PENERAPAN METODE *DECISION TREE* DALAM KLASIFIKASI PENDERITA PENYAKIT DIABETES MENGGUNAKAN ALGORITMA C4.5

Revica Mashitapasha, Fitri Damayanti, Doni Abdul Fatah
Sistem Informasi, Universitas Trunojoyo Madura
Jalan Raya Telang Kamal Bangkalan, Indonesia
190441100084@student.trunojoyo.ac.id

ABSTRAK

Diabetes merupakan penyakit serius yang disebabkan oleh ketidakmampuan pankreas dalam memproduksi insulin secara optimal. Jika tidak dideteksi dan ditangani dengan baik, diabetes dapat menyebabkan komplikasi berat seperti kebutaan, gagal jantung, bahkan kematian. Faktor-faktor penyebab diabetes antara lain keturunan, berat badan, usia, tekanan darah serta gaya hidup. Namun, gejala awal sering kali tidak terlihat sehingga diperlukan metode deteksi dini untuk mengurangi risiko komplikasi. Penelitian ini bertujuan untuk menerapkan metode klasifikasi dalam data mining menggunakan algoritma C4.5 dalam membentuk pohon Keputusan (*Decision Tree*) untuk mengklasifikasikan penderita penyakit diabetes. Penelitian ini menggunakan dataset dari *Kaggle* yang berisi 2.000 data dengan 4 atribut dan 1 label target. Data dilakukan normalisasi dan transformasi sebelum diklasifikasikan menggunakan algoritma Decision Tree C4.5. Pengujian dilakukan menggunakan *confusion matrix* untuk mengevaluasi performa model berdasarkan akurasi, presisi, dan *recall*. Dari hasil pengujian menggunakan confusion matrix, diperoleh akurasi sebesar 77%, presisi sebesar 50%, dan *recall* sebesar 74 %. Hasil ini menunjukkan bahwa model klasifikasi yang diterapkan memiliki kinerja yang cukup baik dalam mengidentifikasi penderita diabetes.

Kata kunci : *Diabetes, Decision Tree, Algoritma C4.5, Klasifikasi, Confusion Matrix*

1. PENDAHULUAN

Diabetes merupakan penyakit yang tidak menular namun termasuk cukup serius bagi manusia yang dikarenakan pankreas tidak mampu menghasilkan insulin secara optimal[1].

Sel tubuh yang tidak bekerja secara normal dalam menyerap glukosa mengakibatkan penumpukan gula dalam darah sehingga hal tersebut menimbulkan gangguan organ tubuh yang lainnya[2].

Diabetes dapat disebabkan oleh beberapa variabel, yaitu obesitas, gaya hidup, makanan yang tidak sehat, tekanan darah, tinggi serta dari riwayat keluarga yang terkena diabetes[1].

Tidak hanya menyebabkan kematian, diabetes juga merupakan penyebab utama terjadinya kebutaan, penyakit jantung, serta gagal ginjal. Organisasi *Internasional Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 463 juta orang pada usia 20-79 tahun di dunia menderita diabetes pada tahun 2019 dimana jumlah itu setara dengan angka prevalensi sebesar 9,3% dari jumlah total penduduk di usia yang sama. Berdasarkan *gender*, IDF memperkirakan prevalensi diabetes di tahun 2019 meningkat 9% pada Perempuan dan 9,65% pada laki-laki. Angka orang yang menderita diabetes diperkirakan akan terus meningkat, mencapai 578 juta pada tahun 2030 dan 700 juta pada tahun 2045[3].

Untuk mengatasi permasalahan tersebut, perlu dilakukan pendeteksian penyakit diabetes sejak dini. Hal ini dikarenakan gejala awal diabetes sering kali tidak terlihat, diharapkan pendeteksian sejak dini ini mampu menurunkan risiko komplikasi yang terkait dengan diabetes[1].

Dalam penelitian ini,pendeteksian dilakukan dengan memanfaatkan *data mining*.

Data mining merupakan pencarian informasi dari kumpulan data yang besar. Data mining dapat mengidentifikasi hubungan dan keterkaitan antara variabel yang mungkin tidak terlihat secara langsung[4].

Data mining memiliki beberapa teknik yang digunakan, seperti regresi, prediksi, klasifikasi, asosiasi, dan clustering[5].

