

PENERAPAN PRUNING PADA ALGORITMA C5.0 UNTUK MENDIAGNOSIS PENYAKIT DIABETES MELITUS

¹Aric Kantono, ²Intan Yuniar Purbasari, ³Fetty Tri Anggraeny

^{2,3}Dosen Program Studi Teknik Informatika, Fakultas Ilmu Komputer,
Universitas Pembangunan Nasional “Veteran” Jawa Timur

Jalan Raya Rungkut Madya, Gunung Anyar, Kota Surabaya, Jawa Timur, 60294

Email: ¹arickantono@gmail.com, ²intanyuniar.if@upnjatim.ac.id, ³fettyanggraeny.if@upnjatim.ac.id

Abstrak. *Diabetes melitus atau kencing manis adalah suatu gangguan kesehatan dengan kondisi dimana jumlah atau konsentrasi glukosa atau gula di dalam darah melebihi atau tidak sesuai dengan keadaan normal. Diabetes melitus merupakan salah satu penyakit yang tidak dapat disembuhkan tetapi dapat dikendalikan. Banyak penderita diabetes melitus yang menganggap remeh kadar gula darah karena kurangnya pengetahuan dan gaya hidup. Pendeteksian dini sebuah penyakit dapat dilakukan oleh sistem yang menggunakan algoritma C5.0. Dimana pada algoritma ini pohon klasifikasi yang terbentuk akan dirampingkan sehingga memudahkan interpretasinya. Dengan tahapan pruning pada algoritma C5.0 membuat tidak ada data yang tidak dapat diklasifikasi. Tahapan pruning menerima masukan berupa atribut yang menjadi node dan masing-masing nilainya yang menjadi cabang node. Untuk menentukan nilai tingkat galat pada setiap node menggunakan sebuah rumus yang didalamnya terdapat nilai z yang merupakan invers dari distribusi kumulatif normal baku dengan parameter a . Besar nilai a mempengaruhi besar nilai z . Dengan menggunakan jumlah data yang sama, uji coba pembentukan pohon klasifikasi dengan variasi nilai $a = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Dari hasil percobaan nilai $a = 0.2$ dipilih sebagai yang paling optimal karena memiliki nilai rata-rata persentase jumlah data salah klasifikasi yang paling kecil diantara empat variasi nilai a lainnya dengan 14.39%. Selain itu, dengan adanya tahapan pruning ini membuat tidak ada data yang tidak dapat diklasifikasikan.*

Kata Kunci : *klasifikasi, algoritma C5.0, pruning, diagnosis, diabetes mellitus*

Diabetes melitus atau kencing manis adalah suatu gangguan kesehatan dengan kondisi dimana jumlah atau konsentrasi glukosa atau gula di dalam darah melebihi atau tidak sesuai dengan keadaan normal, sehingga salah satu hal yang terpenting bagi penderita diabetes melitus adalah pengendalian kadar gula darah. Hal tersebut terjadi karena pankreas tidak dapat memproduksi insulin yang cukup atau karena tubuh tidak mampu secara efektif menggunakan insulin yang telah di produksi. Penyakit diabetes melitus dapat menyerang semua lapisan umur. Diabetes sering disebut dengan *mother of disease* karena dapat menjadi penyebab munculnya penyakit komplikasi seperti hipertensi, penyakit jantung dan pembuluh darah, stroke, gagal ginjal, dan kebutaan, serta kerusakan jangka panjang meliputi gangguan dan kegagalan fungsi berbagai organ terutama mata, ginjal, syaraf, jantung, dan pembuluh darah (Toharin, et al., 2015). Diabetes melitus merupakan salah satu penyakit yang tidak dapat disembuhkan tetapi dapat dikendalikan.

