar jumlahnya dengan jumlah data minoritas. Salah satu metode *undersampling* yaitu *cluster based undersampling*, yang merupakan metode *undersampling* dengan menggunakan *cluster* data untuk mendapatkan nilai ratio data yang digunakan di tiap *cluster* [11].

Berikut merupakan algoritma *cluster based undersampling* yang digunakan pada penelitian ini [12]:

1) Tetapkan nilai k, untuk proses k-means clustering

2) Masukkan data (X)

3) Lakukan proses *k-means clustering* yang menghasilkan *centroid D* berukuran $k\ x\ n$

4) Kerjakan i=1 sampai m

a) Kerjakan c=1 sampai k

i. Hitung jarak euclidean

$$s(i,c) = \sqrt{\sum_{j=1}^{n}(x_{ij} - D_{cj})}$$

5) Kerjakan g=1 sampai sum_g-1
   a) Cari rasio data mayoritas kelas ke-g dengan data minoritas
      $V = \frac{w_{gc}}{w_{1c}}$
   b) Hitung sementara jumlah rasio data mayor kelas ke-g
      $u_g = u_g + V$

6) Proses pengambilan data mayoritas yang akan diambil pada setiap kluster
   a) Kerjakan g=1 sampai sum_g-1
      i. Kerjakan c=1 sampai k
      $l_{gc} = m * num_{minor} * \frac{V_{gc}}{u_g}$

D. Seleksi Atribut

Seleksi atribut merupakan suatu proses pemilihan atribut yang dianggap relevan dalam proses data mining.

1) Correlation based Feature Selection (CFS). CFS adalah teknik heuristic untuk mengevaluasi nilai atau harga subset atribut. Teknik ini mempertimbangkan kegunaan atribut individual bagi prakiraan label kelas dengan level interkorelasi di antara atribut-atribut [13].

Algoritma CFS yang digunakan [13]:

a) Hitung nilai *coeffisien correlation* dengan *Symmetrical Uncertainty*
   $SU = 2.0 * \left(\frac{Gain}{H(X) + H(Y)}\right)$
   Dimana :
   H= entropi atribut
   X= atribut 1
   Y= atribut 2

b) Hitung nilai *Merits*
   $Merits = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$
   k =jumlah atribut
   $\overline{r_{cf}}$ =*correlation class-attribute*
   $\overline{r_{ff}}$ =*intercorrelation attribute-attribute*

c) Lakukan perhitungan Merits untuk semua kemungkinan kombinasi pemilihan atribut untuk menemukan nilai Merits tertinggi.

2) Information Gain. Information Gain diperoleh dari nilai entropi sebelum pemisahan dikurangi dengan nilai entropi setelah pemisahan.

Nilai *Information Gain* pada penelitian ini dihitung dengan algoritma sebagai berikut [14]:

a) Hitung entropi kasus *D*
   $Info(D) = -\sum_{i=1}^{m} p_i * log_2\,(p_i)$
   Dimana :
   Info(D)= nilai entropi kasus *D*
   i= indeks kelas
   m= jumlah kelas
   p_i=probabilitas kemunculau kelas i

b) Hitung entropi kasus *D* atribut *A*
   $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j)$
   Dimana :
   $info_A$(D)= entropi atribut *A* pada kasus *D*
   j= indeks kategori
   v= jumlah kategori dalam atribut
   D_j= jumlah sampel kategori *j*
   D= jumlah seluruh sampel data
   $Info(D_j)$ = entropi sampel untuk kategori *j*

c) Hitung *Information Gain* atribut *A*
   $Gain(A) = Info(D) - Info_A(D)$

E. Decision Tree ID3

ID3 adalah algoritma *decision tree learning* (algoritma pembelajaran pohon keputusan) yang paling dasar dan dikembangkan oleh J. Ross Quinlan. Algoritma ini melakukan pencarian secara rakus/menyeluruh (*greedy*) pada semua kemungkinan pohon keputusan [15].

Algoritma ID3 yang digunakan dalam penelitian ini adalah sebagai berikut [16] :

1) Hitung nilai entropi
   $Entropi(S) = \sum_{i=1}^{k} -p_i\ log_2\, p_i$
   Dimana :
   S =himpunan (*data set*) kasus
   k =jumlah kelas
   i =indeks kelas
   p_i=probabilitas kemunculan kelas *i*

2) Hitung nilai *gain*
   $Gain\ (S, A) = En\ (S) - \sum_{j=1}^{m} \frac{|S_j|}{|S|}\ x\ En\ (S_j)$
   Dimana :
   S=ruang (data) untuk *training*.
   A=atribut.
   $|S_j|$=jumlah sample kategori *j*
   $|S|$=jumlah seluruh sample data.
   $En\ (S_j)$=entropi untuk katgori *j*
   m=banyak kategori pada atribut *A*
   j=indeks kategori

3) Pemilihan atribut yang memiliki nilai *information gain* terbesar untuk dibuat simpul (*node*)

4) Ulangi proses perhitungan *information gain* sampai semua data telah masuk dalam kelas yang sama.

