ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 4, Nomor 2, Tahun 2015, Halaman 237 - 246
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

ALGORITMA *ITERATIVE DICHOTOMISER 3* (ID3) UNTUK MENGIDENTIFIKASI DATA REKAM MEDIS (Studi Kasus Penyakit Diabetes Mellitus Di Balai Kesehatan Kementerian Perindustrian, Jakarta)

Avia Enggar Tyasti[1], Dwi Ispriyanti[2], Abdul Hoyyi[3]
[1]Mahasiswa Jurusan Statistika FSM UNDIP
[2,3]Staff Pengajar Jurusan Statistika FSM UNDIP

ABSTRACT

Iterative Dichotomiser 3 (ID3) Algorithm is a basic decision tree learning algorithm. These algorithms perform a thorough search (greedy) in all possible decision tree. ID3 algorithm can be implemented using a recursive function, (function that calls itself). One of the problems that can be solved using the ID3 algorithm is a classification of diabetic patients. Diabetic is a disease because of the body is not able to control the amount of sugar or glucose in the bloodstream. Classification using ID3 in the case of diabetics produce trees with many vertices to 32 knot where 21 of them is a leaf node and attribute two-hour postprandial glucose fasting elected as the root node in the decision-making tree. Based on the classification performance measurements show that the classification accuracy or measurement accuracy reaches 89,75%. While the measurement accuracy of the classification algorithm ID3 using test samples totaling 84 samples showed an accuracy of 72,619%.

Keywords: ID3 Algortihm, Decision Tree, Diabetes

1. PENDAHULUAN

Algoritma *Iterative Dichotomiser* 3 (ID3) merupakan salah metode dalam data mining. Data Mining mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang penting dalam berbagai bidang, mulai dari bidang akademi, bisnis hingga medis. ID3 adalah algoritma *decision tree learning* (algoritma pembelajaran pohon) yang paling dasar. Algoritma ini melakukan pencarian secara menyeluruh pada semua kemungkinan pohon keputusan. Pembentukan pohon klasifikasi dengan algoritma ID3 melalui dua langkah, yaitu menghitung nilai *entropy* dan menghitung nilai *information gain* dari setiap variabel. ID3 dapat menyelesaikan kasus pada berbagai bidang salah satunya dapat diterapkan pada bidang kesehatan (Santosa, 2007).

Kesehatan merupakan aspek penting dalam kehidupan, banyak permasalahan yang terjadi dalam peningkatan taraf kesehatan masyarakat sehubungan gaya hidup yang kurang sehat *(unhealthy lifestyle)*, akibat dari *unhealthy lifestyle* dapat berujung pada munculnya berbagai macam penyakit. Masalah yang sering terjadi dalam gaya hidup masyarakat tersebut adalah Diabetes Mellitus (DM) yang merupakan penyakit yang disebabkan kadar gula darah yang tinggi. Hal ini menjadi tantangan yang berat pada sistem pelayanan kesehatan di negeri ini (Zahtamal, 2007).

Beberapa metode yang sering digunakan dalam pengklasifikasian adalah Analisis Diskriminan, Regresi Logistik Biner, algoritma *Iterative Dichotomiser 3* (ID3) dan lain-lain. Untuk mengidentifikasi penyakit Diabetes Mellitus tersebut, perlu diketahui ciri-

ciri pasien penyakit Diabetes Mellitus melalui berbagai hasil pengecekan tes laboratorium. Hasil pengecekan tersebut memiliki nilai diskret yang dapat dikategorikan, sehingga pada penelitian ini metode statistik klasifikasi yang digunakan adalah algoritma *Iterative Dichotomiser 3* (ID3).

Beberapa metode statistika yang telah digunakan pada penelitian sebelumnya pada kasus DM dan algoritma *Iterative Dichotomiser 3* (ID3) antara lain adalah "*Faktor-faktor Mempengaruhi Terjadinya Ulkus Diabetikum Pada Pasien Diabetes Melitus Tipe 2 Di RSUD Prof. DR. Margono Soekarjo Purwokerto*" Oleh Ferawati (2014), dan "*Klasifikasi Jurnal Ilmiah Berbahasa Inggris Berdasarkan Abstrak Menggunakan Algoritma ID3*" oleh Wijakso (2013).