Pendeteksian dini sebuah penyakit dapat dilakukan oleh sistem yang menggunakan algoritma pengolahan data. Algoritma C5.0 merupakan salah satu algoritma klasifikasi data. Dengan algoritma C5.0 sebuah sistem dapat dibangun untuk menciptakan sebuah pohon keputusan yang dapat digunakan untuk mendiagnosis sebuah penyakit, salah satunya penyakit diabetes melitus. Pada algoritma C5.0 dapat dilakukan tahapan *pruning* untuk menjadikan pohon keputusan yang lebih baik. *Pruning* menjadi salah satu cara untuk merampingkan struktur dari *decision tree* sehingga proses generalisasi datanya menjadi lebih baik, dan memudahkan interpretasinya (Munawaroh, et al., 2013).

Sebuah penelitian (Anggraeny, et al., 2018) yang menggunakan dataset Pima Indian diabetes melakukan pengisian data sederhana untuk menggantikan nilai kosong pada beberapa variable, seperti pengisian nilai nol (0), rata-rata, median, dan acak. Hasilnya menunjukkan bahwa

nilai kosong yang diisi dengan nol memberikan peningkatan akurasi dibandingkan dengan penghapusan data yang memiliki nilai kosong tersebut. Oleh karena itu, pada penelitian ini, data yang memiliki nilai kosong akan diisi dengan nol.

Pruning menerima masukan berupa atribut yang menjadi *node* dan masing-masing nilainya yang menjadi cabang *node*. Tujuan dari *pruning* ini adalah membuat sebuah *subtree* yang diketahui memiliki nilai tingkat galat yang besar untuk menghasilkan sebuah kelas hasil dari nilai probabilitas yang besar. Sehingga nantinya tidak ada lagi data yang tidak dapat diklasifikasi seperti pada algoritma terdahulunya, ID3 (Patengko, et al., 2016).

Dengan begitu pada penelitian ini akan dilakukan tahapan *pruning* pada algoritma C5.0 untuk mendapatkan pohon klasifikasi terbaik yang nantinya dapat digunakan untuk diagnosis penyakit diabetes melitus sehingga masyarakat dapat mendekteksi penyakit tersebut baik untuk dirinya sendiri maupun orang-orang didekatnya, sehingga lebih dapat menjaga kesehatan dan gaya hidupnya.

Permasalahan

Berdasarkan latar belakang diatas, permasalahan pada penelitian ini adalah bagaimana cara menerapkan *pruning* pada algoritma C5.0 untuk mendiagnosis penyakit diabetes melitus.

Tujuan Penelitian

Tujuan dari penelitian ini adalah membuat pohon keputusan menggunakan algoritma C5.0 yang didalamnya terdapat tahapan *pruning* untuk mendiagnosis penyakit diabetes melitus.

Manfaat Penelitian

Manfaat dari penelitian ini adalah memberikan hasil diagnosis penyakit diabetes melitus berdasarkan kondisi tubuh seseorang menggunakan pohon keputusan yang terbentuk melalui tahapan algoritma C5.0 termasuk *pruning* didalamnya.

I. Metodologi

Penelitian ini dikembangkan melalui pengolahan data kondisi tubuh manusia yang diambil dari dataset *Kaggle* yang nantinya diolah menggunakan algoritma C5.0 dengan tahapan *pruning* didalamnya untuk mengetahui hasil diagnosis penyakit diabetes melitus berdasarkan pohon keputusan yang terbentuk.

C5.0

Algoritma C5.0 merupakan penyempurnaan dari algoritma terdahulu yang dibentuk oleh Ross Quinlan pada tahun 1987, yaitu algoritma ID3 dan C4.5. Strategi pengembangan *decision tree* dengan menggunakan algoritma C5.0 adalah sebagai berikut:

1. Menghitung *information gain* setiap atribut.
2. Membuat *node* berdasarkan atribut yang memiliki *information gain* tertinggi.
3. Cabang dikembangkan untuk tiap nilai yang diketahui dari atribut *node*.
4. Jika sampel seluruhnya berisi kelas yang sama, maka *node* tersebut memiliki *leaf* dan dilabeli dengan kelas tersebut.
5. Jika tidak, hitung *information gain* setiap atribut yang tersisa yang dapat memisahkan *record* ke dalam kelas-kelas individual.
6. Algoritma menggunakan proses yang sama secara *rekursif* atau perulangan membentuk pohon keputusan.
7. Partisi *rekursif* berakhir hanya ketika satu dari kondisi-kondisi berikut terpenuhi:
 - a. Seluruh *record* pada *node* tertentu memiliki kelas yang sama.
 - b. Tidak ada atribut yang tersisa pada *record* yang dapat dipartisi lebih lanjut. Dalam kasus ini suara mayoritas digunakan. *Node* tersebut memiliki *leaf node* dan dilabeli dengan kelas yang menjadi mayoritas dalam *record* yang ada.
 - c. Tidak ada *record* didalam cabang yang kosong. Dalam kasus ini, *leaf* terbentuk dengan mayoritas kelas sebagai label *record* tersebut (Yusuf, 2007).

