



Penerapan *Algoritma C4.5* dalam Mengidentifikasi Karakteristik Pasien Beresiko Diabetes

Nuradha¹, Andi riski ramadani², Hazriani³, Yuyun⁴

^{1,2,3}Sistem Komputer, Program Pascasarjana Universitas Handayani Makassar

⁴Badan Riset dan Inovasi Nasional (BRIN)

nuradhagaffar@gmail.com

Abstract

Diabetes Mellitus is a disease characterized by an increase in glucose as well as an abnormal rise in blood sugar concentration due to insulin deficiency. The International Diabetes Federation (IDF) reports that in 2021, approximately 540 million people worldwide were affected by diabetes, and this number is expected to increase further if the general public's lack of awareness about symptoms that can trigger the diabetes disease continues. This research aims to implement the C4.5 algorithm in predicting diabetes mellitus based on acquired data. The amount of data used is 300 records, where 90% of the data serves as training data and the remaining 10% is test data. The data consists of 6 attributes: age, gender, hypertension, glucose, heart disease, and BMI (Body Mass Index). Based on Gain calculations, the Glucose attribute becomes the root of the decision tree. The tested data in this study achieved an accuracy rate of 77%, precision of 82%, and recall of 64%.

Keywords: C4.5 algorithm, diabetes, decision tree

Abstrak

Diabetes Melitus adalah penyakit yang ditandai dengan kenaikan glukosa serta kenaikan konsentrasi gula darah yang abnormal disebabkan oleh defisiensi insulin. Organisasi *Internasional Diabetes Federation* (IDF) melaporkan bahwa Pada tahun 2021 sebanyak 540 juta jiwa di seluruh dunia yang menderita diabetes, angka tersebut akan semakin meningkat jika ketidaktahuan masyarakat umum tentang gejala yang dapat memicu penyakit diabetes. Penelitian ini bertujuan untuk mengimplementasikan algoritma C4.5 dalam memprediksi penyakit diabetes mellitus, berdasarkan data yang telah diperoleh. Jumlah data yang digunakan adalah 300 data, dimana 90% data merupakan data training dan 10% nya merupakan data uji, data tersebut memiliki 6 atribut yaitu umur, gender, hipertensi, glukosa, penyakit jantung, dan BMI (Body Massa Index), berdasarkan hasil perhitungan *Gain*, atribut Glukosa menjadi akar utama pohon keputusan. Data yang telah di uji pada penelitian ini memiliki nilai akurasi sebesar 77%, *precision* 82% dan *recall* 64%.

Kata kunci: Algoritma C4.5, diabetes, pohon keputusan

1. Pendahuluan

Diabetes atau biasa disebut kencing manis adalah salah satu penyakit yang mematikan, penyakit diabetes dapat membunuh lebih dari satu juta jiwa setiap tahunnya dan siapapun dapat terkena [1]. Diabetes adalah penyakit yang ditandai dengan tingginya kadar gula darah yang disebabkan oleh sensitivitas insuli yang rendah dari jaringan [2]. *Organisasi Internasional Diabetes Federation* (IDF) melaporkan bahwa pada tahun 2021 sebanyak 540 juta jiwa di seluruh dunia menderita diabetes atau setara dengan angka prevalensi sebesar 10,5% populasi orang dewasa (20-79) menderita diabetes. Keterlambatan dalam diagnosa penyakit diabetes dapat menyebabkan meningkatnya jumlah

penderita diabetes. Meningkatnya jumlah penderita diabetes dikarenakan diabetes dikenal sebagai silent killer. Hal ini mengacu pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes, penderita kadang mengetahui ketika sudah terjadi komplikasi tanpa adanya gejala di awal. Untuk menekan jumlah penderita penyakit diabetes yang semakin bertambah, bisa dilakukan deteksi dini yang dapat dilakukan oleh tenaga ahli [3].

Untuk melakukan deteksi dini penyakit diabetes, dapat dikembangkan suatu sistem untuk memprediksi penyakit dengan memanfaatkan berbagai metode, salah satu metode yang digunakan yaitu metode data mining

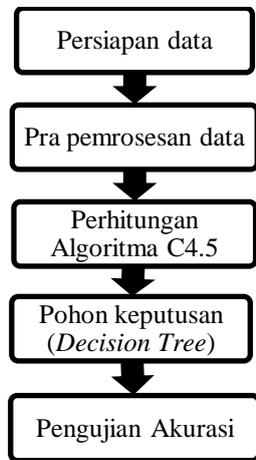
dengan prinsip klasifikasi. Metode ini dapat mengolah data berjumlah besar.