2. TINJAUAN PUSTAKA

2.1 Data Mining

Data mining, sering juga disebut *Knowledge Discovery in Database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan (Santosa, 2007).

2.2 *Decision tree*

Menurut Santosa (2007), *decision tree* sesuai digunakan untuk kasus-kasus dimana outputnya bernilai diskret. Walaupun terdapat banyak variasi model *decision tree* dengan tingkat kemampuan dengan syarat yang berbeda, pada umumnya beberapa ciri kasus berikut cocok untuk diterapkan pada *decision tree* :

1. Data dinyatakan dengan pasangan atribut dan nilainya. Misalnya atribut satu data adalah temperatur dan nilainya adalah dingin.
2. Label atau output data biasanya bernilai diskret.
3. Untuk membuat *decision tree*, perlu diperhatikan :
    1. Atribut mana yang akan dipilih untuk pemisahan obyek
    2. Urutan atribut mana yang akan dipilih terlebih dahulu.
    3. Struktur *tree*
    4. Kriteria pemberhentian

2.3 Algoritma *Iterative Dichotomiser* 3 (ID3)

Iterative Dichotomiser 3 (ID3) adalah algoritma *decision tree learning* (algoritma pembelajaran pohon keputusan) yang paling dasar. Algoritma ini melakukan pencarian secara menyeluruh (*greedy*) pada semua kemungkinan pohon keputusan. Salah satu algoritma induksi pohon keputusan yaitu ID3 (*Iterative Dichotomiser 3)*. ID3 dikembangkan oleh J. Ross Quinlan. Algoritma ID3 dapat diimplementasikan menggunakan fungsi *rekursif* (fungsi yang memanggil dirinya sendiri). Algoritma ID3 berusaha membangun *decision tree* (pohon keputusan) secara *top-down* (dari atas ke bawah) (David, 2004).

2.3.1 *Entropy*

Menurut Rokach dan Maimoon (2008), *information gain* atau biasa disebut *gain info* adalah kriteria pemisahan yang menggunakan pengukuran *entropy*. Untuk mendapatkan *information gain* dari suatu atribut dibutuhkan *entropy* keseluruhan kelas atau *Entropy(S)*. Menurut Han *et al.* (2011), secara matematis *entropy* dirumuskan sebagai berikut

$$Entropy\ (S) = \sum_{i=1}^{c} p_i\ log_2\ p_i$$

dengan, S adalah himpunan kelas klasifikasi
c adalah banyaknya kelas klasifikasi
p_i adalah proporsi untuk kelas *i*

2.3.2 *Information Gain*

Setelah mendapatkan nilai *entropy*, maka dapat diukur efektivitas suatu atribut dalam mengklasifikasikan data yang disebut sebagai *information gain*. Secara matematis, *information gain* dari suatu atribut A, dituliskan sebagai berikut :

$$\text{Gain (S,A)} = Entropy\ (S) - \sum_{v \in Values(A)} \frac{S_v}{S}\ Entropy\ (S_v)$$

dengan, A adalah atribut
v menyatakan suatu nilai yang mungkin untuk atribut A
Values (A) adalah himpunan nilai-nilai yang mungkin untuk atribut A
S_v adalah sub-himpunan kelas klasifikasi
Entropy(S_v) adalah *entropy* untuk sampel-sampel yang memiliki nilai *v*

Atribut yang mempunyai nilai *information gain* paling tinggi dibanding dengan atribut yang lain, dipilih sebagai pemilah.

2.3.3 Ketepatan Pohon Klasifikasi

Menurut Prasetyo (2012), matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 1 merupakan matriks konfusi yang melakukan klasifikasi masalah biner.