Penelitian ini menggunakan teknik klasifikasi untuk menentukan atau mendeteksi penyakit diabetes berdasarkan beberapa variabel yang telah diperoleh.

Klasifikasi merupakan pengelompokkan suatu data ke dalam kelas atau label berdasarkan atribut yang sama[4].

Klasifikasi biasanya dilakukan dengan menggunakan algoritma *machine learning*. Algoritma ini akan mempelajari pola atau hubungan antar fitur-fitur yang dimiliki oleh objek-objek dalam data latih atau *data training* dan kemudian membuat model atau aturan-aturan untuk memprediksi label atau kelas dari objek-objek yang belum diidentifikasi kelasnya dalam data uji atau *data testing*.

Di dalam klasifikasi, salah satu metode yang cukup umum untuk digunakan yaitu metode *decision tree, decision tree* merupakan metode yang merepresentasikan ketentuan dari banyaknya data ke dalam bentuk pohon keputusan. Pohon keputusan adalah struktur yang membagi jumlah data besar menjadi sejumlah kecil data[6].

Setiap titik percabangan pada pohon ini akan mewakili suatu atribut data dimana atribut tersebut akan membagi data ke dalam kelompok-kelompok

yang lebih kecil. Proses klasifikasi dimulai dari akar pohon terus mengikuti cabang-cabang hingga mencapai cabang yang terkecil yang menunjukkan kategori dari data tersebut.

Dalam penelitian yang dikemukakan oleh Andy Supriyadi pada tahun 2023, penelitian ini menggunakan data dari sumber daya manusia (SDM) Universitas Sebelas Maret serta wawancara yang dilakukan oleh peneliti yang berjumlah 8 atribut. Penelitian ini mengulik mengenai perbandingan antara metode *decision tree* dan *naïve bayes*,penelitian ini mengukur tingkat klasifikasi dosen berprestasi dalam lingkup Universitas Sebelas Maret. Metode *decision tree* menunjukkan kinerja yang lebih baik dibanding dengan metode *naïve bayes*. Metode *decision tree* mencapai akurasi sebesar 95,80%, *precision* 96,90% dan *recall* 95,90%. Sedangkan metode *naïvebayes* hanya mencapai akurasi sebesar 94,80% dengan *precision* sebesar 95,25% dan *recall* sebesar 95,90%. Hal ini mengindikasi bahwa *decision tree* lebih handal dalam mengklasifikasikan dosen berprestasi dibandingkan dengan *naïve bayes*[7].

2. TINJAUAN PUSTAKA

2.1. Diabetes Melitus

Diabetes melitus merupakan salah satu jenis penyakit yang tidak menular namun berbahaya dan banyak membunuh jutaan orang setiap tahunnya, menurut *World Health Organization* (WHO) jumlah penderita diabetes terus bertambah hingga empat kali lipat lebih banyak setiap tahunnya dibanding tahun-tahun sebelumnya[8].

Diabetes berpotensi paling tinggi menimbulkan penyakit komplikasi (memicu timbulnya penyakit yang lain) dikarenakan berkaitan dengan tingginya kadar gula di dalam darah sehingga berdampak langsung pada rusaknya pembuluh darah, saraf, dan struktur internal lainnya[9].

Oleh karena itu, sangat penting untuk mendeteksi diabetes sejak dini serta menerapkan gaya hidup sehat dan menjalani pengobatan yang tepat sehingga mencegah komplikasi penyakit serius[10].

2.2. *Data Mining*

Data mining merupakan proses Analisa terhadap data yang berjumlah besar yang bertujuan untuk menemukan suatu hubungan yang jelas sehingga dapat ditarik kesimpulan yang berguna bagi pemilik data[5].

Pada pemrosesan data dengan jumlah besar, data mining akan sangat diperlukan sehingga dapat dihasilkan sebuah informasi baru yang dapat bermanfaat sebagai sumber pengambilan keputusan selanjutnya. Berdasarkan tugasnya, *data mining* dikategorikan menjadi lima jenis utama yaitu memahami karakteristik data secara mendalam, mendeskripsikan data dengan menggunakan model statistik, membangun model untuk memprediksi nilai atau kelas suatu variabel, menemukan pola serta aturan yang digunakan, serta melakukan pemanggilan data[4].

