Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika dan Nilai Islami)
Vol.1, No.1, Juli 2017, Hal. 485-491
p-ISSN: **2580-4596; e-ISSN: 2580-460X**

Model *Machine Learning* CART Diabetes Melitus

Ria Dhea Layla Nur Karisma[1], Bambang Widjanarko Otok[2]
* UIN Maulana Malik Ibrahim Malang
** Institut Teknologi Sepuluh Nopember Surabaya

Info Artikel

Riwayat Artikel:

Diterima: 15 Mei 2017
Direvisi: 1 Juni 2017
Diterbitkan: 31 Juli 2017

Kata Kunci:

Diabetes Tipe I
Diabetes Tipe II
Dibetes Gestasional
CART

ABSTRAK

Penyakit Diabetes secara perlahan dapat menimbulkan masalah yang dikenal dengan *the silent killer*. Penyakit Diabetes disebabkan oleh kerusakan pada hormon insulin Tipe penyakit Diabetes ada tiga jenis, yaitu Diabetes tipe I yang disebabkan oleh kurangnya produksi insulin, tipe II yang disebabkan oleh produksi hormon insulin yang berlebihan, dan Gestasional yaitu hiperglekemia yang terjadi selama kehamilan. Metode CART (*Classification and Regression Tree*) merupakan salah satu metode yang digunakan untuk pengklasifikasian. Metode CART dapat digunakan untuk data yang memiliki skala kontinu maupun rasio. Data yang digunakan pada penelitian ini merupakan data skunder dari penderita Diabetes Melitus tipe II dan bukan Tipe II. Variabel respon yaitu penderita Diabetes tipe II dan bukan tipe II, dengan variabel prediktor riwayat keluarga, usia, jenis kelamin, obesitas, pola makan, dan aktivitas fisik (olahraga). Faktor-faktor yang mempengaruhi penderita Diabetes Melitus menurut metode CART riwayat keluarga, obesitas, usia, dan jenis kelamin.

Korespondensi:

Ria Dhea Layla Nur Karisma,
Jurusan Matematika,
UIN Maulana Malik Ibrahim Malang,
Jl. Gajayana No. 50 Malang, Jawa Timur, Indonesia 65144
Email: riadhea@uin-malang.ac.id

1. PENDAHULUAN

Penyakit Diabetes Melitus merupakan salah satu penyakit tidak menular (PTM) utama dengan jumlah penderitanya mengalami peningkatan. Diabetes Melitus (DM) disebut juga "*the silent killer*" karena secara perlahan dapat menimbulkan masalah yang serius dan menyebabkan kematian. Indonesia menempati urutan ke-4 terbesar jumlah penderita penyakit ini dan terus mengalami peningkatan tinggi setiap tahunnya [1]. Hal ini disebabkan kurangnya pengetahuan masyarakat mengenai gejala-gejalanya. Jenis diabetes menurut [2] ada tiga jenis Diabetes Melitus yaitu Diabetes tipe I, Diabetes tipe II, dan Diabetes Gastasional.

Salah satu metode yang digunakan untuk menggolongkan Diabetes Melitus berdasarkan faktor yang mempengaruhi Classification and Regression Trees (CART). Metode CART merupakan metode yang dapat diterapkan pada data dalam jumlah besar, variabel yang sangat banyak serta melalui prosedur pemilah biner [3]. Pada penelitian ini mengklasifikasikan penderita Diabetes Melitus yang dilakukan dengan menerapkan metode CART. Data yang digunakan dalam penelitian data sekunder mengenai penderita DM yang berasal dari RSUD Kabupaten Gorontalo. Mengingat adanya kultur atau budaya pernikahan sedarah di provinsi Gorontalo, tingkat pendidikan yang rendah, kondisi ekonomi yang rendah, dan minimnya sarana kesehatan seperti rumah sakit sehingga angka Diabetes Melitus di provinsi tersebut meningkat setiap tahunnya [4]. Berdasarkan latar belakang yang telah dipaparkan penelitian ini membahas mengenai klasifikasi perdasarkan faktor yang mempengaruhi penderita Diabetes Melitus dengan metode CART.