Tabel 1. Matriks Konfusi

f_{ij}		Kelas Hasil Prediksi (j)	
		Kelas = A	Kelas = B
Kelas asli (i)	Kelas = A	f_{11}	f_{10}
	Kelas = B	f_{01}	f_{00}

Dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar, yaitu $(f_{11} + f_{00})$, dan data yang diklasifikasikan secara salah, yaitu $(f_{10} + f_{01})$. Maka dapat dihitung tingkat akurasi dan tingkat kesalahan prediksi :

$$\text{Akurasi} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}}$$

2.4 Diabetes Mellitus

Diabetes adalah suatu penyakit karena tubuh tidak mampu mengendalikan jumlah gula atau glukosa dalam aliran darah. Terdapat dua tipe diabetes, yaitu diabetes tipe 1 dan diabetes tipe 2 (Toruan, 2012).

2.4.1 Penyebab Diabetes

Faktor utama diabetes tipe 1 disebabkan oleh faktor turunan alias gen yang diturunkan dari garis ibu atau ayah. Produk urine berlebihan, rasa haus yang tak kunjung hilang, nafsu makan yang terus meningkat berat badan menurun drastis dan rasa lelah yang tak kunjung hilang. Sedangkan pada diabetes tipe 2, faktor utama diabetes tipe ini tidak lain adalah kegemukan. Meski tidak menutup kemungkinan faktor gen juga berperan penting. (Toruan, 2012).

3. METODOLOGI PENELITIAN

3.1 Data dan variabel Penelitian

Jenis data yang digunakan dalam penelitian ini merupakan data sekunder. Data tersebut merupakan data rekam medis pegawai Kementrian Perindustrian yang berobat di Balai Kesehatan Kementerian Perindustrian mulai bulan Juli 2014 sampai dengan September 2014 dengan jumlah data sebanyak 416 data. Penjelasan atribut dalam penelitian ini dapat dilihat pada Tabel 2 dan Tabel 3.

3.2 Langkah-langkah Analisis

Langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut :

1. Membuat deskripsi data.
2. Membagi data menjadi sampel pelatihan dan sampel pengujian dengan melakukan beberapa kali percobaan dengan melihat hasil akurasi yang paling tinggi.
3. Mengkonstruksikan pohon keputusan algoritma ID3 dengan menghitung nilai *entropy* dan *information gain* dari masing-masing atribut.
4. Melakukan analisis terhadap hasil pohon keputusan yang terbentuk dan menghitung nilai akurasi pohon.
5. Mengidentifikasikan data rekam medis pasien positif diabetes mellitus dan negatif diabetes mellitus.
6. Menguji pohon keputusan menggunakan sampel pengujian.

Tabel 2. Kriteria Jenis Kelamin dan Usia Pasien

Atribut	Keterangan
Diabetes	Positif Negatif
Jenis Kelamin	Perempuan Laki-laki
Usia Pasien	26-35 = dewasa awal 36-45 = dewasa akhir 46-55 = lansia awal 56-65 = lansia akhir Sumber (Depkes, 2009)

Tabel 3. Kriteria Diabetes Mellitus

Atribut	Keterangan
Glukosa darah puasa (mg/dL)	80-109 = baik 110-125 = sedang ≥ 126 = buruk
Glukosa darah 2 jam (mg/dL)	80-144 = baik 145-179 = sedang ≥180 = buruk
HDL (mg/dL)	>45 = baik ≤45 = buruk
LDL (mg/dL)	<100 = baik 100-129 = sedang ≥130 = buruk
Trigliserida (mg/dL)	<150 = baik 150-199 = sedang ≥200 = buruk
hbA1c	<6,5 = baik 6,5-8 = sedang >8 = buruk

Sumber (Toruan, 2012)

4. PEMBAHASAN

4.1. Statistika Deskriptif

Berikut ringkasan data pada penelitian tugas akhir ini :

Deskripsi data berikut menunjukan informasi mengenai status diabetes pasien di Balai Kesehatan Kementerian Perindustrian Jakarta Periode Juli 2014 sampai dengan September 2014.

Tabel 4. Status Diabetes Pasien

Status Pasien	Jumlah	%
Positif	241	57,933
Negatif	175	42,067

Tabel 5. Status Diabetes Pasien Berdasarkan Jenis Kelamin.