II. METODE PENELITIAN

A. Persiapan Data

Data yang digunakan pada penelitian ini berasal dari Rumah Sakit Pusat Pertamina Jakarta yang memuat data dari tahun 2013 - tahun 2015. Atribut yang digunakan dapat dilihat pada TABEL I.

TABEL I
ATRIBUT DATA

No	Atribut	Kode	Tipe
1	Umur	umur	Numerik
2	Jenis kelamin	sex	Nominal
3	Glukosa darah puasa	glun	Numerik
4	Glukosa darah 2 jam	gpost	Numerik
5	Glukosa urin puasa	urn	Numerik
6	Glukosa urin 2 jam	upost	Numerik
7	Aseton urin puasa	actn	Numerik
8	Aseton urin 2 jam	actpp	Numerik
9	Kolesterol LDL	ldl	Numerik
10	Kolesterol HDL	hdl	Numerik
11	Kolesterol total	chol	Numerik
12	Trigliserida	tg	Numerik
13	Kelas	class	Nominal

B. Pra Pemrosesan

1) Data cleaning. Pada tahap ini dilakukan penanganan *missing value* menggunakan *series mean*.

2) Data transformation. Pada tahap ini, data diubah kedalam bentuk kategori. Pembagian kategori yang digunakan dapat dilihat pada TABEL II.

TABEL II
PEMBAGIAN KATEGORI

No	Atribut	Kategori	Ket	*Map*
1	Umur	MUDA	>25	1
		DEWASA	25-59	2
		TUA	>59	3
2	Jenis Kelamin	L	Laki-Laki	1
		P	Perempuan	2
3	Glukosa Darah Puasa	BAIK	>101	1
		SEDANG	101-125	2
		BURUK	>125	3
4	Glukosa Darah 2 Jam	BAIK	>145	1
		SEDANG	145-179	2
		BURUK	>179	3
5	Kolesterol Total	BAIK	<200	1
		SEDANG	200-239	2
		BURUK	>239	3
6	Kolesterol HDL	BAIK	L : >40, P : >50	1
		BURUK	L : <41, P : <51	2
7	Kolesterol LDL	BAIK	<100	1
		SEDANG	100-129	2
		BURUK	>129	3
8	Trigliserida	BAIK	>150	1
		SEDANG	150-199	2
		BURUK	>199	3
9	Glukosa Urin Puasa	NEGATIF	= 0	1
		POSITIF	> 0	2
10	Glukosa Urin 2 Jam	NEGATIF	= 0	1
		POSITIF	> 0	2
11	Aseton Urin Puasa	NEGATIF	= 0	1
		POSITIF	> 0	2
12	Aseton Urin 2 Jam	NEGATIF	= 0	1
		POSITIF	> 0	2
13	Kelas	YA	Terjangkit	1
		TIDAK	Tidak terjangkit	2

3) Data selection. Tahap ini dilakukan penanganan *imbalance data* dan pemilihan atribut

a) Penanganan *Imbalance data*
Proses ini berhasil melakukan reduksi jumlah data kelas Tidak (Kelas Mayoritas) yang semula 2441 menjadi 1435 menggunakan *cluster based undersampling*.

b) Seleksi Atribute
Pemilihan atribut melibatkan algoritma *Correlation based Feature Selection* (CFS) dan algoritma *Information Gain*. Dilakukan pemilihan untuk 1-11 atribut menggunakan CFS dan *Information Gain*

C. Pelatihan

Langkah perhitungan pelatihan menggunakan ID3 adalah sebagai berikut:

1) Langkah pertama hitung jumlah total, kelas "Ya" dan kelas "Tidak" untuk setiap kategori.