Data mining membantu menemukan pola dan informasi baru dalam data dengan mengidentifikasi hubungan keterkaitan antar variabel yang mungkin tidak terlihat secara langsung. Pemrosesan data mining biasanya terdiri dari beberapa langkah mulai dari pemilihan data secara relevan, pemrosesan data, pengolahan data, dan analisis data. Semua Teknik ini digunakan untuk menyelesaikan proses data mining seperti klasifikasi, prediksi, clustering, regresi, asosiasi, dan lain-lain.

2.3. Klasifikasi

Klasifikasi merupakan proses pengelompokan data berdasarkan aturan tertentu yang diperoleh dari pembelajaran data latih. Klasifikasi data yang akan diuji dapat digunakan untuk mengkategorikan data uji. Proses klasifikasi ini termasuk dalam pembelajaran terawasi dimana komputer menggunakan atribut sebagai penanda karakteristik data yang sering disebut label atau kelas[11].

Tujuan klasifikasi adalah untuk memprediksi kelas atau kategori dari suatu data, seperti misalnya memprediksi apakah suatu email merupakan spam atau tidak, atau mengidentifikasi jenis penyakit berdasarkan gejalanya[4].

Dalam klasifikasi, terdapat target variabel kategori, misalnya penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, sedang dan rendah[12].

2.4. *Decision Tree*

Decision tree merupakan algoritma *machine learning* yang menggunakan seperangkat aturan untuk membuat keputusan dengan struktur seperti pohon. Algoritma ini memodelkan kemungkinan hasil, biaya sumber daya, utilitas, dan risiko atau kemungkinan yang mungkin terjadi[13].

Dalam model ini, setiap titik atau *node* pada pohon mencerminkan keputusan atau penilaian terhadap suatu atribut, cabang-cabangnya mencerminkan hasil dari penilaian tersebut, dan daun mencerminkan keputusan akhir atau hasil[14]. .Menurut[15], terdapat beberapa jenis *node* yang ada di *decision tree*, yaitu:

a. *Root node* (*node* akar): *node* awal yang mewakili dataset secara keseluruhan dan memecah data sesuai dengan kondisi yang paling informatif.
b. *Internal rode* (*node* internal): titik percabangan yang menguji atribut tertentu dalam dataset, Dimana setiap cabang menunjukkan hasil pengujian tersebut.
c. *Leaf node* (*node* daun): *node* akhir yang menunjukkan Keputusan atau hasil dari proses. Setiap jalur yang menghubungkan akar ke daun merupakan aturan klasifikasi atau prediksi.

2.5. Algoritma C4.5

Algoritma C4.5 merupakan algoritma klasifikasi yang berfungsi untuk menghasilkan prediksi berdasarkan data yang sudah dikategorikan[14].

Algoritma C4.5 merupakan pengembangan dari algoritma ID3 dengan prinsip kerja yang mirip. Namun yang membedakan dengan algoritma C4.5 yaitu pada langkah perhitungan, di mana algoritma C4.5 dimulai dengan mencari nilai *entropy* dan memilih atribut berdasarkan *gain ratio* tertinggi[16].

Algoritma ini memungkinkan pembangunan pohon keputusan secara iteratif yang memungkinkan pengambilan keputusan berdasarkan atribut dan nilai-nilai yang ada dalam data[17].

2.6. *Confusion Matrix*

Confusion matrix merupakan metode umum yang digunakan untuk menghitung akurasi pada konsep *data mining*. *Confusion matrix* berguna untuk menelaah sejauh mana pengkategorian dapat mengenali *tuple* dari kelas yang berlawanan dengan baik. Tabel *confusion matrix* terdiri dari dua Tingkat yaitu Tingkat yang dianggap positif dan Tingkat yang dianggap negatif[11].

Berikut ini adalah tabel *confusion matrix* yang dipaparkan pada Tabel 1:

Tabel 1. *Confusion matrix*

Confusion Matrix		***Nilai Prediksi***	
		Positif	***Negatif***
Nilai Sebenarnya	*Positif*	TP	FP
	Negatif	FN	TN

Berdasarkan tabel *confusion matrix* diatas:

a. TP berarti hasil klasifikasi positif yang diprediksi dengan benar oleh pengklasifikasi.
b. FP berarti data yang sebenarnya negatif namun diprediksi positif oleh pengklasifikasi.
c. FN berarti data yang sebenarnya positif namun diprediksi negatif oleh pengklasifikasi.
d. TN berarti hasil klasifikasi negatif yang diprediksi dengan benar oleh pengklasifikasi.