Jenis Kelamin	Positif	Negatif	Total
Laki-laki	147	104	251
Perempuan	94	71	165
Total	241	175	416

Tabel 6. Status Diabetes Pasien Berdasarkan Atribut yang digunakan

Atribut	Minimum	Maksimum	Rataan
Usia (tahun)	28	64	48,25
Glukosa Puasa (mg/dL)	80	290	143,75
Glukosa 2 Jam PP (mg/dL)	82	389	183,27
Trygliserida (mg/dL)	38	301	144,56
HDL (mg/dL)	29	178	58,8
LDL (mg/dL)	47	200	111,06
hbA1c (%)	4,9	10,3	7,2

4.2. Algoritma *Itterative Dichotomizer 3* (ID3)

Langkah awal sebelum melakukan pengolahan data adalah membagi data menjadi data *training* dan data *testing*. Dalam penelitian ini, data dipartisi sebesar 80% untuk sampel pelatihan atau sebanyak 332 data dan 20% untuk sampel pengujian atau sebanyak 84 data.

4.2.1 Konstruksi Algoritma ID3

Berikut ini adalah perhitungan mencari nilai *entropy* dan *information gain* pada simpul akar menggunakan sampel pelatihan dengan Algorithma ID3 untuk mengkonstruksikan pohon keputusan. Perhitungannya adalah sebagai berikut :

1. Menghitung proporsi masing-masing kelas

Tabel 7. Proporsi Masing-masing Kelas

Kelas	Jumlah	Proporsi
Positif	186	0,56
Negatif	146	0,44
Total (S)	332	1,00

2. Menghitung nilai *entropy* kelas yang disimbolkan *E(S)*

Pada penelitian ini S adalah himpunan dari kelas klasifikasi positif dan negatif. Kelas klasifikasi positif dengan kode 1 dan kelas klasifikasi negatif dengan kode 2 sehingga diperoleh :

$$\textit{Entropy (S)} = \sum_i^c - p_i \log_2 p_i$$

$$\textit{Entropy}\ (1,2) = -\left(\frac{186}{332}\right).\ log_2\left(\frac{186}{332}\right) - \left(\frac{146}{332}\right).log_2\left(\frac{146}{332}\right)$$

$$= 0{,}989 \text{ bits}$$

3. Menghitung frekuensi masing-masing kategori pada atribut glukosa 2 jam PP berdasarkan kelasnya

Tabel 8. Frekuensi Masing-masing Kategori pada Atribut Glukosa 2 Jam PP Berdasarkan Kelasnya.

Glukosa 2 jam PP	Frekuensi		Total
	Positif	Negatif	
Baik	18	126	144
Buruk	158	20	178
Sedang	10	0	10
Total	186	146	332

4. Menghitung nilai *entropy* pada atribut glukosa 2 jam PP

Entropy (Baik, 1, 2) = $-\left(\frac{18}{144}\right).\ log_2\left(\frac{18}{144}\right)-\left(\frac{126}{144}\right).log_2\left(\frac{126}{144}\right)$ =0,543

Entropy (Buruk, 1, 2) = $-\left(\frac{158}{178}\right).log_2\left(\frac{158}{178}\right)-\left(\frac{20}{178}\right).log_2\left(\frac{20}{178}\right)$ = 0,506

Entropy (Sedang, 1, 2) = $-\left(\frac{10}{10}\right).log_2\left(\frac{10}{10}\right)-\left(\frac{0}{10}\right).log_2\left(\frac{0}{10}\right)=0$

5. Menghitung nilai *information gain*

Gain(S,Glukosa_2_PP) = *Entropy(S)*-$\sum_{v\in(baik,buruk,sedang)}\frac{|Sv|}{|S|}Entropy(S_V)$

= *Entropy (S)* $-\frac{144}{332}$ *Entropy*(S_{baik}) - $\frac{178}{332}$ *Entropy*(S_{buruk}) $-\frac{10}{332}$ *Entropy* (S_{sedang})

Gain (S,Glukosa_2_PP) = 0,989 – ($\frac{144}{332}$. 0,543) – ($\frac{178}{332}$. 0,506) – ($\frac{10}{332}$. 0) = 0,481

Berikut ini adalah hasil perhitungan mencari nilai *entropy* dan *information gain* dari semua atribut untuk menentukan pemilah terbaik.