2) Hitung nilai entropi untuk setiap kategori

$$Entropi(S) = \sum_{i=1}^{k} (-p_i \ * log_2 \, p_i)$$

Sehingga :

$$Entropi(Total) = \left(-\frac{1230}{2460} * log_2 \frac{1230}{2460}\right) + \left(-\frac{1230}{2460} * log_2 \frac{1230}{2460}\right) = 1$$

$$Entropi(MUDA) = \left(-\frac{0}{0} * log_2 \frac{0}{0}\right) + \left(-\frac{0}{0} * log_2 \frac{0}{0}\right) = 0$$

$$Entropi(DEWASA) = \left(-\frac{486}{1118} * log_2 \frac{486}{1118}\right) + \left(-\frac{632}{1118} * log_2 \frac{632}{1118}\right) = 0.9877$$

$$Entropi(TUA) = \left(-\frac{744}{1342} log_2 \frac{744}{1342}\right) + \left(-\frac{598}{1342} log_2 \frac{598}{1342}\right) = 0.9914$$

3) Hitung nilai information gain untuk setiap atribut

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \left(\frac{|S_i|}{|S|}\ x\ Entropi\ (S_i)\right)$$

Sehingga :

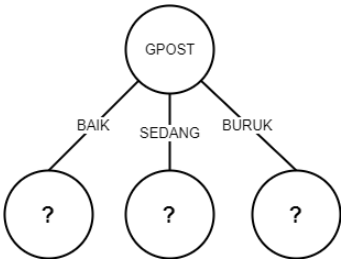
$$Gain(umur) = 1 - \left(\left(\frac{0}{2460} * 0\right) + \left(\frac{1118}{2460} * 0.9877\right) + \left(\frac{1342}{2460} * 0.9914\right)\right) = 0.0103$$

4) Lakukan untuk semua atribut, maka diperoleh :

Gain(umur) = 0.0103
Gain(sex) = 0.0033
Gain(glun) = 0.3352
Gain(gpost) = 0.3687
Gain(chol) = 0.0069
Gain(hdl) = 0.002
Gain(ldl) = 0.0176
Gain(tg) = 0.0071
Gain(urn) = 0.092
Gain(upost) = 0.2103
Gain(actn) = 0.0158
Gain(actpp) = 0.0127

5) Nilai *Gain* tertinggi dimiliki oleh atribut gpost (glukosa darah 2 jam). Maka atribut gpost ditetapkan sebagai node root.

6) Jika semua sampel berada dalam kelas yang sama, maka *node* akan menjadi daun dan dilabeli menjadi kelas. Jika tidak, maka akan dilakukan perhitungan *gain* dengan *node root* kategori pada atribut dengan nilai gain tertinggi pada. Pada gpost tidak ada kategorinya yang terletak pada kelas yang sama, maka tidak ada yang menjadi daun (maka dilabeli tanda "?" yang berarti dilakukan perhitungan selanjutnya). Ilustrasi pembentukan *node* dapat dilihat pada Gambar 1.



Gambar 1. Ilustrasi Pembentukan Node

7) Lakukan perhitungan sampai semua sampel pada *node* menghasilkan suatu kelas dan tidak ada atribut lainnya yang dapat digunakan untuk mempartisi sampel lebih lanjut

III. HASIL DAN PEMBAHASAN

Hasil pemilihan atribut menggunakan algoritma CFS dalam kodisi ketika data tidak imbang dapat diihat pada TABEL III sedang untuk kondisi data imbang pada TABEL IV.

TABEL III
HASIL PEMILIHAN ATRIBUT PADA DATA TIDAK IMBANG MENGGUNAKAN CFS

No	Atribut yang Digunakan
1	Gpost
2	glun, gpost
3	glun, gpost, upost,
4	glun, gpost, upost, tg
5	umur, glun, gpost, upost, actpp
6	umur, glun, gpost, upost, actpp, tg
7	umur, sex, glun, gpost, upost, actpp, tg
8	umur, sex, glun, gpost, upost, actpp, ldl, tg
9	umur, sex, glun, gpost, upost, actpp, ldl, hdl, tg
10	umur, sex, glun, gpost, upost, urn, actpp, ldl, hdl, tg
11	umur, sex, glun, gpost, upost, actn, urn, actpp, ldl, hdl, tg
12	umur, sex, glun, gpost, upost, actn, urn, actpp, ldl, hdl, chol, tg

TABEL IV
HASIL PEMILIHAN ATRIBUT PADA DATA IMBANG MENGGUNAKAN CFS

No	Kombinasi Atribut
1	Gpost
2	glun, gpost
3	glun, gpost, upost
4	glun, gpost, upost, actn
5	glun, gpost, upost, actn, ldl
6	sex, glun, gpost, upost, actpp, ldl
7	umur, sex, glun, gpost, upost, actpp, ldl
8	umur, sex, glun, gpost, upost, actpp, ldl, hdl
9	umur, sex, glun, gpost, upost, actpp, ldl, hdl, tg
10	umur, sex, glun, gpost, upost, urn, actpp, ldl, hdl, tg
11	umur, sex, glun, gpost, upost, actn, urn, actpp, ldl, hdl, tg
12	umur, sex, glun, gpost, upost, actn, urn, actpp, ldl, hdl, chol, tg

Hasil pemilihan atribut menggunakan algoritma Information Gain dalam kodisi data tidak imbang dapat diihat pada TABEL V dan pada data imbang dapat dilihat pada TABEL VI.