Teknik klasifikasi data mining dapat membantu untuk menentukan penyakit diabetes melitus atau kencing manis. Hasil dari klasifikasi data mining dapat digunakan untuk bahan tambahan edukasi masyarakat tentang bahaya penyakit diabetes mellitus ini[4]. Banyak algoritma yang bisa digunakan dalam klasifikasi contohnya Algoritma C4.5, Algoritma Naïve Bayes, Algoritma Support Vektor Machine dan lain sebagainya.

Algoritma yang digunakan untuk menentukan penyakit diabetes adalah algoritma C4.5. Algoritma C4.5 dapat digunakan untuk membantu dalam mengklasifikasikan diagnosa penyakit diabetes mellitus yang terbagi dalam beberapa atribut. Berdasarkan permasalahan tersebut maka akan diterapkan algoritma C4.5 untuk memprediksi penyakit diabetes.

2. Metode Penelitian

Pada gambar 1 metode penelitian terlihat tahapan metode dalam menyelesaikan penelitian ini yaitu persiapan data, pra pemrosesan, perhitungan algoritma C4.5, pohon keputusan (*Decision Tree*) dan pengujian Akurasi.



Gambar 1. Tahapan metode penelitian

2.1. Persiapan Data

Data yang digunakan pada penelitian ini diperoleh dari Kaggle, yakni *Diabetes prediction dataset* oleh mohammed Mustafa. Data yang digunakan adalah data medis dari pasien diabetes yang memuat data diabetes tahun 2019. Dataset yang digunakan yaitu 300 data terdiri 6 atribut yaitu umur, gender, hipertensi, glukosa, penyakit jantung, dan BMI (Body Massa Index), selain itu terdapat satu variabel Outcome yang terdiri dari dua kelas keputusan: positif (1) dan Negatif (0).

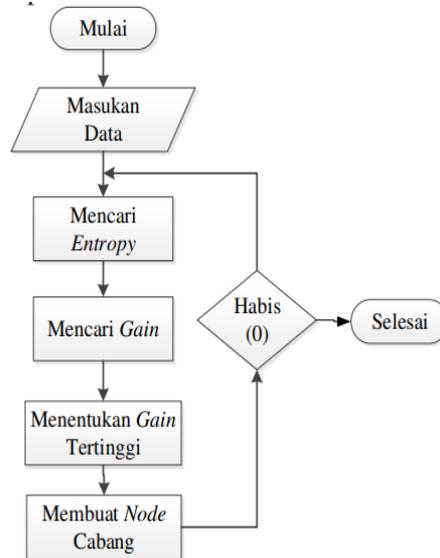
2.2. Pra Pemrosesan Data

Dilakukan pra proses data yaitu data diubah ke dalam bentuk kategori. Pembagian kategori yang digunakan dapat dilihat pada tabel berikut:

No	Atribut	Nilai	Kategori
1	Umur	21-59	Dewasa
		>=60	Lansia
		<=140	Baik
2	Glukosa	141-199	Sedang
		>=200	Buruk
3	BMI	<=18.5	Kurang
		18.6-29.9	Normal
		>=30	Obesitas

2.3. Perhitungan Algoritma C.45

Algoritma C4.5 adalah sebuah algoritma yang digunakan untuk membangun *decision tree* atau pohon keputusan, pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan digunakan untuk menemukan hubungan antara sejumlah calon variabel input dengan variabel target [6]. Proses pada pohon keputusan adalah mengubah bentuk data (tabel) ke dalam model pohon, kemudian mengubah model pohon menjadi rule, dan menyederhanakan rule. Sebuah pohon keputusan merupakan hasil dari perhitungan algoritma C4.5 [7]. Flowchart Algoritma C4.5 dalam membangun sebuah pohon keputusan terlihat pada gambar 2.



Gambar 2. Flowchart Algoritma C4.5 [8]

Pada gambar 2, hal yang dilakukan pertama kali adalah memasukkan data yang telah ditransformasi ke beberapa atribut, kemudian menghitung nilai entropy dan nilai gain, untuk memperoleh gain tertinggi yang akan menjadi atribut akar atau root dari pohon keputusan, untuk menghitung nilai entropy dapat dilakukan dengan menggunakan rumus [9], persamaan 1

$$Entropy(S) = \sum_{i=1}^n - pi * log_2 pi \tag{1}$$

Pada persamaan 1, S merupakan himpunan kasus, n merupakan Jumlah partisi dalam atribut dan pi adalah proporsi dari Si terhadap S.