2. TINJAUAN PUSTAKA

2.1. Diabetes Melitus

Diabetes melitus atau diabetes adalah penyakit kronis, yang terjadi ketika pankreas tidak menghasilkan insulin yang cukup, atau ketika tubuh tidak dapat secara efektif menggunakan insulin yang dihasilkan. Sehingga menyebabkan peningkatan konsentrasi glukosa dalam darah (hiperglikemia). Beberapa jenis diabetes yang terjadi [2].

1. Diabetes tipe 1 (sebelumnya dikenal sebagai diabetes insulin-dependent atau anak-onset) ditandai oleh kurangnya produksi insulin.
2. Diabetes tipe 2 (sebelumnya disebut non-insulin-dependent atau diabetes onset dewasa) disebabkan oleh penggunaan efektif tubuh insulin. Hal ini sering terjadi karena kelebihan berat badan dan kurangnya aktivitas fisik.
3. Gestational diabetes adalah hiperglikemia yang ditemukan pertama kali selama kehamilan.

Ruang lingkup diabetes dibagi atas dua faktor, yaitu

1. Faktor risiko yang tidak dapat dimodifikasi (*unmodiable risk factors*), terdiri dari
   a. Ras dan Etnik
   b. Umur
      Kategori umur dalam ini hal ini adalah semua masyarakat yang berumur 20 tahun ke atas.
   c. Riwayat keluarga dengan diabetes melitus.
   d. Riwayat melahirkan bayi dengan berat (BB) lahir > 4.000 gram atau riwayat pernah menderita diabetes gestasional/kehamilan dengan diabetes melitus.
   e. Riwayat lahir dengan berat badan lahir rendah (BBLR) (< 2.500 gram)
2. Faktor risiko yang dapat dimodifikasi (*modifiable risk factors*)
   a. Berat Badan Lebih
      Berat badan lebih atau sering disebut dengan obesitas didefinisikan sebagai akumulasi lemak yang berlebih yang memiiki resiko bagi kesehatan. Ukuran seseorang mengalami obesitas adalah indeks masa tubuh (IMT). Berikut klasifikasi berdasarkan IMT.

Tabel 1. Klasifikasi Indeks Massa Tubuh (IMT)

Klasifikasi IMT		**Kg/m^2**
Kurus	Kekurangan berat badan tingkat berat	< 17,0
	Kekurangan berat badan tingkat ringan	17,0 – 18,4
Normal		18,5 – 25,0
Kegemukan	Kelebihan berat badan tingkat ringan	> 25,0 – 27,0
	Kelebihan berat badan tingkat ringan	> 27,0

   b. Obesitas Abdominal/Sentral (Lingkar perut untuk pria > 90 cm, wanita > 80 cm).
   c. Kurangnya Aktivitas Fisik
   d. Diet Tak Seimbang, dengan Tinggi Gula dan Rendah Serat
   e. Merokok

2.2. *Classification and Regression Trees* (CART)

CART merupakan metodologi statistik nonparametrik yang dikembangkan untuk topik analisis klasifikasi, baik untuk variabel respon kategorik maupun kontinu [5]. CART menghasilkan suatu pohon klasifikasi jika variabel responnya kategorik, dan menghasilkan pohon regresi jika variabel responnya kontinu. Tujuan utama CART adalah mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. Metode CART merupakan metode yang bisa diterapkan untuk himpunan data yang mempunyai jumlah besar, variabel yang sangat banyak dan dengan skala variabel campuran melalui prosedur pemilahan biner.