Atribut dengan nilai *information gain* tertinggi dipilih sebagai *parent* bagi *node*

selanjutnya. Algoritma ini membentuk pohon keputusan dengan cara pembagian dan menguasai sampel secara *rekursif* dari atas ke bawah. Algoritma ini dimulai dengan semua data yang dijadikan akar dari pohon keputusan sedangkan atribut yang dipilih menjadi pembagi bagi sampel tersebut.

Pruning pada C5.0

Pruning menjadi salah satu cara untuk merampingkan struktur dari *decision tree* sehingga proses generalisasi datanya menjadi lebih baik, dan memudahkan interpretasinya. Tahapan ini hanya mengolah data latih untuk membentuk pohon keputusan dan tidak membutuhkan data uji untuk mengukur tingkat galat selama proses *pruning*. Pengukuran tingkat galat dihitung menggunakan persamaan (1) berikut:

$$e = \frac{f + \frac{z^2}{2N} + z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \quad (1)$$

dimana:

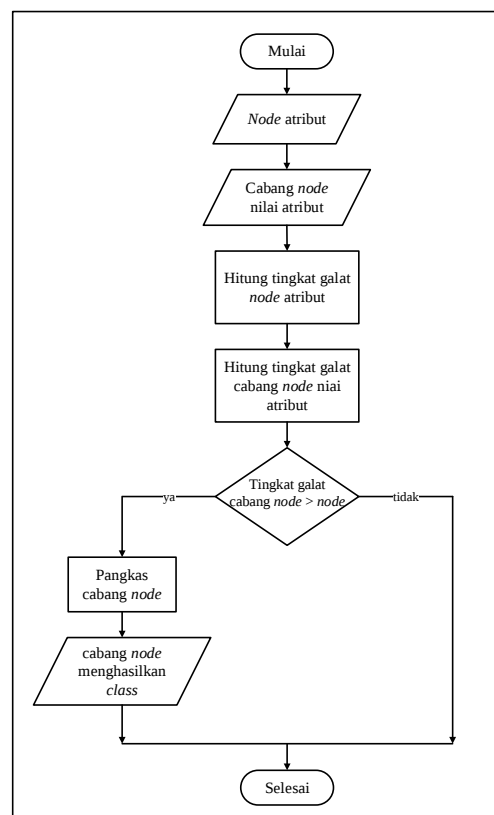
e = tingkat galat pada *node* (*error rate*).

f = perbandingan jumlah data yang salah dalam klasifikasi dengan jumlah keseluruhan data pada *node*, $\frac{E}{N}$, dengan E adalah jumlah galat.

N = jumlah data pada *node*.

z = nilai *invers* dari distribusi kumulatif normal baku dengan α sebagai parameter sesuai masukan dari pengguna.

Cara kerja *pre-pruning* adalah dengan menghitung dulu nilai *information gain* untuk mengetahui nilai *parent* dan *child*. Setelah *parent* dan *child* diketahui kemudian dihitung nilai erornya menggunakan persamaan (1), jika nilai eror *child* lebih kecil *parent* maka *parent* membentuk *subtree* lagi, tapi sebaliknya jika nilai eror *child* lebih besar dari *parent* maka *pruning* dilakukan dan pembentukan *subtree* berhenti (Munawaroh, et al., 2013).