Sedangkan untuk menghitung nilai gain dalam memilih sebuah atribut menjadi akar, dilakukan perhitungan nilai atribut yang ada, digunakan rumus [9], persamaan 2

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i) \quad (2)$$

Pada persamaan 2, S merupakan Himpunan kasus, A merupakan atribut, n adalah Jumlah partisi dalam atribut, |S_i| merupakan Jumlah kasus pada partisi ke-I, dan |S| adalah Jumlah kasus. Kemudian dilanjutkan dengan menentukan gain tertinggi yang akan menjadi atribut akar dari pohon keputusan, setelah gain tertinggi didapatkan maka dilanjutkan dengan proses pembuatan node cabang masing-masing nilai, jika setiap kasus dalam cabang tersebut telah berada satu kelas yang sama maka proses perhitungan selesai, tapi jika kasus berbeda kelas maka kembali ke perhitungan entropy dan begitu seterusnya hingga semua kasus berada dalam kelas yang sama. Setelah perhitungan selesai, maka dihasilkan pohon keputusan (*decision tree*).

2.4. Pengujian

Pengujian dilakukan untuk menentukan tingkat akurasi dari pohon keputusan algoritma C4.5. dalam pengujian ini dataset dibagi menjadi dua bagian, yaitu sebanyak 100 data untuk dijadikan sebagai data training dan sebanyak 50 data dijadikan sebagai data uji. Pengujian tingkat akurasi ini dilakukan dengan menggunakan metode Confusion matrix.

Confusion matrix adalah metode yang digunakan untuk melakukan perhitungan tingkat akurasi pada data mining. *Confusion matrix* memuat informasi tentang klasifikasi yang diprediksi dengan benar oleh sebuah sistem klasifikasi. Terdapat tiga parameter yang dihitung, yaitu akurasi, recall, dan precision[10].

Pada pengukuran performa model klasifikasi menggunakan confusion matrix, terdapat empat nilai yang digunakan untuk merepresentasikan hasil klasifikasi yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). True Positive adalah jumlah data positif yang diprediksi benar, nilai True Negative adalah jumlah data negatif yang diprediksi benar, False Positive adalah jumlah data jumlah data negatif yang diprediksi positif, dan False negatif adalah jumlah data positif yang diprediksi negatif. Berdasarkan nilai tersebut dapat diperoleh nilai akurasi, precision, dan recall. Akurasi menggambarkan seberapa akurat algoritma dalam mengklasifikasikan data. Nilai akurasi dapat diperoleh dengan persamaan 3 berikut [11].

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Presisi adalah rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif, nilai presisi diperoleh dengan menggunakan rumus pada persamaan 4 berikut [12].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif, nilai recall diperoleh dengan menggunakan rumus pada persamaan 5 berikut [12].

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

F-Measure adalah metrik pengukuran yang mengkombinasikan presisi dan recall untuk menghasilkan pengukuran tunggal yang merepresentasikan keefektifan suatu klasifikasi, *f1-score* merupakan rata-rata dari presisi dan recall [13]. *f1-score* menyediakan hasil pengukuran presisi dan recall yang seimbang, perhitungan *f1-score* dilakukan menggunakan rumus pada persamaan (6) [14].

$$F - \text{Measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

3. Hasil dan Pembahasan

3.1. Persiapan Data

Dataset yang digunakan ini berisi data dari pasien diabetes dan non diabetes dari 300 dataset yang terlihat pada tabel 2 yang merupakan data awal.

Tabel 2. Data awal

N	um	Gend	Hiperte	Gluko	jantu	BM	Diabe
o	ur	er	nsi	sa	ng	I	tes
1	13	Male	0	85	0	19.9	0
2	14	Male	0	100	0	19.69	0
3	14	Female	0	260	0	14.99	1
4	14	Female	0	280	0	15.13	1
5	16	Male	0	220	0	15.69	1
6	16	Male	0	200	0	15.63	1
7	17	Female	0	200	0	17.91	1
8	17	Female	0	260	0	16.44	1
9	17	Male	0	280	0	15.4	1
10	18	Female	0	280	0	14.32	1
29	80	Female	1	160	0	20.8	0
29	80	Female	1	160	0	20.8	0
29	80	Male	1	145	0	22.68	0
29	80	Male	1	160	0	21.52	0
29	80	Male	1	145	0	22.61	0
29	80	Female	1	159	1	23.24	1
29	80	Female	1	159	1	24.24	1
6	80	Male	1	155	1	19.19	1
29	80	Female	1	155	1	24.24	1
7	80	Female	1	145	1	89	1

29	80	Female	1	145	0	22.05	1
29	80	Female	0	145	1	36.7	1
30	80	Male	0	159	0	21.7	1

Berikutnya adalah hasil perubahan data dari data numerik menjadi data kategori, yang akan di analisis menggunakan algoritma C4.5 yang terlihat pada tabel 3 yang merupakan data kategori.