Langkah-langkah penerapan Algoritma CART adalah sebagai berikut

1. Pembentukan pohon klasifikasi
   Proses pembentukan pohon klasifikasi terdiri atas 3 tahapan, yaitu
   a. Pemilihan Pemilah (*Classifier)*
      Pemilihan pemilah tergantung pada jenis pohon atau tergantung pada jenis variabel responnya. Mengukur tingkat keheterogenan suatu kelas dari suatu simpul tertentu dalam pohon klasifikasi disebut *impurity measure i (t)*. Ukuran ini akan membantu menemukan fungsi pemilah yang optimal. Fungsi keheterogenan *i*(t) adalah sebagai berikut [5].

i. Indeks Gini
$$i(t) = \sum_{i \neq j} p(i|j)p(j|t) \quad (2.1)$$
ii. Indeks Informasi
$$i(t) = -\sum_j p(j|t) \log[p(j|t)] \quad (2.2)$$
iii. Indeks Twoing
$$i(t) = \frac{p_L p_R}{4}\left[\sum \left| p(j|t_L) - p\,(j|t_R) \right|\right]^{-2} \quad (2.3)$$
iv. Indeks Entropi
$$i(t) = N_j(t) \log[p(j|t)] \quad (2.4)$$

$p(j|t)$ adalah peluang j pada simpul t. *Goodness of Split* $\emptyset(s,t)$ didefinisikan sebagai penurunan keheterogenan.
$$\emptyset(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (2.5)$$
Pemilah yang menghasilkan nilai $\Delta i(s,t)$ lebih tinggi merupakan pemilah yang lebih baik t_L dan t_R merupakan partisi dari simpul *t* menjadi dua himpunan bagian saling lepas dimana $p_L$dan p_R adalah proporsi masing-masing peluang simpul. Karena $t_L \cup t_R = t$ maka nilai $\Delta i(s,t)$ merepresentasikan perubahan dari keheterogenan dalam simpul *t* yang hanya disebabkan oleh pemilah *s*. Jika simpul yang diperoleh merupakan kelas yang tidak homogen, prosedur yang sama diulangi sampai pohon klasifikasi menjadi suatu konfigurasi tertentu, dan memenuhi syarat berikut
$$\Delta i(s^*, t_1) = max_{s \subset S} \Delta i(s, t_1) \quad (2.6)$$

b. Penentuan Simpul Terminal
Suatu simpul *t* akan menjadi simpul terminal atau tidak, akan dipilah kembali bila pada simpul *t* tidak terdapat penurunan keheterogenan dengan adanya batasan minimum *n* seperti hanya terdapat satu pengamatan pada tiap simpul anak.

c. Penandaan Label Kelas
Penandaan label kelas pada simpul terminal dilakukan berdasarkan aturan jumlah terbanyak, yaitu
$$p(j_o|t) = max_j p(j|t) = max_j \frac{N_j(t)}{N(t)} \quad (2.7)$$
dengan,
$p(j|t)$: proporsi kelas *j* pada simpul *t*,
$N_j(t)$: jumlah pengamatan kelas *j* pada simpul *t*
$N(t)$: jumlah pengamatan pada simpul *t*.
j_o : label kelas simpul terminal *t*, $j_o$yang memberi nilai dugaan kesalahan pengklasifikasian pada simpul *t* terbesar.
Proses pembentukan pohon klasifikasi berhenti saat terdapat hanya satu pengamatan dalam tiap tiap simpul anak atau adanya batasan minimum *n*, semua pengamatan dalam tiap simpul anak identik, dan adanya batasan jumlah level atau kedalaman pohon maksimal.

2. Pemangkasan pohon klasifikasi
Pemangkasan dilakukan dengan jalan memangkas bagian pohon yang kurang penting sehingga didapatkan pohon optimal. Ukuran pemangkasan yang digunakan untuk memperoleh ukuran pohon yang layak adalah C*ost complexity minimum* [3]. Sub pohon dari pohon terbesar T_{max} ($T < T_{max}$) yang merupakan ukuran *cost complexity* yaitu
$$R_\alpha(t) = R(T) + \alpha\left|\tilde{T}\right| \quad (2.8)$$
dimana,
$R(T)$: *Resubtitusion Estimate* (Proporsi kesalahan pada sub pohon),
α : kompleksitas parameter (*complexity parameter)*,
$\left|\tilde{T}\right|$: ukuran banyaknya simpul terminal pohon *T*
$R_\alpha(t)$: merupakan kombinasi linear biaya dan kompleksitas pohon yang dibentuk dengan menambahkan *cost penalty* bagi kompleksitas terhadap biaya kesalahan klasifikasi pohon.
Cost complexity prunning menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan. $R_\alpha(t)$ pada seluruh pohon bagian, atau untuk setiap nilai α, dicari pohon bagian $T(\alpha) < T_{max}$ yang meminimumkan $R_\alpha(t)$, algoritma (2.8). Jika $R(T)$ digunakan sebagai kriteria penentuan pohon optimal maka akan cenderung pohon terbesar adalah T_1, sebab semakin besar pohon, maka semakin kecil nilai $R(T)$.