Tabel 9. Nilai *Informartion Gain*

No	Atribut	*Gain*
1	Jenis Kelamin	0,00154
2	Usia	0,17752
3	Glukosa Puasa	0,45519
4	**Glukosa 2 jam pp**	**0,48191**
5	Trygserida	0,11609
6	HDL	0,09264
7	LDL	0,04619
8	HBA1C	0,47976

Berdasarkan Tabel 9, dapat diketahui bahwa atribut glukosa 2 jam pp adalah atribut dengan nilai *information gain* terbesar dengan nilai 0,48191, maka atribut glukosa 2 jam pp merupakan *the best classifer*.

4.2.2 Analisis Pohon Keputusan

Hasil Algoritma ID3 untuk mengidentifikasi data rekam medis pasien dengan studi kasus status diabetes pasien di Balai Kesehatan Kementerian Perindustrian dengan periode Bulan Juli 2014 sampai dengan September 2014 dengan atribut jenis kelamin, usia, glukosa puasa, glukosa dua jam setelah makan, kadar trygliserida, kadar HDL, kadar LDL, dan kadar hbA1c. Berikut ini informasi yang dapat diperoleh dari hasil klasifikasi menggunakan Algorithma ID3 berdasarkan Lampiran 3:

1. Pada penelitian ini banyak dari seluruh simpul yang terbentuk sebanyak 32 simpul.
2. Simpul daun merepresentasikan kelas yang terbentuk. Pada penelitian ini terbentuk sebanyak 21 simpul daun, ini artinya terdapat 21 karakteristik status diabetes pasien yang melakukan rekam medis di Balai Kesehatan Kementerian Perindustrian Jakarta.

3. Atribut glukosa 2 jam pp terpilih sebagai pemilah terbaik terhadap simpul akar berdasarkan nilai *information gain* yang terbesar.

4.2.3 Pengukuran Ketepatan Hasil Klasifikasi Algoritma ID3 berdasarkan Data Training

Setelah didapatkan secara utuh hasil klasifikasi Algoritma ID3 berupa pohon keputusan, langkah selanjutnya adalah mengukur ketepatan hasil klasifikasi yang terbentuk. Ketepatan klasifikasi maupun kesalahan klasifikasi dirangkum dalam tabel matriks konfusi sebagai berikut:

Tabel 10. Hasil Matriks Konfusi Algoritma ID3 Menggunakan Data *Training*

	Prediksi (Positif)	Prediksi (Negatif)	Total
(Positif)	173	13	186
(Negatif)	21	125	146
Total	194	138	332

Dapat dilihat pada Tabel 10 bahwa sebanyak 173 kasus dengan status diabetes positif dan sebanyak 125 kasus dengan status diabetes negatif diklasifikasikan secara tepat. Kemudian sebanyak 13 kasus dengan status diabetes positif diklasifikasikan kedalam status diabetes negatif, sehingga hal ini disebut kesalahan klasifikasi. Sebanyak 21 kasus dengan status diabetes negatif diklasifikasin kedalam status diabetes positif maka hal ini disebut kesalahan klasifikasi.

Akurasi atau persentasi dari keseluruhan kasus yang diklasifikasikan secara tepat pada konstruksi pohon ini adalah sebagai berikut :

$$\text{Akurasi} = \frac{173+125}{332} \times 100\%$$
$$= \frac{298}{332} \times 100\%$$
$$= 89{,}759\%$$

4.2.4 Hasil Pohon Keputusan Menggunakan Data *Testing*

Setelah didapatkan hasil konstruksi pohon dengan nilai akurasi mencapai 89,759%, maka untuk mengetahui apakah hasil konstruksi pohon baik digunakan untuk memprediksi kemungkinan kelas pada kasus-kasus selanjutnya, pohon konstruksi Algoritma ID3 tersebut diujikan dengan memasukkan data testing kedalam pohon konstruksi. Ukuran sampel pengujian adalah sebanyak 84 kasus. Tabel matriks konfusi pada sampel pengujian sebagai berikut :