Tabel 3. Data Kategori

No	umur	Gender	Hipertensi	Glukosa	Penyakit jantung	BMI	Diabetes
1	Remaja	Perempuan	Tidak	Buruk	Tidak	Obesitas	Positif
2	Remaja	Laki-laki	Tidak	Buruk	Tidak	Kurang	Positif
3	Remaja	Laki-laki	Tidak	Baik	Tidak	Kurang	Negatif
4	Remaja	Laki-laki	Tidak	Baik	Tidak	Kurang	Negatif
5	Remaja	Laki-laki	Tidak	Baik	Tidak	Kurang	Negatif
6	Remaja	Perempuan	Tidak	Baik	Tidak	Kurang	Negatif
7	Remaja	Perempuan	Tidak	Baik	Tidak	Kurang	Negatif
8	Remaja	Laki-laki	Ya	Sedang	Tidak	Obesitas	Positif
9	Remaja	Perempuan	Tidak	Baik	Tidak	Kurang	Negatif
10	Remaja	Laki-laki	Tidak	Baik	Tidak	Kurang	Negatif
293	Dewasa	Laki-laki	Tidak	Buruk	Tidak	Obesitas	Positif
294	Dewasa	Laki-laki	Tidak	Baik	Tidak	Kurang	Negatif
295	Dewasa	Perempuan	Tidak	Baik	Tidak	Obesitas	Negatif
296	Dewasa	Perempuan	Tidak	Buruk	Tidak	Obesitas	Positif
297	Dewasa	Laki-laki	Tidak	Buruk	Ya	Obesitas	Positif
298	Lansia	Laki-laki	Ya	Sedang	Tidak	Normal	Negatif
299	Lansia	Laki-laki	Ya	Sedang	Tidak	Normal	Negatif
300	Lansia	Laki-laki	Ya	Sedang	Tidak	Normal	Negatif

Setelah data dikategorikan seperti tabel 3 di atas, maka selanjutnya adalah dengan membagi dataset menjadi data yang berbeda yaitu data training dan serta data uji, diambil 270 data untuk training data dan 30 data untuk data uji.

3.2. Perhitungan Algoritma C4.5

Setelah data ditransformasi ke dalam bentuk kategori, maka selanjutnya adalah melakukan perhitungan menggunakan algoritma C4.5 untuk menghasilkan decision tree atau pohon keputusan. Pohon keputusan dibuat untuk mengetahui algoritma C4.5 dapat mendiagnosa penyakit diabetes.

Langkah pertama adalah mencari nilai entropy dan gain, proses pencarian entropy total dan gain dilakukan

dengan menggunakan rumus hitung entropy dan gain pada masing-masing atribut data. Berikut adalah contoh perhitungan nilai entropy pada atribut glukosa(sedang) serta gain atribut glukosa pada persamaan 7 dan 8.

$$Entropy(sedang) = \left(-\frac{31}{98}\right) * \log_2\left(\frac{31}{98}\right) + \left(-\frac{67}{98}\right) * \log_2\left(\frac{67}{98}\right) = 0.9003 \tag{7}$$

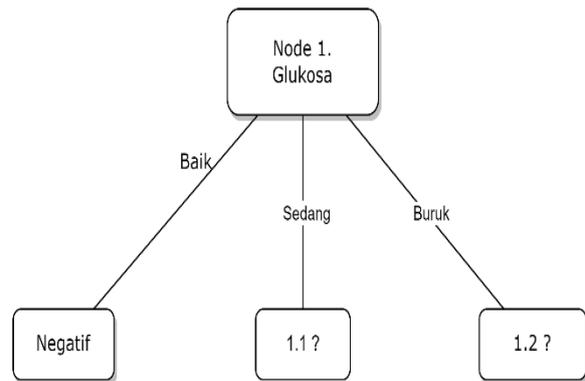
$$Gain(Glukosa) = (0.86872) - \left(\left(\frac{19}{130}\right) * 0\right) + \left(\left(\frac{98}{130}\right) * 0.9003\right) + \left(\left(\frac{13}{130}\right) * 0\right) = 1.6021 \tag{8}$$

Perhitungan nilai entropy dan gain di atas adalah pada atribut glukosa, untuk atribut lain dilakukan hal yang sama seperti diatas, tabel 4 merupakan tabel hasil perhitungan nilai entropy dan gain.

