ISBN: 978-602-60280-1-3

Penerapan *Data Mining* untuk Mengidentifikasi Penyakit Diabetes Mellitus dengan Menggunakan Algoritme Iterative Dichotomiser 3 (ID3)

[1st]Okriandy Nugroho, [2nd]Bagus Adhi Kusuma, [3rd]Zanuar Rifa'i, [4th]Tri Astuti, [5th]Uswatuh Khasanah, [6th]Rizky Wahyudi

Teknik Informatika
STMIK Amikom Purwokerto
Jl. Let. Jend. Pol. Soemarto, Watumas, Purwanegara, Purwokerto Utara

ABSTRAK-Algoritma Iterative Dichotomier 3 (ID3) adalah algotitma pembalajaran decision tree dasar. Algoritma ini melakukan pencarian menyeluruh di semua pohon keputusan yang memungkinkan. Algoritma dapat diimplementasikan menggunakan fungsi rekursif (fungsi yang memanggil dirinya sendiri). Salah satu masalah yang dapat diselesaikan dengan menggunakan algoritma ID3 adalah klasifikasi pasien diabetes. Diabetes adalah penyakit karena tubuh tidak mampu mengontrol jumlah gula atau glukosa dalam aliran darah. Klsifikasi menggunakan ID3 dalalm kasus penderita diabetes menghasilkan pohon dengan banyak simpul menjadi 32 knot dimana 21 diantaranya adalah simpul daun dan atribut puasa glukosa postpradinal dua jam terpilih sebagai simpul akar di pohon pengambilan keputusan. Berdasarkan hasil pengukuran kinerja klasifikasi menunjukkan bahwa akurasi klasifikasi atau akurasi pengukuran mencapai 89,75%. Sedangkan keakuratan pengukuran algoritma klasifikasi ID3 menggunakan sampel uji yang berjumlah 84 sampel menunjukkan akurasi 72,619%.

Keywords: ID3, Diabetes Mellitus, Data Mining

I. PENDAHULUAN

Algoritma Iterative Dichotomiser 3 (ID3) adalah salah satu metode dalam data mining. ID3 diperkenalkan pertama kali oleh Ross Quinlan (1979. ID3 merepresentasikan konsep-konsep dalam bentuk pohon keputusan. Algoritma ini melakukan pencarian secara menyeluruh pada semua kemungkinan pohon kehidupan. Pembuatan pohon klasifikasi ID3 dilakukan melalui 2 langkah, yaitu menghitung nilai dari enntropy dan menghitung nilai information gain dari variabel-variabel. ID3 dapat menyelesaikan kasus pada berbagai bidang salah satunya bidang kesehatan.

Kesehatan merupakan aspek yang sangat penting dalam kehidupan. Banyak permasalahan yang terjadi dalam kesehatan masyarakat terutama gaya hidup yang kurang sehat. Akibat gaya hidup yang kurang sehat akhirnyamuncullah berbagai macam penyakit. Masalah yang sering terjadi adalah penyakit Diabetes Mellitus (DM). Penyakit Diabetes Mellitus (DM) adalah penyakit yang disebabkan karena kadar gula yang tinggi.

Untuk mengidentifikasi penyakit tersebut perlu mengetahui ciri-ciri pasien melalui berbagai pengecekan dan tes laborat. Hasil pengecekan tersebut mempunyai nilai diskret yangbisa dikategorikan sehingga pada penelitain ini metode yang digunakan adalah metode Algoritma Iterative Dichotomiser 3 (ID3).

II. METODOLOGI PENELITIAN

1. Data dan Variable Penelitian

Jenis data yang digunakan dalam penelitian ini adalah data sekunder. Data ini merupakan data rekam medis pasien yang berobat di Klinik Tanjung mulai bulan Oktober 2017 sampai bulan Desember 2017 dengan jumlah data sebanyak 416 data. Penjelasan atribut pada tabel 1 da 2.

ISBN: 978-602-60280-1-3

2. Langkah-langkah Analisis

Langkah-langkah yang dilakukan adalah sebagai berikut :

1. Membuat deskripsi data.
2. Membagi data menjadi 2 sampel, yaitu sampel pelatihan dan pengujian dengan melakukan beberapa kali percobaan dengan melihat hasil akurasi yang paling tinggi.
3. Mengkonstruksi pohon keputusan algoritma ID3 dengan menghitung nilai dari entropy dan information gain dari masing-masing atribut.
4. Melakukan analisa terhadap hasil pohon keputusan yang terbentuk dan menghitung nilai akurasi pohon tersebut.
5. Mengidentifikasi data rekam medis pasien yang positif diabetes melliyus dan negative diabetes mellitus.
6. Menguji pohon keputusan menggunakan sampel pengujian.