TABEL V
HASIL PEMILIHAN ATRIBUT PADA DATA TIDAK IMBANG MENGGUNAKAN *INFORMATION GAIN*

No	Atribut yang Digunakan
1	Gpost
2	gpost, glun
3	gpost, glun, upost
4	gpost, glun, upost, urn
5	gpost, glun, upost, urn, tg
6	gpost, glun, upost, urn, tg, umur
7	gpost, glun, upost, urn, tg, umur, actpp
8	gpost, glun, upost, urn, tg, umur, actpp, sex
9	gpost, glun, upost, urn, tg, umur, actpp, sex, ldl
10	gpost, glun, upost, urn, tg, umur, actpp, sex, ldl, actn
11	gpost, glun, upost, urn, tg, umur, actpp, sex, ldl, actn, chol
12	gpost, glun, upost, urn, tg, umur, actpp, sex, ldl, actn, chol, hdl

TABEL VI
HASIL PEMILIHAN ATRIBUT PADA DATA IMBANG MENGGUNAKAN *INFORMATION GAIN*

No	Atribut yang digunakan
1	gpost
2	gpost, glun
3	gpost, glun, upost
4	gpost, glun, upost, urn
5	gpost, glun, upost, urn, actn
6	gpost, glun, upost, urn, actn, actpp
7	gpost, glun, upost, urn, actn, actpp, ldl
8	gpost, glun, upost, urn, actn, actpp, ldl, umur
9	gpost, glun, upost, urn, actn, actpp, ldl, umur, tg
10	gpost, glun, upost, urn, actn, actpp, ldl, umur, tg, sex
11	gpost, glun, upost, urn, actn, actpp, ldl, umur, tg, sex, chol
12	gpost, glun, upost, urn, actn, actpp, ldl, umur, tg, sex, chol, hdl

Perbandingan hasil pengujian dengan menggunakan data imbang dan data tidak imbang dapat dilihat pada TABEL VII.

TABEL VII
HASIL UJI DATA TIDAK IMBANG DENGAN DATA IMBANG

Kasus	Akurasi	*Sensitifity*	*Specificity*
Imbang	63.21	49.76	71.26
Tidak imbang	78.82	77.28	80.35

Perbandingan hasil pengujian penggunaan CFS dan Information Gain pada data tidak imbang dapat dilihat pada TABEL VIII. Dengan ak(akurasi), se(*sensitifity*), dan sp(*specificity*).

TABEL VIII
HASIL UJI DENGAN CFS DAN *INFORMATION GAIN* PADA DATA TIDAK IMBANG

Atr	CFS			Information Gain		
	ak	*Se*	*Sp*	Ak	*Se*	*Sp*
1	74.28	69.32	77.35	74.28	69.32	77.35
2	74.25	72.47	75.47	74.25	72.47	75.47
3	74.25	72.47	75.47	74.25	72.47	75.47
4	74.33	72.32	75.67	74.51	72.16	76.04
5	74.46	71.62	76.3	73.81	71.12	75.51
6	73.09	69.61	75.36	73.58	69.6	76.13
7	72.78	70.44	74.39	73.01	68.85	75.69
8	69.35	64.01	72.54	72.45	70.16	74.07
9	67.75	61.53	71.75	68.66	62.47	72.48
10	67.11	59.96	71.54	67.98	61.39	72.1
11	66.2	57.61	71.48	64.78	53.71	71.46
12	63.21	49.76	71.26	63.21	49.76	71.26

Perbandingan hasil pengujian penggunaan CFS dan Information Gain pada data tidak imbang dapat dilihat pada TABEL IX. Dengan ak(akurasi), se(*sensitifity*), dan sp(*specificity*).