Gambar 1. Tahapan *pruning*

Seperti pada gambar (1), keluaran pada tahapan *pruning* ini adalah berupa cabang *node* memiliki *output class* jika terjadi proses pemangkasan. Jika pada tahapan ini cabang *node* belum juga menghasilkan *output class*, maka akan dilanjutkan dengan menghitung *information gain* untuk setiap atribut tersisa, sesuai dengan tahapan algoritma C5.0.

II. Hasil dan Pembahasan

Pohon keputusan untuk mendiagnosis penyakit diabetes melitus menggunakan algoritma C5.0 yang didalamnya terdapat tahapan *pruning*. Pada persamaan (1) *pruning*, terdapat nilai z yang merupakan *invers* dari distribusi kumulatif normal baku. Dimana kata baku tersebut merujuk pada nilai standart distribusi kumulatif, yaitu kondisi dimana rata-rata sama dengan nol dan standart deviasi sama dengan

satu. Sedangkan α digunakan sebagai parameter untuk menentukan nilai z tersebut.

Untuk itu, akan dicoba sebanyak lima nilai untuk α tersebut sebagai parameter. Nilai yang akan dicoba adalah kelipatan 0.1, yaitu $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Sebagai pembanding dari nilai-nilai tersebut, akan dilakukan *running* program sebanyak lima kali untuk masing-masing nilai dan akan dicari nilai rata-rata dari persentase salah klasifikasi. Hal ini dilakukan untuk pemilihan nilai α yang optimal untuk klasifikasi menggunakan algoritma C5.0.

Persiapan Data

Data penelitian yang digunakan pada penelitian ini diambil dari *kaggle* (Pima Indians Diabetes Database | Kaggle) (Kaggle, 2016). Adapun keterangan dari dataset yang digunakan sebagai berikut:

1. Jumlah : 768 data
2. Format : data kondisi tubuh dalam ekstensi .xls
3. Bentuk : numerik
4. Atribut : terdapat 8 atribut dengan 1 atribut *class*
5. Waktu Pengambilan : 20 Februari 2019

Dataset ini berasal dari National Institute of Diabetes and Digestive and Kidney Diseases, India, yang digunakan untuk mendiagnosis apakah pasien menderita diabetes melitus atau tidak, berdasarkan pengukuran komputasi tertentu. Secara khusus, semua pasien pada dataset ini adalah wanita setidaknya berumur 21 tahun dari warisan suku India Pima.

Dari dataset yang digunakan diatas akan melalui *preprocessing data* yaitu pengambilan data yang relevan dengan gejala penyakit diabetes melitus yang akan digunakan untuk penelitian dan mengubah data numerik menjadi kategorik. Setelah data telah bertipe kategorik, data diambil sebanyak 70% dari data keseluruhan atau 538 data untuk yang nantinya diolah menjadi pohon keputusan dan melalui tahapan *pruning*.

Pruning

Setelah menghitung nilai *information gain* setiap atribut dan memperoleh atribut yang

memiliki nilai *information gain* tertinggi, maka secara otomatis pula diperoleh *node* dari atribut tersebut dan cabang *node* dari semua nilai atributnya. Pada tahapan ini akan menghitung nilai tingkat galat pada *node* dan cabang *node* tersebut dan keduanya akan dibandingkan. Hasil tingkat galat keduanya akan menentukan suatu cabang *node* apakah akan menghasilkan *output* atau tidak. Oleh karena itu, hasil perhitungan *pruning* ini akan disimpan pada tabel “tree”. Dimana pada tabel ini menyimpan data pohon keputusan yang nantinya terbentuk. Persamaan (1) *pruning* akan diterapkan pada program melalui baris perintah yang disajikan pada kode (1).

```
Sub tingkat_galat(j_nol, j_satu, jumlah)
//nilai alfa = 0.2
z = -0.841621234
N = jumlah

//mencari jumlah class terkecil untuk
nilai E, f=E/N
If j_nol < j_satu Then
    f = j_nol / N
Else
    f = j_satu / N
End If

//deklarasi atribut pembantu untuk
menampung hasil pembagian
Dim pembagi(5) As Decimal
pembagi(0) = (z ^ 2) / (2 * N)
pembagi(1) = f / N
pembagi(2) = (f ^ 2) / N
pembagi(3) = (z ^ 2) / (4 * (N ^ 2))
pembagi(4) = (z ^ 2) / N

//memasukkan variabel-variabel ke
rumus pruning
galat = Round((f + (pembagi(0)) + z *
(Math.Sqrt((pembagi(1)) -
(pembagi(2)) + (pembagi(3))))
/ 1 + (pembagi(4))), 6)

End Sub
```