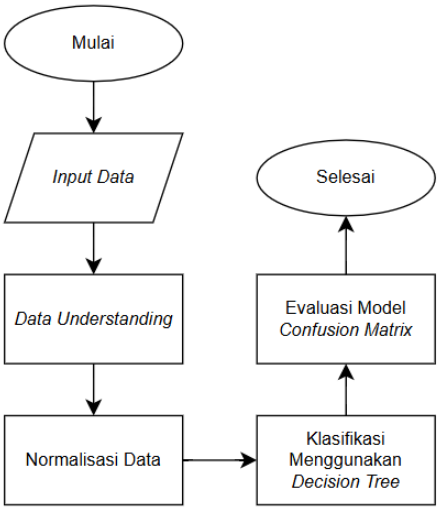
2.7. Penelitian Terdahulu

Aldiyansyah, Ade Irma Purnamasari, dan Irfan Ali, pada tahun 2024 meneliti tentang perbandingan Tingkat akurasi dalam klasifikasi penerima bantuan social dengan menggunakan metode *decision tree*. Hasil pengujiannya memiliki akurasi sebesar 97,86% dengan nilai presisi untuk kelas 0 dan kelas 1 sebesar 92,30%. *Recall* kelas 0 mencapai 99,20%, sementara kelas 1 mencapai 85,71%.*F1- score* kelas 0 sebesar 98.81%, dan kelas 1 sebesar 88.88%[18].

Pada tahun 2025 J.M. Havidz Yasaf Al-Afghoni meneliti tentang klasifikasi jenis benih kacang yang bertujuan untuk mengelompokkan jenis kacang yang belum memiliki label. Penelitian ini menggunakan metode *decision tree* dan berhasil mendapatkan hasil evaluasi menggunakan *confusion matrix* adalah Akurasi: 94% Presisi: 94% *Recall*: 94% *F1-Score*: 94%[4].

Pada tahun 2023 iddrus dan dewi Wulan sari melakukan penelitian yang bertujuan untuk memprediksi mahasiswa *drop out* yang terdapat di Universitas Wiraraja. Peneliti menggunakan metode *decision tree*. Berdasarkan penelitian yang dilakukan menggunakan rapidminer mendapat hasil pengujian berupa nilai akurasi sebesar 81,82%[5].

3. METODE PENELITIAN



Gambar 1. *Flowchart* alur penelitian

3.1. Input Data

Data yang diperoleh merupakan dataset tabel dengan format csv hal ini cukup mempermudah karena data telah tertata sebagai tabel. Data berasal dari *Kaggle* dengan judul data berupa "Klasifikasi Penyakit Diabetes" yang berisi 2.000 data, 8 kelas, dan 1 label target.

3.2. *Data Understanding*

Data understanding (Pemahaman Data) merupakan tahap awal dari proses analisis data.tujuan dari tahap ini adalah mengumpulkan, memahami, dan menganalisis data yang akan digunakan lebih lanjut. *Point* ini juga dapat memberikan deskripsi tentang data, seperti informasi umum seperti ukuran, format, jenis data (numerik, kategori, teks), serta variabelnya. Dalam hal ini dapat dicari dimensi data, tipe dataset, jumlah atribut unik, hitungan data tiap kelas, dan data yang tidak lengkap.