Tabel 4. Hasil Hitung Gain Tertinggi (Node 1)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropy	Gain
Total Glukosa	270	126	144	0.996	
a					1.887
Baik	23	0	23	0	
Sedang	202	94	108	0.9965	
Buruk	45	32	13	0.8673	

Berdasarkan hasil perhitungan diatas, diketahui bahwa nilai gain terbesar adalah pada atribut glukosa, sehingga atribut glukosa menjadi node (akar). Pada gambar 3 adalah gambar pohon sementara dari tabel 4 hasil yang merupakan hasil hitung Gain tertinggi.



Gambar 3. Pohon Keputusan Sementara (Node 1)

Selanjutnya adalah mencari node cabang atribut glukosa = sedang yaitu node 1.1 dan atribut glukosa=buruk yaitu node 1.2, dengan mencari nilai gain tertinggi untuk node cabang. Berikut adalah tabel hasil perhitungan nilai gain tertinggi, atribut glukosa=sedang, dan glukosa = Buruk yang terlihat pada tabel 5 dan tabel 6.

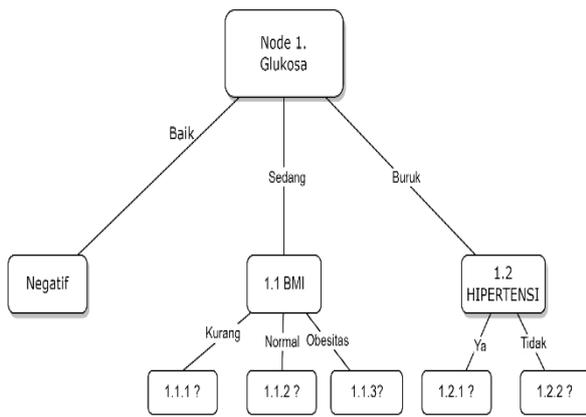
Tabel 5. Hasil Hitung Gain Tertinggi (Node 1.1)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropy	Gain
Total BMI	98	31	67	0.90	
Kurang	11	3	8	0.84	
Normal	156	60	96	0.96	
Obesitas	35	31	4	0.51	

Tabel 6. Hasil Hitung *Gain* Tertinggi (Node 1.2)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropi hy	Gain
Total Hipertensi	98	31	67	0.90	1.28
Ya	9	4	5	0.99	
Tidak	36	28	8	0.76	

Berdasarkan tabel 5 dan 6 hasil perhitungan di atas dapat diketahui bahwa atribut yang menjadi node cabang dari glukosa=sedang adalah BMI, dimana BMI terdiri dari tiga sub atribut yaitu kurang, normal, dan obesitas. Sub atribut tersebut belum memiliki label, sehingga perlu dilakukan perhitungan lebih lanjut dan akan menjadi node cabang selanjutnya yaitu node 1.1.1, 1.1.2, dan node 1.1.3. sedangkan node cabang dari glukosa =buruk adalah hipertensi, dimana sub atribut dari hipertensi belum memiliki label, sehingga sub atribut tersebut menjadi node cabang selanjutnya yaitu node 1.2.1 dan 1.2.2 yang ditunjukkan pada gambar 4 yang merupakan pohon keputusan sementara.



Gambar 4. Pohon Keputusan Sementara

Tabel 7 merupakan hasil perhitungan node 1.1.1, 1.1.2, dan node 1.1.3 yang merupakan data dari gender dimana terdiri dari dua data yaitu perempuan dan laki-laki.