Kode 1. Baris perintah menghitung tingkat galat *node*

Pada kode (1) variabel *j_nol*, *j_satu*, dan *jumlah* adalah variabel yang menyimpan nilai jumlah data yang memiliki *class Hasil* = 0, jumlah data yang memiliki *class Hasil* = 1, dan jumlah keduanya. Selain menggunakan tiga parameter tersebut, untuk menghitung nilai tingkat galat diperlukan juga nilai α atau alfa

yang nantinya mempengaruhi besar nilai z yang merupakan nilai *invers* dari distribusi kumulatif normal baku. Dari tahap pertama perhitungan data latih awal dan dengan nilai $\alpha = 0.1$, maka diperoleh nilai galat untuk *node* dan masing-masing cabang *node*-nya seperti pada gambar (2).

level	node	nama_sifat	nilai_sifat	output	node_induk	tingkat_galat	tingkat_galat_induk
0	1	glukosa	rendah	--	--	0.165563	3.32946
0	2	glukosa	normal	--	--	0.2003	3.32946
0	3	glukosa	tinggi	--	--	0.364609	3.32946

Gambar 2. Hasil perhitungan tingkat galat untuk *pruning*

Gambar (2) menunjukkan hasil perhitungan tingkat galat dan disimpan pada basis data tabel “tree”. Kolom “tingkat_galat” merupakan nilai tingkat galat untuk setiap cabang *node* dan akan dibandingkan dengan nilai tingkat galat di atasnya pada kolom “tingkat_galat_induk”. Dari gambar (2) tersebut juga dapat diketahui bahwa masing-masing cabang *node* tidak memiliki *output class* dan tidak akan dipangkat karena masing-masing nilai tingkat galatnya lebih kecil dari pada nilai tingkat galat *node* induknya.

Pembentukan Pohon Klasifikasi

Pembentukan pohon keputusan C5.0 berdasarkan pada nilai *information gain* pada setiap atribut. Dari data latih yang digunakan, pada langkah awal pembentukan pohon klasifikasi atau penentuan *root* akan dihitung nilai *information gain* dari delapan atribut, yaitu atribut *kehamilan*, *glukosa*, *bp*, *st*, *insulin*, *bmi*, *dpg*, dan *umur*. Dimana dari proses perhitungan tersebut diperoleh pohon klasifikasi seperti pada Lampiran (1). Lampiran (1) merupakan pohon klasifikasi yang didapatkan melalui proses perhitungan data latih awal yang ditampilkan menggunakan *TreeView* pada bahasa pemrograman VB.NET.

Hasil Percobaan Nilai α

Seperti yang telah dijelaskan sebelumnya, uji coba nilai α pada tahapan *pruning* menggunakan 5 variasi nilai yaitu $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ dengan dataset yang sama dan pembagian data latih dan data uji yang sama pula. Sebanyak 538 data digunakan sebagai data latih untuk menguji pengaruh nilai α yang berbeda pada tahapan *pruning*. Besar nilai α

mempengaruhi besar nilai z pada persamaan (1) *pruning* yang merupakan *invers* dari distribusi kumulatif normal baku. Besar nilai α dan z tersaji pada tabel (1).

Tabel 1. Nilai α dan z

	Nilai				
α	0.1	0.2	0.3	0.4	0.5
z	-1.282	-0.842	-0.524	-0.253	0

Uji coba ini dilakukan dengan cara *running* program untuk masing-masing variasi nilai α sebanyak lima kali. Sedangkan yang dijadikan pembanding adalah rata-rata persentase salah klasifikasi data latih pohon klasifikasi *final* yang terpilih disetiap *running* program. Dengan begitu, nilai α yang menghasilkan rata-rata persentase salah klasifikasi terendah yang dinyatakan paling optimal. Tabel (2) merupakan hasil dari uji coba nilai α , dengan i adalah iterasi *running* program.