3.3. Normalisasi Data

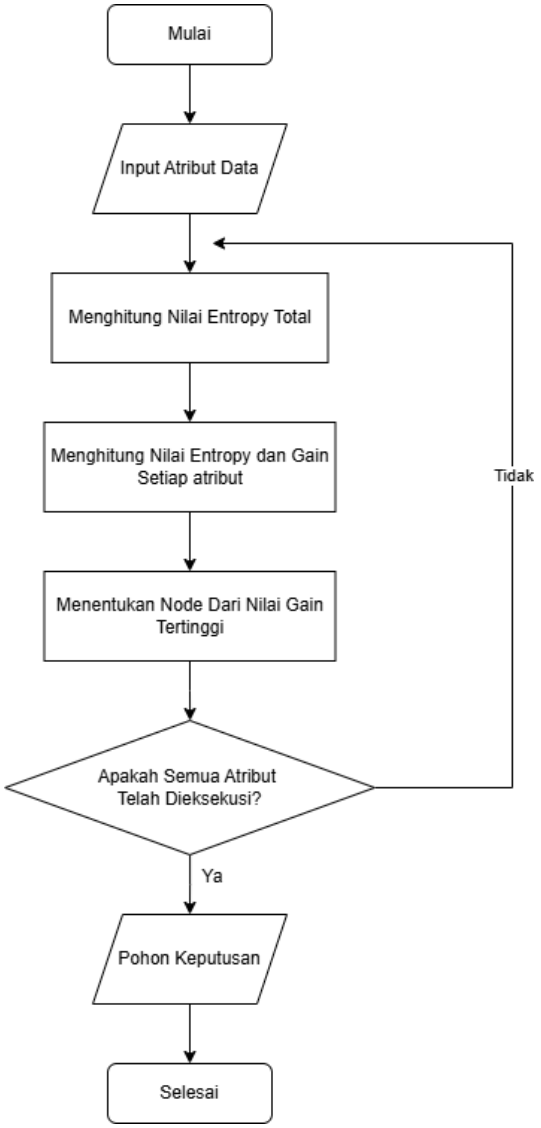
Normalisasi merupakan tahapan mengubah skala atau rentang kelas pada dataset sehingga kelas tersebut memiliki nilai yang sebanding. Didalam normalisasi data juga dilakukan pengelompokkan data menjadi beberapa label, tujuan dari normalisasi data adalah menyederhanakan tipe data sehingga data tersebut bisa lakukan proses klasifikasi dengan lebih mudah.

3.4. Klasifikasi *Decision Tree*

Proses pada *decision tree* yaitu mengubah bentuk data tabel dengan atribut yang menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan *tree*. *Decision tree* merupakan salah satu

metode klasifikasi yang terdiri dari Kumpulan node yang dihubungkan oleh cabang bergerak ke bawah dari *root node* hingga *leaf node*[19].

Berikut ini merupakan flowchart tahapan dalam penentuan node menggunakan Algoritma C4.5:



Gambar 2. *Flowchart* metode algoritma C4.5

Terdapat beberapa tahapan dalam membuat pohon Keputusan, yaitu:

a. Mempersiapkan *data training*. *Data training* merupakan data yang telah diambil dari data yang sebelumnya sudah dikelompokkan ke dalam kelas-kelas tertentu.
b. Menghitung *node* dari *tree*. *Node* akan diambil dari atribut yang akan dipilih dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang tertinggi akan menjadi *node* pertama. Sebelum menghitung nilai *gain* dari atribut, harus menghitung dulu nilai *entropy*nya.untuk menghitung nilai *entropy*, dapat digunakan rumus:

$$Entropy\ (S) = \sum_{i=1}^{n} -p^i\ log^2(p^i) \quad (1)$$

Keterangan:
S = Himpunan Kasus
n = Jumlah Partisi S
pi = Proporsi Si tterhadap S

c. Menghitung nilai *gain* menggunakan persamaan 2.

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

Keterangan:
S = Himpunan kasus
A = Fitur
.n = Jumlah Partisi S
|Si| = Proporsi Si terhadap S
|S| = Jumlah kasus dalam S

d. Agar semua *record* terpartisi, ulangi Langkah 2 dan 3.
e. Proses partisi *decision tree* akan berhenti Ketika:
   a. Semua *record* dalam simpul N mendapatkan kelas yang sama.
   b. Di dalam *record* tidak lagi terdapat atribut yang dipartisi.
   c. Tidak terdapat *record* pada *node* yang kosong

3.5. Evaluasi *Confusion Matrix*

Confusion matrix disebut sebagai "tabel kontingensi" yang merupakan tabel yang digunakan untuk mengevaluasi performa model klasifikasi. *Confusion matrix* menggambarkan hasil prediksi model klasifikasi dengan membandingkan prediksi yang dibuat oleh model dengan nilai data yang sebenarnya diamati. *Confusion matrix* memberikan informasi yang bermanfaat untuk mengevaluasi kinerja model *decision tree*. Berbagai *matrix* evaluasi yang dapat dihitung dari *confusion matrix*, yaitu:

a. Akurasi: rasio keseluruhan dari prediksi benar terhadap jumlah total atribut.
b. Presisi: rasio prediksi benar positif terhadap total prediksi positif (TP / (TP + FP)).
c. *Recall*: rasio prediksi benar positif terhadap total kelas positif sebenarnya (TP / (TP + FN)).