TABEL IX
HASIL UJI DENGAN CFS DAN *INFORMATION GAIN* PADA DATA IMBANG

Atr	CFS			Information Gain		
	Ak	*Se*	*Sp*	Ak	*Se*	*Sp*
1	83.1	84.53	81.67	83.1	84.53	81.67
2	84.29	86.76	81.81	84.29	86.76	81.81
3	84.29	86.76	81.81	84.29	86.76	81.81
4	84.36	85.99	82.72	84.39	86.06	82.72
5	84.77	87.18	82.37	84.49	86.13	82.86
6	83.94	85.78	82.09	84.53	86.2	82.86
7	83.83	85.71	81.95	84.74	87.18	82.3
8	83.8	84.04	83.55	84.01	85.78	82.23

9	81.22	81.11	81.32	83.14	84.04	82.23
10	80.77	79.72	81.81	80.59	79.3	81.88
11	80.59	79.3	81.88	77.94	76.17	79.72
12	78.82	77.28	80.35	78.82	77.28	80.35

IV. PENUTUP

A. Kesimpulan

Kesimpulan yang dapat diambil dari hasil penelitian tugas akhir ini adalah:

1) Penggunaan data yang imbang pada penelitian ini meningkatkan performa prediksi menggunakan algoritma ID3.

2) Penggunaan metode seleksi atribut *Correlation based Feature Selection* dan *Information Gain* sama-sama dapat meningkatkan performa prediksi menggunakan algortima ID3. Keduanya terlihat sedikit perbedaan, namun performa tertinggi didapatkan pada penggunaan *Correlation based Feature Selection* dengan 5 atribut yaitu gpost, glun, upost, urn, dan actn dengan rata-rata akurasi 84.77, rata-rata *sensitifity* 87.18, dan rata-rata *specificity* 82.37.

B. Saran

Penggunaan seleksi atribut terbukti meningkatkan performa algoritma ID3 pada kasus ini. Untuk pengembangan penelitian selajutnya, dapat dicoba menggunakan metode seleksi atribut lain seperti *forward selection*, *backward elimination*, dsb.

DAFTAR PUSTAKA

[1] Kemenkes. 2014. *Situasi dan Analisis Diabetes*. Pusat Data dan Informasi Kementerian Kesehatan RI, Jakarta.

[2] Jasri, M. 2017. *Klasifikasi Penyakit Diabetes Mellitus Tipe 2 Dengan Metode Algoritma C4.5*. STT Nurul Jadid, Probolinggo.

[3] Kamagi, D.H. dan Hansun, S. 2014. *Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa*. UMN, Tangerang

[4] Santoso, T.B. 2013. *Analisa dan Penerepan Metode C4.5 untuk Prediksi Loyalitas Pelanggan*. Universitas Satya Negara Indonesia. Jakarta.

[5] Mutmainah, P.A.M. 2015. *Decision Tree Menggunakan Algoritma ID3 untuk Melakukan Deteksi Penyakit Kanker Payudara*. Universitas Dian Nuswantoro, Semarang.

[6] Rozzaqi, A.R. 2015. *Naïve Bayes dan Filtering Feature Selection Information Gain untuk Prediksi Ketepatan Kelulusan Mahasiswa*. Universitas PGRI, Semarang.

[7] Sathta, S dan Rajesh, A. 2016. *An Effective Prediction of Diabetics Using ID3 Classification Algorithm*. Tamilnadu, St.Peter's University, India.

[8] Djatna, T. dan Morimoto, Y. 2011. *Pembandingan stabilitas algoritma seleksi fitur menggunakan transformasi ranking normal*. IPB, Bogor.

[9] Depkes. 2008 *Pedoman Pengendalian Diabetes Melitus dan Penyakit Metabolik*. Direktorat Pengendalian PTM, Jakarta.

[10] Han, J dan Kamber, M. 2006. *Data Mining Concept and Tehniques 2nd Edition*. Morgan Kauffman, San Francisco.

[11] Yen, SJ dan Lee, YS. 2009. *Cluster-based under-sampling approaches for imbalanced data distributions.* Expert Systems with Applications 36(11), pp. 5718-5727.

[12] Fikri M.L. 2017. *Model Jaringan Syaraf Tiruan Radial Basis Function Dalam Deteksi Penyakit Ginjal Kronis*. Univversitas Diponegoro, Semarang.

[13] Hall, M.A. 2000. *Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning*. University of Waikato, Hamilton.

[14] Dinakaran, S. dan Thangaiah, P.R.J. 2013. *Role of Attribute Selection in Classification Algorithms*. International Journal of Scientific & Engineering Research ,Volume 4, Issue 6, June-2013.

[15] Utama, T.D. 2014. *Implementasi Algoritma Iterative Dichotomiser 3 Pada Penyeleksian Program Mahasiswa Wirausaha UNS*. UNS, Surakarta.

[16] Himawan, D. 2014. *Aplikasi Data Mining Menggunakan Algoritma ID3 Untuk Mengklasifikasi Kelulusan Mahasiswa Pada Universitas Dian Nuswantoro*. Universitas Dian Nuswantoro, Semarang.