3. Penentuan Pohon Klasifikasi Optimal
Pohon klasifikasi yang berukuran besar memberikan nilai penduga pengganti paling kecil, sehingga pohon tersebut cenderung dipilih untuk menduga nilai respon. Ukuran pohon yang besar akan menyebabkan nilai kompleksitas yang tinggi karena struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan nilai penduga pengganti cukup kecil.

3. METODE PENELITIAN

3.1. Sumber Data

Data yang digunakan dalam penelitian ini merupakan data skunder yang berasal dari data pasien rawat inap penderita DM di RSUD Kabupaten Gorontalo tahun 2011.

3.2. Identifikasi Variabel

Tabel 2. Variabel Penelitian

No	Variabel	Skala	Keterangan
1.	Status (Y)	Nominal	0: Bukan penderita DM tipe II 1: Penderita DM tipe II
2.	Riwayat Keluarga (X_1)	Nominal	0: Tidak memiliki keturunan diabetes 1: Memiliki keturunan diabetes
3.	Umur (X_2)	Ratio	
4.	Jenis kelamin (X_3)	Nominal	0: jenis kelamin laki-laki 1: perempuan
5.	Obesitas (X_4)	Nominal	0: tidak menderita obesitas 1: menderita obesitas
6.	Pola makan (X_5)	Nominal	0: pola diet untuk mencegah diabetes 1: tidak memenuhi kriteria sehat
7.	Olah raga (X_6)	Nominal	Olahraga teratur 3-4 kali/minggu, setidaknya 20 sampai 30 menit (misalnya jalan kaki, senam) [1]. 0: aktif berolahraga 1: kurang aktif berolahraga

3.3. Metode Analisis Data

Analisis CART menggunakan langkah-langkah sebagai berikut

a. Kombinasi yang digunakan data *learning* dan data *testing*, yaitu (95% : 5%), (90% : 10%), (85% : 15%), (80% : 20%), dan (75% : 25%).
b. Pembentukan (*growing*) pohon klasifikasi maksimal,
c. Pemangkasan pohon klasifikasi yang paling kecil dengan kriteria kompleksitas kesalahan (*cost complexity*) yang minimum
d. Memilih pohon terbaik

3.4. Pembahasan

Hasil penelitian yang dilakukan beserta analisisnya adalah sebagai berikut.

3.4.1 Statistika Deskriptif

Menurut status penderita Diabetes Melitus. Penderita pasien tipe II merupakan penderita paling banyak yaitu 516 orang. Penderita yang memiliki riwayat keluarga menderita DM, yaitu 468 penderita. Usia yang paling banyak menderita Diabetes Melitus yaitu usia 43 sampai 64 tahun yaitu 368 pasien. Penderita Diabetes Melitus yang paling banyak adalah kelamin perempuan, yaitu 307 penderita. Penderita Diabetes Melitus paling banyak menderita obesitas atau kelebihan berat badan yaitu 487 orang. Penderita dengan pola diet lebih banyak menderita Diabetes Melitus tipe II yaitu 440 penderita. Aktif fisik yaitu berolahraga lebih banyak menderita DM yaitu 456 pasien.