Persamaan yang digunakan untuk menghitung akurasi, presisi, dan *recall* diperlihatkan pada persamaan 3,4, dan 5.

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN}\ x\ 100\% \quad (3)$$

$$presisi = \frac{TP}{TP + FP}\ x\ 100\% \quad (4)$$

$$recall = \frac{TP}{TP + FN}\ x\ 100\% \quad (5)$$

Keterangan:
TP = True Positif
TN = True Negatif
FP = False Positif
FN = False Negatif

Confusion matrix membantu mengevaluasi dan memahami bagaimana model *decision tree* melakukan prediksi kelas target dan memberikan informasi detail tentang keakuratan dan kesalahannya. Hal ini membantu dalam evaluasi pemahaman performa model serta memungkinkan perbaikan atau penyesuaian yang diperlukan.

4. HASIL DAN PEMBAHASAN

4.1. Normalisasi Data

Berdasarkan data yang telah didapat, terdapat beberapa kelas atau label yang tidak diperlukan. Dalam penelitian ini, hanya akan fokus mencantumkan kelas yang dinilai berhubungan langsung dengan indicator diagnosa penyakit diabetes Dimana terdapat 4 kelas yaitu: *Glucose, Blood Pressure, BMI, dan Age* serta 1 label target yaitu *Outcome*. Berikut ini merupakan data mentah dalam penelitian ini:

Tabel 2. Data mentah

Glucose	BloodPressure	BMI	Age	Outcome
138	62	33.6	47	ya
84	82	38.2	23	tidak
145	0	44.2	31	ya
135	68	42.3	24	ya
139	62	40.7	21	tidak
173	78	46.5	58	tidak
99	72	25.6	28	tidak
194	80	26.1	67	tidak
83	65	36.8	24	tidak
89	90	33.5	42	tidak
99	68	32.8	33	tidak
125	70	28.9	45	ya
80	0	0	22	tidak
166	74	26.6	66	tidak
110	68	26	30	tidak

Setelah melewati analisis dan beberapa tahapan percobaan menggunakan data mentah, untuk mendapatkan hasil perhitungan menggunakan algoritma *decision tree* C4.5 perlu dilakukan transformasi data yang bertujuan untuk memfokuskan data dan kelas yang digunakan dalam perhitungan. Berikut ini merupakan tabel acuan transformasi data:

Tabel 3. Acuan transformasi data

Atribut	Nilai	Kategori
Glucose	<140	Baik
	141-199	Sedang
	>200	Buruk
Blood Pressure	<80	Normal
	81-89	Prahipertensi
	90-119	Hipertensi
	>120	Krisis
BMI	<18,5	Kurang
	18,6-29,9	Normal
	>30	Obesitas
Umur	21-59	Dewasa
	>60	Lansia

Perubahan data ini dilakukan berdasarkan acuan dari website Kementrian Kesehatan (KemenKes) RI, bahwa kadar glukosa dikategorikan baik jika berjumlah kurang dari 140 mg/dL, kategori sedang jika berjumlah 141-199 mg/dL, dan dikategorikan buruk jika bernilai lebih dari 200mg/dL. *Blood Pressure* dikategorikan normal jika nilainya kurang dari 80 mmhg, dikategorikan pra hipertensi jika ada dalam rentang nilai 81-89 mmhg, dikategorikan hipertensi jika berada direntang nilai 90-119 mmhg, dan dikatakan krisis jika *blood pressure* seseorang berada diatas 120 mmhg. Selanjutnya untuk *body massa index* (BMI) seseorang dikategorikan kurang (*underweight*) jika seseorang memiliki bmi kurang dari 18,5, dikategorikan normal jika seseorang memiliki bmi 18,6-29,9, dan dikategorikan obesitas (*overweight*) jika seseorang memiliki BMI lebih dari 30. Seseorang akan dikategorikan dewasa jika berumur 21-59 dan dikategorikan lansia jika berumur lebih dari 60. Hasil dari perubahan data numerik menjadi data kategori dari 2000 data diperoleh data seleksi yang dapat dilihat pada Tabel 4.