Tebel 1. Kriteria Jenis Kelamin dan Usia Pasien

Atribut	Keterangan
Diabetes	Positif
	Negatif
Jenis Kelamin	Perempuan
	Laki-laki
Usia	26 – 35 = Dewasa awal
	36 – 45 = Dewasa akhir
	46 – 55 = Lansia awal
	56 – 65 = Lansia akhir

Tabel 2. Kriteria Diabetes Mellitus

Atribut	Keterangan
Glukosa darah puasa (mg/dL)	80 – 100 = 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
Trigiserida (mg/dL)	< 150 = baik
	150 – 199 = sedang
	≥ 200 = buruk
hbAlc	< 6,5 = baik
	6,5 – 8 = sedang
	➢ 8 = buruk

III. HASIL DAN PEMBAHASAN

1. Statistika Deskriptif

Deskripsi data berikut menunjukkan informasi mengenai status penyakit diabetes pasien di Klinik Tanjung Periode Oktober 2017 sampai Desember 2017.

Tabel 3. Status Diabetes Pasien

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

Tabel 4. Status Diabetes Pasien Berdasarkan Jenis Kelamin

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

Tabel 5. 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	1183,27
Trigliserida (mg/dL)	38	301	144,56
HDL (mg/dL)	29	178	58,8
LDL (mg/dL)	47	200	111,06
hbAcl (%)	4,9	10,3	7,2

2. Algoritma Itterative Dichotomizer 3 (ID3)

Langkah awal sebelum melakukan pengolahan data adalah membagi data menjadi 2 data, yaiti data training dan testing. Dalam penelitian ini data dipartisi sebesar 80% untuk sampel pelatihan atau sebanyak 332 data dan 20% untuk sampel pengujian atau sebanyak 84 data.

3. Konstruksi Algoritma ID3

Berikut ini perhitungan mencari nilai dari entropy dan information gain pada simpul akar menggunakan sampel pelatihan dengan Algoritma ID3 untuk mengkonstruksi pohon keputusan. Perhitungannya adalah sebabgai berikut :

ISBN: 978-602-60280-1-3

a. Menghitung proporsi masing-masing kelas.

Tabel 6. Proporsi Masing-masing Kelas

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

b. Menghitung niali entropy kelas yang disimbolkan *E* (S).

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

Entropy (S) $= \sum_i^c - pi \ log_2 \ pi$

Entropy (1,2) =

$-(\frac{186}{332}) \ . \ \log 2 \ (\frac{186}{332}) - (\frac{146}{332}) \ . \ \log 2 \ (\frac{146}{332})$ = 0,989 bits

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

Tabel 7. 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

d. Menghitung nilai entropy pada atribut glukosa 2 jam PP :

- Entropy (Baik,1, 2) =

$$-(\frac{18}{144}) \ . log2 \ (\frac{18}{144}) - (\frac{126}{144}) \ . \ log2 \ (\frac{126}{144}) = \mathbf{0,543}$$

- Entropy (Buruk, 1, 2) =

$-(\frac{158}{178}) \ . log2 \ (\frac{158}{178}) - (\frac{20}{178}) \ . log2 \ (\frac{20}{178}) = \mathbf{0,506}$

- $Entropy$ (Sedang, 1, 2) =

$$-(\frac{10}{10}) \ . log2 (\frac{10}{10}) - (\frac{0}{10}) = \mathbf{0}$$

e. Menghitung information gain :

- Gain (S,Glukosa_2_PP) =

$$0{,}0989 - (\frac{144}{332} \ . 0{,}543) - (\frac{178}{332} \ . 0{,}506) - (\frac{10}{332} \ . 0) = \mathbf{0,481}$$

Berikut ini adalah hasil perhitungan mencari nilai entropy dan informasi gain dari semua atribut untuk menentukan pemilah terbaik :

Tabel 8. Nilai Information 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	Trygiserida	0,11609
6	HDL	0,09264
7	LDL	0,04619
8	hbAlc	0,47976

Berdasarkan Tabel 8, dapat deiketahui bahwa atribut glukosa 2 jam PP adalah atribut dengan nilai information gain terbesar, yaitu 0,48191. Maka atribut tersebut merupakan *the best classifer*.