3.4.2 Analisis CART (*Classification and Regression Trees*)

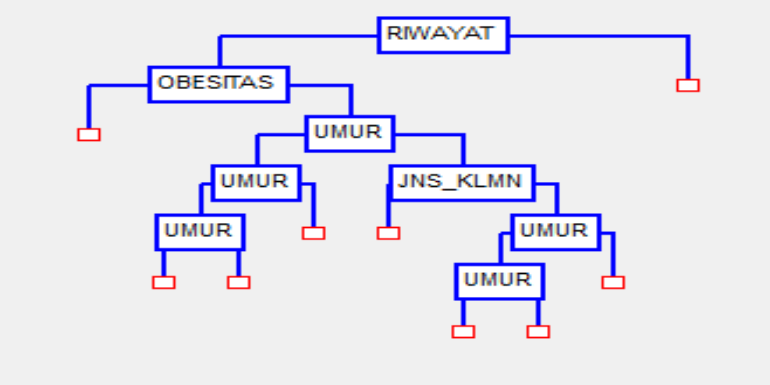
Data *learning* digunakan untuk permodelan sedangkan data *testing* digunakan untuk validasi model. Tabel 3 berikut merupakan perbandingan ketepatan klasifikasi antar beberapa kombinasi data. Kombinasi data *learning* dan *testing* tertinggi dicapai data *learning* 90% dan *testing* 10%. Sehingga kombinasi data *learning* dan *testing* inilah yang digunakan pada analisis selanjutnya.

Tabel 3. Perbandingan Ketepatan Klasifikasi Antar Kombinasi Data

No	Kombinasi Data (%)		Ketepatan Klasifikasi (%)	
	Learning	Testing	Learning	Testing
1	95	5	91,50	89,3
2*	**90**	**10**	**91,30**	**92,90**
3	85	15	91,80	89,30
4	80	20	91,30	92,00
5	75	25	91,90	90,00

1. Pohon Klasifikasi Maksimal

Pohon klasifikasi maksimal adalah pohon klasifikasi dengan jumlah simpul terminal terbanyak. Metode pemilih pada penelitian ini menggunakan Indeks Gini. Pemilah pertama terjadi pada variabel Riwayat.



Gambar 1. Pohon Klasifikasi Maksimal CART

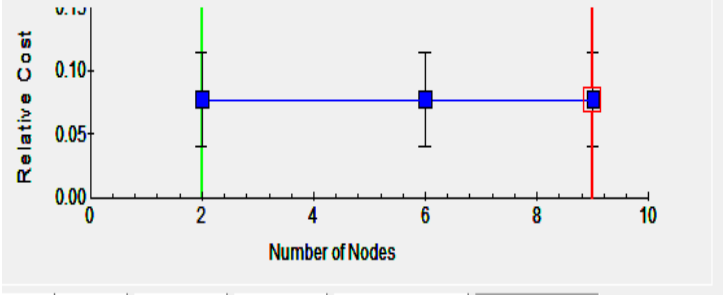
Variabel prediktor yang masuk dalam klasifikasi pohon maksimal yang terbentuk adalah variabel Riwayat, Usia, Jenis Kelamin, dan Obesitas. Variabel Riwayat merupakan pemilah yang memiliki peranan utama dalam pembentukan pohon maksimal dan merupakan variabel yang sangat dominan dalam pengelompokan. Tabel 4 berikut menunjukkan skor variabel prediktor dalam phon klasifikasi maksimum

Tabel 4. Skor Variabel Prediktor dalam Pohon Klaifikasi Maksimal

Variabel	Skor
Riwayat	100%
Aktivitas Fisik	36.595%
Pola Makan	36.595%
Obesitas	36.595%
Umur	9.316%
Jenis Kelamin	3.096%

2. Pemangkasan Pohon Klasifikasi Maksimal (*Pruning)*

Pemangkasan dilakukan berdasarkan aturan *cost complexty minimum* dan menggunakan penduga sampel uji (*test sample estimate*).