Tabel 4. Transformasi data

Glucose	BloodPressure	BMI	Age
baik	normal	obesitas	dewasa
baik	prahipertensi	obesitas	dewasa
sedang	normal	obesitas	dewasa
baik	normal	obesitas	dewasa
baik	normal	obesitas	dewasa
sedang	normal	obesitas	dewasa
baik	normal	normal	dewasa
sedang	normal	normal	lansia
baik	normal	obesitas	dewasa
baik	hipertensi	obesitas	dewasa
baik	normal	obesitas	dewasa
baik	normal	normal	dewasa
baik	normal	kurang	dewasa
sedang	normal	normal	lansia

4.2. Klasifikasi *Decision Tree* C4.5

Langkah awal dalam menentukan pohon Keputusan dengan menggunakan algoritma C4.5 yaitu dengan menentukan *data training* dan *data testing*. Umumnya pembagian data dibagi menjadi 80% *data training* dan 20% *data testing*. Jadi pada penelitian ini menggunakan 1600 data sebagai *data training* dan 400 data sebagai *data testing*.

Selanjutnya adalah menentukan nilai *entropy* dan *gain* pada masing-masing *node*. Untuk menghitung nilai *entropy* dan *gain* pada *node* pertama dapat menggunakan rumus pada persamaan (1) untuk mencari nilai *entropy*. Berikutnya untuk menghitung nilai *gain* menggunakan rumus pada persamaan (2). Berikut merupakan hasil tabel perhitungan *entropy* dan *gain* pada *node* 1:

Tabel 5. Nilai *entropy* dan *gain node 1*

Ket.	Total	Ya	Tidak	Entro	Gain
	1600	546	1054	0.92601	
Glucose					
Baik	1191	274	917	0.778114	0.581955
Sedang	409	272	137	0.919917	
Buruk	0	0	0		
BloodP					
Normal	1266	390	876	0.890918	0.014135
Prahipertensi	219	102	117	0.996613	
Hipertensi	113	54	59	0.998587	
Krisis	2	0	2		
BMI					
Kurang	29	2	27	0.362051	0.062193
Normal	587	104	483	0.673851	
Obesitas	984	440	544	0.991927	
Age					
Dewasa	1527	527	1000	0.929635	0.00105
Lansia	73	19	54	0.82716	

Dari tabel 5 dapat dilihat bahwa *gain* tertinggi ada pada atribut *glucose* yakni 0,581955 sehingga dapat disimpulkan aebagai akar dari pohon keputusan atau *node* awal yaitu *glucose*. Kemudian dilakukan kembali perhitungan nilai *entropy* dan *gain* untuk menentukan *node* 1.1. nilai yang dihitung berdasarkan atribut *glucose* baik dan sedang, dikarenakan tidak ada data *glucose* buruk. Perhitungan nilai *entropy* dan *gain* dapat dilihat pada tabel 6.

Tabel 6. Nilai *entropy* dan *gain node 1.1*

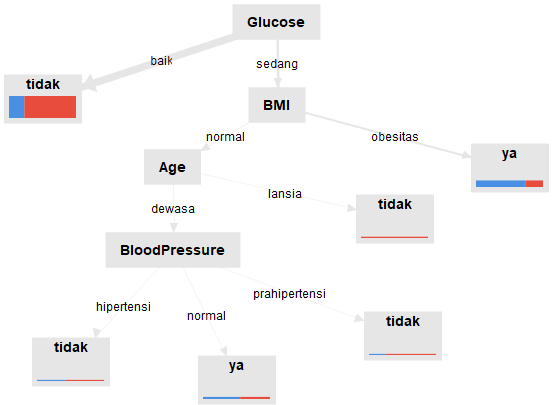
***Glucose* (Baik)**	**Total**	**Ya**	**Tidak**	***Entropy***	***Gain***
	1492	343	1149	0.777824	
BloodP					
Normal	987	209	778	0.74482	0.162814
Prahipertensi	134	46	88	0.927926	
Hipertensi	68	19	49	0.854648	
Krisis	2	0	2		
BMI					
Kurang	29	2	27	0.362051	0.185797
Normal	493	64	429	0.556937	
Obesitas	669	208	461	0.894223	
Age					
Dewasa	1156	271	885	0.785653	0.159202
Lansia	35	3	32	0.422001	

Berdasarkan hasil perhitungan tabel 6, dapat dilihat nilai *gain* tertinggi terdapat pada atribut BMI yakni sebesar 0,185797 sehingga dapat disimpulkan bahwa BMI dijadikan *node* 1.1. Selanjutnya perhitungan nilai *entropy* dan *gain node* 1.2 dapat dilihat pada tabel 7.