Tabel 11. Hasil Matriks Konfusi Sampel Pengujian Menggunakan Data *Testing*

Aktual	Prediksi (Positif)	Prediksi (Negatif)	Total
Positif	51	4	55
Negatif	19	10	29
Total	70	14	84

Berdasarkan Tabel 11 nilai akurasi Algoritma ID3 pada sampel pengujian adalah sebagai berikut :

$$\text{Akurasi} = \frac{51+10}{84} \times 100\%$$
$$= \frac{61}{84} \times 100\%$$
$$= 72{,}619\%$$

Berdasarkan hasil tingkat akurasi pohon klasifikasi dalam mengklasifikasikan data diperoleh tingkat akurasi sebesar 72,619% dan dengan tingkat kesalahan memprediksi sebesar 27,381%, sehingga hasil konstruksi pohon cukup baik digunakan untuk memprediksi kemungkinan kelas pada kasus-kasus selanjutnya.

5. KESIMPULAN

Dari hasil analisis dapat diambil kesimpulan sebagai berikut:

1. Konstruksi pohon keputusan yang terbentuk menggunakan Algoritma ID3 menghasilkan pohon dengan banyak simpul mencapai 32 simpul dimana 21 diantaranya adalah simpul daun dan atribut glukosa puasa dua jam *postprandial* terpilih sebagai simpul akar dalam pembuatan pohon keputusan.
2. Berdasarkan pengukuran kinerja klasifikasi menunjukkan bahwa akurasi atau ukuran ketepatan klasifikasi mencapai 89,759 %. Berdasarkan pengukuran akurasi hasil klasifikasi Algoritma ID3 menggunakan sampel pengujian yang berjumlah 84 sampel menunjukkan akurasi sebesar 72,619%.

DAFTAR PUSTAKA

David, Mcg. 2004. *Tutorial: The ID3 Decision Tree Algorithm*. Monash University Faculty of Information Technology.

Ferawati, I. 2014. *Faktor-faktor Mempengaruhi Terjadinya Ulkus Diabetikum Pada Pasien Diabetes Melitus Tipe 2 Di RSUD Prof. DR. Margono Soekarjo Purwokerto*. Skripsi. Tidak Dipublikasikan. Universitas Jenderal Soedirman: Purwokerto.

Han, J, Kamber,M and Pei, J. 2011. *Data Mining Concepts and Technique*. Third Edition. Elsevier, Inc. *Massachusetts.*

Hardiwino.2012.http://ilmu-kesehatan-masyarakat.blogspot.com/2012/5/kategori-umur.html?m=1 diakses pada tanggal 8 November 2014

Prasetyo, E. 2014. *Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB*. Andi: Yogyakarta.

Prasetyo, E. 2012. *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. C.V Andi Offset: Yogyakarta.

Rangkuti, Y. R. 2011. *Hubungan Antara Diabetes Melitus Tipe 2 dengan Retinopati Diabetik Dikaji Dari hbA1c Sebagai Parameter Kontrol Gula Darah*. Tesis. Tidak Dipublikasikan. Universitas Sumatera Utara: Medan.

Rokach, L and Maimon, O. 2008. *Data Mining With Decision Trees : Theory and Applications*. World Scientific Publishing Co. Pte. Ltd. Singapura.

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

Toruan, P. L. 2012. *Diabetes Sakit Tapi Sehat*. Transmedia : Jakarta.

Wijakso, B. 2013. *Klasiikasi Jurnal Ilmiah Berbahasa Inggris Berdasarkan Abstrak Menggunakan Algoritma ID3.* Skripsi. Tidak Dipublikasikan. Universitas Brawijaya: Malang.

Zahtamal, *et al*. 2007. *Faktor-faktor Resiko Pasien Diabetes Mellitus Vol. 23, No. 3, hal. 142-147*. Berita Kedokteran Masyarakat.