Gambar 2. Pemangkasan Pohon Maksimal (*Prunning)*

Pohon klasifikasi maksimal mengha-silkan penduga pengganti (*resubstitution relative cost*) yang kecil yaitu sebesar 0,056 dengan biaya kesalahan sebesar (*relative cost*) 0,077±0,037 atau antara 0,114-0,04. Nilai *relative cost* pada pohon maksimal memiliki nilai yang sama dengan nilai *relative cost* pada pohon optimal. Sehingga sudah tidak perlu dilakukan pemangkasan.

3. Pohon Klasifikasi Optimal

Pohon klasifikasi optimal diperoleh melalui langkah pemangkasan yang telah dilakukan sebelumnya. Karena nilai *relative cost* pada pohon maksimal memiliki nilai yang sama dengan nilai *relative cost* pada pohon optimal maka tidak perlu pemangkasan. Sehingga klasifikasi pohon maksimal yang digunakan dalam pengklasifikasian. Simpul terminal yang terbentuk adalah 9 simpul terminal, sebanyak 5 simpul terminal diprediksi sebagai kategori bukan penderita DM II dan 4 simpul terminal diprediksi sebagai kategori penderita DM tipe II.

Tabel 5. Ketepatan Klasifikasi Data *Learning*

Kelas Aktual	Prediksi Kelas		Ketepatan Klasifikasi (%)
	0	1	
0	40	0	40
1	26	438	464
Ketepatan Klasifikasi Keseluruhan			94,8

Pohon klasifikasi yang terbentuk mampu memprediksi dengan tepat pengamatan sebesar 94,8%. Penderita bukan DM tipe II yang salah diklasifikasikan ke dalam kelompok penderita DM II yaitu 0 pengamatan, atau tidak ada penderita bukan DM tipe II yang salah diklasifikasikan ke dalam kelompok penderita DM tipe II. Sedangkan, 26 penderita DM tipe II yang salah diklasifikasikan ke dalam kelompok penderita bukan DM tipe II.

4. Validasi Pohon Klasifikasi

Ketepatan pohon klasifikasi sebesar 94,8%, Validasi dilakukan dengan memasukkan data *testing* sebanyak 56 data ke dalam model pohon klasifikasi yang terbentuk sebelumnya dari data *learning*. Data *testing* yang digunakan adalah 10% .

Tabel 6. Ketepatan Klasifikasi Data *Testing*

Kelas Aktual	Prediksi Kelas		Ketepatan Klasifikasi (%)
	0	1	
0	4	0	100
1	4	48	92,3
Ketepatan Klasifikasi Keseluruhan			92,9

Tabel 6 menunjukkan bahwa sebanyak ketepatan klasifikasi pohon memiliki ketepatan klasifikasi sebesar 92,9%, artinya model pohon yang telah terbentuk memiliki keakuratan hasil prediksi sebesar 92,9%.

4. KESIMPULAN

Faktor yang mempengaruhi penderita Diabetes Melitus dengan pendekatan metode CART yaitu riwayat keluarga, obesitas usia, dan jenis kelamin.

REFERENSI

[1] Dinas Kesehatan. (2008). [Online]. Avaible: (http://www.depkes. go.id /downloads/ BULETIN%20PTM.pdf, diakses tanggal 8 Maret 2013)

[2] WHO (World Health Organitation). 2012. [Online]. Avaible: (http://www.who.int/ ,diakses tanggal 9 Oktober 2012)

[3] Lewis, M.D dan Roger, J. (2000). An Introduction to Classification and Regression Tree (CART) Analysis. Presented at the 2000 Anual Meeting of Society For Academy Emergency Medicine in San Fransisco, California [Online]. Avaible: (http://www.google.co.id/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CC8QFjAA&url=hkdi akses tanggal 4 Maret 2013)

[4] Kompas. (2001). [Online] Avaible: (http://kesehatan.kompasiana.com/seksologi/2011/08/08/rendahnya-tingkat-pendidikan-dan-pengetahuan-seks-mempengaruhi-meningkatnya-kasus-inses-384853.html,diakses tanggal 27 Mei 2013)

[5] Breiman L., Friedman J.H Olshen R.A & Stone C.J. (1993). Classification and Regression Tree. New York, NY: Chapman And Hall