Tabel 7. Hasil Hitung *Gain* Tertinggi (Node 1.1.1)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropi hy	Gain
Total Gender	11	3	8	0.84	1.53
Perempuan	3	0	3	0	
Laki-Laki	8	3	5	0.95	

Dari hasil perhitungan node 1.1.1 yaitu dapat dilihat bahwa atribut yang menjadi node cabang selanjutnya adalah atribut gender, dimana atribut gender memiliki dua sub atribut, sub atribut Perempuan telah memiliki label yaitu negatif, sedangkan gender=laki-laki belum memiliki label, dan akan menjadi node cabang selanjutnya yaitu 1.1.1.1. Untuk Perhitungan node 1.1.1.1 dilakukan hal yang sama, dari hasil perhitungan node 1.1.1.1, atribut yang memiliki *gain* tertinggi adalah penyakit jantung. Atribut penyakit jantung memiliki dua sub atribut yaitu ya dan tidak, penyakit jantung = ya sudah memiliki label yaitu positif sedangkan penyakit jantung = tidak belum memiliki label, itu artinya masih

samar dan perlu dilakukan perhitungan untuk node selanjutnya yaitu node 1.1.1.1 dan tabel 8 menunjukkan hasil hitung *Gain* tertinggi berdasarkan total umur yang terbagi menjadi tiga kriteria yaitu remaja, dewasa dan lansia.

Tabel 8. Hasil Hitung *Gain* Tertinggi (Node 1.1.2)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropi hy	Gain
Total Umur	156	60	96	0.96	1.72
Remaja	28	1	27	0.22	
Dewasa	20	12	8	0.97	
Lansia	108	47	61	0.98	

Tabel hasil perhitungan node 1.1.2 tersebut menghasilkan nilai *gain* tertinggi terdapat pada atribut umur, dimana atribut umur memiliki tiga sub atribut yang belum memiliki label, itu artinya atribut umur menjadi node selanjutnya yaitu node 1.1.2.1, 1.1.2.2, dan 1.1.2.3. Pada tabel 9 menunjukkan hasil hitung terhadap total hipertensi.

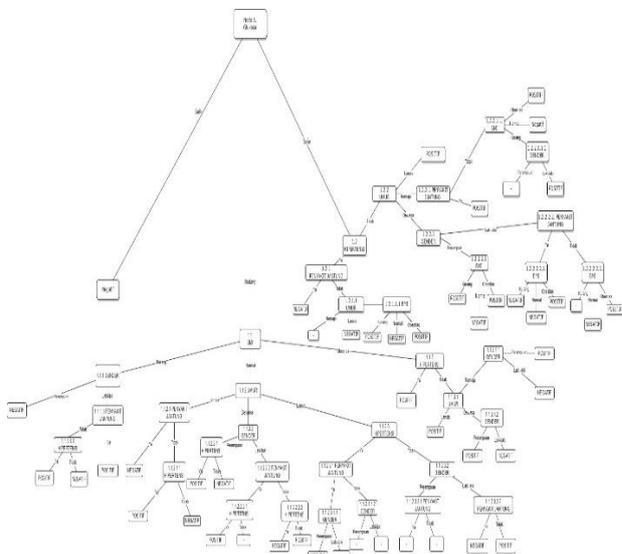
Tabel 9. Hasil Hitung *Gain* Tertinggi (Node 1.1.3)

	Jml kasus(S)	Positif(S1)	Negatif(S2)	Entropi hy	Gain
Total Hipertensi	35	31	4	0.51	0.94
Ya	13	13	0	0	
Tidak	22	18	4	0.68	

Tabel hasil perhitungan node 1.1.3 tersebut menghasilkan nilai *gain* tertinggi terdapat pada atribut hipertensi, dimana atribut hipertensi memiliki 2 sub atribut, hipertensi=ya memiliki label positif sedangkan hipertensi = tidak belum memiliki label, itu artinya atribut hipertensi = tidak menjadi node selanjutnya yaitu node 1.1.3.1, dalam hal ini perhitungan node selanjutnya dilakukan hal yang sama seperti perhitungan node sebelumnya.

3.3. Pohon Keputusan (*Decision Tree*)

Berdasarkan hasil perhitungan node pada gambar 5, maka dapat digambarkan pohon keputusan akhir.



Gambar 5. Pohon Keputusan Akhir

Rule/aturan pada tabel 10 menunjukkan aturan pohon keputusan.