4. Analisa Pohon Keputusan

Hasil Algoritma ID3 untuk mengidentifikasi data rekam medis pasien di Klinik Tanjung periode Oktober 2017 sampai dengan Desember 2017 dengan atribut jenis kelamin, usia, glukosa puasa, glukosa 2 jam setelah makan, kadar trygiserida, kadar HDL, kadar LDL, dan kadar hbAlc. Berikut ini adalah informasi yang diperoleh dari hasil klasifikasimenggunakan Algoritma ID3 :

a. Dari penelitian ini banyaknya simpul yang terbentuk sebanyak 32 simpul.
b. Simpul daun merepresentasikan kelas yang terbentuk. Pada penelitian ini terbentuk sebanyak 21 simpul, itu artinya terdapat 21 karakteristik status diabetes pasien yang melakukan rekam medis di Klinik Tanjung.
c. Atribut glukosa 2 jam PP terpilih menjadi simpul akar berdasarkan nilai information gain yang terbesar

.

5. Pengukuran Ketepatan Hasil Klasifikasi Algoritma ID3 Berdasarkan Data Training

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

ISBN: 978-602-60280-1-3

Tabel 9. Hasil Matriks Konfusi Algoritma ID3 Menggunakan Data Training

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

Dapat dilihat pada Tabel 9 bahwa sebanyak 173 kasus dengan status positif diabetes dan 125 kasus dengan kasus negatif diabetes diklasifikasikan secara tepat. Kemudian sebanyak 13 kasus dengan status positif diabetes diklasifikasikan kedalam status negatif diabetes, sehingga hal ini disebut kesalahan klasifikasi. Sebanyak 21 kasus dengan kasus negatif diabetes diklasifikasikan kedalam status positif diabetes maka hal ini disebut kesalahan klasifikasi. Akurasi dari keseluruhan kasus yang diklasifikasi secara tepat pada kondtruksi pohon ini ialah sebagai berikut.

$$\frac{173 + 125}{332}\ x\ 100\% = \frac{298}{332}\ x\ 100\%$$

= 89,759% (Akurasi)

6. Hasil Pohon Keputusan Menggunakan Data Testing

Setelah didapatkan hasil konstruksi pohon dengan nilai akurasi sebesar 89,758%, maka untuk mengetahui apakah hasil dari konstroksi pohon baik untuk prediksi kemungkinan kelas pada kasus-kasus selanjutnya, pohon konstruksi Algoritma ID3 tersebut diuji dengan memasukkan data testing kedalam data konstruksi. Ukuran sampel pengujian ialah sebanyak 83 kasus. Tabel matriks konstruksi pada sampel pengujian ialah sebagai berikut :

Tabel 10. Hasil Matriks Konfusi Sampel Pengujian Menggunakan Data Testing

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

Berdasarkan tabel 10, nilai akurasi Algoritma ID3 sebagai berikut :

$$\frac{51 + 10}{84}\ x\ 100\% = \frac{61}{84}\ x\ 100\%$$

= 72,619% (Akurasi)

IV. KESIMPULAN

Dari hasil hasil analisa dapat diambil kesimpulan yaitu konstruksi pohon keputusan yang dibuat menggunakan Algoritma ID3 menghasilkan pohon dengan simpul sebanyak 32 simpul dimana 21 diantaranya ialah simpul daun dan atribut glukosa puasa 2 jam *postprandial* terpilih sebagai simpul akar dalam pembuatan pohon keputusan. Berdasarkan kinerja klasifikasi menunjukkan bahwa akurasi mencapai 89,759%. Dan berdasarkan pengukuran akurasi dari hasil klasifikasi Algoritma ID3 menggunakan sampel pengujian yang berjumlah 84 sampel menunjukkan akurasi sebesar 72,619%.

DAFTAR PUSTAKA

[1] Ferawati, I. 2014. Faktor-faktor Mempengaruhi Terjadinya Ulkus Diabetikum Pada Pasien Diabetes Mellitus Tipe 2 Di RSUD Prof. DR. Margono Soekardjo Purwokerto. Skripsi. Tidak Dipublikasikan. Universitas Jendral Soedirman: Purwokerto.

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

[3]Hardiwino.2012.http//ilmu-kesehatan-masyarakat.blogspot.com/2012/5/kategori-umur.html?m=1

[4] Prsetyo, E. 2014. Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB. Andi: Yogyakarta.

[5] Rangkuti, Y, R. 2011 Hubungan Antara Diabetes Mellitus Tipe 2 dengan Retinopati Diabetik Dikaji Dari hbAlc Sevagai Parameter Kontrol Gula Darah. Tesis.

[6] Santosa, B. 2007. Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Graha Ilmu: Yogyakarta.

[7] Toruan, P. L. 2012. Diabetes Sakit Tapi Sehat. Transmedia : Jakarta.