Tabel 2. Hasil uji coba nilai

$i \backslash \alpha$	0.1	0.2	0.3	0.4	0.5
ke-1	13.94%	9.29%	7.43%	13.75%	24.16%
ke-2	18.59%	17.47%	20.45%	23.42%	22.12%
ke-3	15.43%	16.17%	18.96%	22.86%	22.12%
ke-4	14.87%	17.66%	21.19%	16.36%	22.68%
ke-5	21.19%	11.34%	13.38%	18.77%	23.79%
Rata-rata	16.80%	14.39%	16.28%	19.03%	22.97%

Pada tabel (2) menyajikan rata-rata hasil persentase salah klasifikasi setiap variasi nilai α yang diambil dari lima kali iterasi *running* program disetiap variasi nilainya. Tabel tersebut menunjukkan bahwa nilai α juga sangat mempengaruhi hasil klasifikasi dari pohon klasifikasi yang terbentuk. Nilai $\alpha = 0.2$ dipilih sebagai yang paling optimal karena memiliki nilai rata-rata persentase salah klasifikasi yang paling kecil diantara empat variasi nilai α lainnya dengan 14.39%.

III. Simpulan

Berdasarkan hasil uji coba penelitian yang telah dilakukan, diperoleh sejumlah simpulan antara lain:

1. Sistem diagnosis penyakit diabetes melitus dapat dibangun menggunakan algoritma C5.0.
2. Nilai α pada tahapan *pruning* paling optimal adalah sebesar 0.2 yang didapat dari hasil pengujian lima variasi nilai α , $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5\}$.
3. Dengan tambahan tahapan *pruning* pada algoritma C5.0 membuat tidak ada data yang tidak dapat diklasifikasi.

IV. Daftar Pustaka

1. Anggraeny, F. et al., 2018. *Analysis of Simple Data Imputation in Disease Dataset*. Denpasar, Atlantis Press, pp. 471-475.
2. Anon., n.d. *Pima Indians Diabetes Database* | Kaggle. [Online]
Available at:
<https://www.kaggle.com/uciml/pima-indians-diabetes-database>
[Accessed 4 Maret 2019].
3. Dwiyantri, Y., Herdianti, A. & Puspitasari, S. Y., 2017. Memprediksi Status Berlangganan Klien Bank Pada Kampanye Pemasaran Langsung Dengan Menggunakan Metode Klasifikasi Dengan Algoritma C5.0. *e_Proceeding of Engineering*, pp. 31-38.
4. Munawaroh, H., Khotimah, B. & Kustiyahningsih, Y., 2013. Perbandingan Algoritma Id3 Dan C5.0 Dalam Identifikasi Penjurusan Siswa Sma. *Jurnal Sarjana Teknik Informatika*, 1(1), pp. 1-12.
5. Patengko, A., Purnomo, H. & Tanone, R., 2016. *Sistem Pakar Diagnosa Penyakit Diabetes Melitus Menggunakan Algoritma Iterative Dichotomizer (ID3) Berbasis Android*, Salatiga: Universitas Kristen Satya Wacana.
6. Toharin, S. N., Cahyati S.KM, M. W. & Zainafree MH.Kes., d. I., 2015. Hubungan Modifikasi Gaya Hidup Dan Kepatuhan Konsumsi Obat Antidiabetik Dengan Kadar Gula Darah Pada Penderita Diabetes Melitus Tipe 2 Di Rs Qim Batang Tahun 2013. *Unnes Journal of Public Health*, 4(2), pp. 153-161.
7. Yusuf, Y., 2007. *Perbandingan Performansi Algoritma Decision Tree C5.0, CART, Dan CHAID: Kasus Prediksi Status Resiko Kredit Di Bank X*. Yogyakarta, Universitas Katolik Parahyangan, pp. B59-B62.

Lampiran 1. Pohon Klasifikasi