Tabel 10. Aturan Pohon Keputusan

Rule/Aturan	
R1	IF Glukosa = Baik THEN diabetes= Negatif
R2	IF Glukosa = Sedang AND BMI = Kurang AND Gender = Perempuan THEN diabetes = Negatif
R3	IF Glukosa = Sedang AND BMI = Kurang AND Gender = Laki-laki AND Penyakit jantung = ya THEN diabetes = Positif
R4	IF Glukosa = Sedang AND BMI = Kurang AND Gender = Laki-laki AND Penyakit jantung = tidak AND hipertensi=ya THEN diabetes = Positif
R5	IF Glukosa = Sedang AND BMI = Kurang AND Gender = Laki-laki AND Penyakit jantung = tidak AND hipertensi=Tidak THEN diabetes = Negatif
R30	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Ya THEN diabetes = Negatif
R31	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Tidak AND Umur = Remaja THEN Diabetes = (-)
R32	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Tidak AND Umur = Lansia THEN Diabetes = Negatif
R33	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Tidak AND Umur = Dewasa AND BMI = Kurang THEN Diabetes = Positif
R34	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Tidak AND Umur = Dewasa AND BMI= Normal THEN Diabetes = Negatif
R35	IF Glukosa = Buruk AND Hipertensi = Ya AND Penyakit Jantung= Tidak AND Umur = Dewasa AND BMI = Obesitas THEN Diabetes = Positif
R36	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Lansia THEN Diabetes = Positif
R37	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Remaja AND Penyakit Jantung=Ya THEN Diabetes = Positif
R38	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Remaja AND Penyakit Jantung=Tidak AND BMI=Obesitas THEN Diabetes = Positif
R39	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Remaja AND Penyakit Jantung=Tidak AND BMI= Normal THEN Diabetes = Negatif
R40	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Remaja AND Penyakit Jantung=Tidak AND BMI= Kurang AND Gender= Perempuan THEN Diabetes = (-)
R41	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Remaja AND Penyakit Jantung=Tidak AND BMI= Kurang AND Gender= Laki-laki THEN Diabetes = Positif
R42	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Perempuan AND BMI= Kurang THEN Diabetes = Positif
R43	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Perempuan AND BMI= Normal THEN Diabetes = Negatif
R44	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Perempuan AND BMI= Obesitas THEN Diabetes = Positif
R45	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Ya AND BMI=Kurang THEN Diabetes = Negatif

R46	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Ya AND BMI=Normal THEN Diabetes = Negatif
R47	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Ya AND BMI=Obesitas THEN Diabetes = Positif
R48	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Tidak AND BMI=Kurang THEN Diabetes = (-)
R49	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Tidak AND BMI=Normal THEN Diabetes = Negatif
R50	IF Glukosa = Buruk AND Hipertensi = Tidak AND Umur= Dewasa AND Gender= Laki-laki AND Penyakit Jantung=Tidak AND BMI=Obesitas THEN Diabetes = Positif

3.4. Uji Akurasi

Setelah proses training terhadap 270 data dan menghasilkan pohon keputusan, maka selanjutnya dilakukan pengujian hasil klasifikasi terhadap data uji yaitu 30 data uji yang terlihat pada tabel 11.

Tabel 11. Data Uji

N o	um ur	Gend er	Hipe r tns i	Glu kos a	Penyakit jantung	BM I	Dia bete s	Pre diks i
1	Re maj a	Laki-laki	Tidak	Bai k	Tidak	Nor mal	Neg atif	Neg atif
2	Re maj a	Laki-laki	Tidak	Bai k	Tidak	Nor mal	Neg atif	Neg atif
3	Re maj a	Pere mpua n	Tidak	Bur uk	Tidak	Obe sitas	Posi tif	Posi tif
4	Re maj a	Pere mpua n	Tidak	Bur uk	Tidak	Obe sitas	Posi tif	Posi tif
5	Re maj a	Laki-laki	Tidak	Bur uk	Tidak	Kur ang	Posi tif	Posi tif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Nor mal	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Nor mal	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Nor mal	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Kur ang	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Nor mal	Posi tif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Kur ang	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Ya	Sed ang	Ya	Kur ang	Neg atif	Neg atif
2	Lan sia	Pere mpua n	Tidak	Sed ang	Ya	Obe sitas	Posi tif	Posi tif
2	Lan sia	Laki-laki	Tidak	Sed ang	Tidak	Nor mal	Posi tif	Posi tif
3	Lan sia	Laki-laki	Tidak	Sed ang	Tidak	Nor mal	Neg atif	Posi tif
0	Lan sia	Laki-laki	Tidak	Sed ang	Tidak	Nor mal	Neg atif	Posi tif

Data uji pada tabel 10, kami melakukan uji akurasi dengan menggunakan metode *confusion matrix*. Pada

tabel 12 menunjukkan perhitungan hasil *confussion matrix* dan nilai akurasi .

Tabel 12. Perhitungan *Confussion Matrix*

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	9	5
Negatif	2	14

Dari tabel 11 diketahui bahwa kelas TP (*True Positif*) yaitu 9 data, kelas TN (*True Negatif*) yaitu 14 data, FP (*False Positif*) yaitu 2, dan FN (*False Negatif*) yaitu 5, maka didapatkan nilai akurasi yaitu 77%, *Precision* 82%, dan *recall* 64%.

4. Kesimpulan

Berdasarkan hasil implementasi algoritma C4.5 terhadap data diabetes adalah algoritma C4.5 dengan 6 parameter atribut yaitu glukosa, umur, hipertensi, penyakit jantung, BMI (*Body Massa Index*), dan gender, atribut glukosa menjadi parameter utama. Hasil uji akurasi berdasarkan 300 dataset yang dibagi menjadi 90%:10% didapatkan nilai akurasi 77%, *precision* 82%, dan *Recall* 64% serta memiliki nilai *F-Measure* yaitu 72%.

Daftar Rujukan

- [1] A. Najib, D. Nurcahyono, And R. P. P. Setiawan, "Klasifikasi Diagnosa Penyakit Diabetes Mellitus (Dm) Menggunakan Algoritma C4.4," *Just Ti J. Sains Terap. Teknol. Inf.*, Vol. 11, No. 2, P. 47, Jul. 2019, Doi: 10.46964/Justti.V11i2.153.
- [2] M. Sukmadani Rusdi, "Hipoglikemia Pada Pasien Diabetes Melitus," *J. Syifa Sci. Clin. Res.*, Vol. 2, No. 2, Pp. 83–90, Aug. 2020, Doi: 10.37311/Jsscr.V2i2.4575.
- [3] O. Ozougwu, "The Pathogenesis And Pathophysiology Of Type 1 And Type 2 Diabetes Mellitus," *J. Physiol. Pathophysiol.*, Vol. 4, No. 4, Pp. 46–57, Sep. 2013, Doi: 10.5897/Jpap2013.0001.
- [4] I. Menarianti, "Klasifikasi Data Mining Dalam Menentukan Pemberian Kredit Bagi Nasabah Koperasi," *J. Ilm. Teknosains*, Vol. 1, No. 1, 2015.
- [5] A. A. Robbani, A. M. Siregar, And D. S. Kusumaningrum, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5," No. 1, 2022.
- [6] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, Vol. 14, No. 2, P. 217, Sep. 2018, Doi: 10.33480/Pilar.V14i2.926.
- [7] Y. Mardi, "Data Mining: Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, Vol. 2, No. 2, Pp. 213–219, Feb. 2017, Doi: 10.22202/Ei.2016 V2i2.1465.
- [8] Y. A. Fiandra, S. Defit, And Y. Yuhandri, "Penerapan Algoritma C4.5 Untuk Klasifikasi Data Rekam Medis Berdasarkan International Classification Diseases (Icd-10)," *J. Resti Rekayasa Sist. Dan Teknol. Inf.*, Vol. 1, No. 2, Pp. 82–89, Oct. 2017, Doi: 10.29207/Resti.V1i2.48.
- [9] B. Sugara, D. Widyatmoko, B. S. Prakoso, And D. M. Saputro, "Penerapan Algoritma C4.5 Untuk Deteksi Dini Gangguan Autisme Pada Anak," 2018.
- [10] E. M. Sipayung, H. Maharani, And I. Zefanya, "Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier," *J. Sist. Inf.*, Vol. 8, No. 2, 2016.
- [11] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, And R. Budiarto, "Evaluating Trust Prediction And Confusion Matrix Measures For Web Services Ranking," *Ieee Access*, Vol. 8, Pp. 90847–90861, 2020, Doi: 10.1109/Access.2020.2994222.

- [12] E. W. H. Candana, I. G. A. Gunadi, And D. G. H. Divayana, "Jurnal Ilmu Komputer Indonesia (Jik) , Volume 6 , No : 2, November 2021 P-Issn : 2615-2703 (Print) Dan E-Issn: 2615-2711 (Online)," Vol. 6, 2021.
- [13] R. Novendri, R. Andreswari, And O. N. Pratiwi, "Implementasi Data Mining Untuk Memprediksi Customer Churn Menggunakan Algoritma Naive Bayes".
- [14] T. T. Hanifa And S. Al-Faraby, "Analisis Churn Prediction Pada Data Pelanggan Pt. Telekomunikasi Dengan Logistic Regression Dan Underbagging".