Tabel 7. Nilai *entropy* dan *gain node 1.2*

***Glucose* (Sedang)**	**Total**	**Ya**	**Tidak**	***Entropy***	***Gain***
	508	341	167	0.91363	
BloodP					
Normal	279	181	98	0.93518	0.1773933
Prahipertensi	85	56	29	0.92594	
Hipertensi	45	35	10	0.76421	
***Glucose* (Sedang)**	**Total**	**Ya**	**Tidak**	***Entropy***	***Gain***
	508	341	167	0.91363	
Krisis	0	0	0		
BMI					
Kurang	0	0	0		
Normal	94	40	54	0.98394	0.2156794
Obesitas	315	232	83	0.83197	
Umur					
Dewasa	371	256	115	0.89314	0.1879065
Lansia	38	16	22	0.9819	

Berdasarkan hasil perhitungan tabel 7, dihasilkan nilai *gain* tertinggi ada pada atribut BMI yakni 0,2156794 sehingga dapat disimpulkan bahwa BMI dijadikan *node* 1.2. Perhitungan *entropy* dan *gain* akan diulang dan diteruskan hingga *node* terakhir sehingga akan menghasilkan pohon Keputusan sebagai berikut:



Gambar 3. Pohon keputusan

Berdasarkan pohon keputusan pada gambar 2 menghasilkan *rule* sebagai berikut:

a. Jika glukosa Baik maka label = Tidak Diabetes
b. Jika glukosa Sedang, BMI Obesitas maka label = Diabetes
c. Jika glukosa Sedang, BMI Normal, dan Age Lansia maka label = Tidak Diabetes
d. Jika glukosa Sedang, BMI Normal, Age Dewasa, dan Blood Pressure Normal maka label = Diabetes
e. Jika glukosa Sedang, BMI Normal, Age Dewasa, dan Blood Pressure Prahipertensi maka label = Tidak Diabetes
f. Jika glukosa Sedang, BMI Normal, Age Dewasa, dan Blood Pressure Hipertensi maka label = Tidak Diabetes

4.3. Evaluasi *Confusion Matrix*

Dari hasil perhitungan dan pohon Keputusan yang telah didapat, selanjutnya akan dievaluasi menggunakan kriteria seperti akurasi, presisi, dan *recall*. Evaluasi *confusion matrix* menggunakan *software rapidminer*, hasil evaluasi data yang ada adalah sebagai berikut:

accuracy: 76.50%

	true tidak	true ya	class precision
pred. tidak	238	70	77.27%
pred. ya	24	68	73.91%
class recall	90.84%	49.28%	

Gambar 4. Hasil evaluasi *performance*

Berdasarkan hasil evaluasi Algoritma *Decision Tree* C4.5 dengan *data training* sebanyak 80% dan *data testing* sebanyak 20%, di dapatkan nilai *confusion matrix* dari setiap kelas. Nilai-nilai tersebut merupakan nilai True Positif(TP), True Negatif(TN), False Positif(FP), dan False Negatif(FN). Untuk mendapatkan hasil evaluasi model, perhitungan manual perlu dilakukan dengan menggunakan persamaan 3, 4, dan 5 diatas untuk menghitung nilai akurasi, presisi, dan *recall.* Berikut perhitungannya:

$$akurasi = \frac{238+70}{70+238+24+68}\ x\ 100\% = 77\ \%$$

$$presisi = \frac{70}{70+\ 68}\ x\ 100\% = 50{,}7\ \%$$

$$recall = \frac{70}{70+\ 24}\ x\ 100\% = 74\ \%$$

Berdasarkan perhitungan manual, didapatkan hasil nilai akurasi sebesar 77%, presisi sebesar 50,7%, dan nilai *recall* sebesar 74%. Hal tersebut menunjukkan hasil klasifikasi yang dilakukan termasuk dalam kelas cukup baik.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian ini, penerapan algoritma Decision Tree C4.5 dalam klasifikasi penderita diabetes menunjukkan kinerja yang cukup baik dengan akurasi sebesar 76,5%. Dalam penelitian ini, Algoritma C4.5 digunakan untuk menentukan apakah seseorang memiliki diabetes atau tidak. Perhitungan menggunakan Algoritma C4.5 membentuk *decision tree* yang memiliki 6 *rule* diharapkan bisa menjadi suatu informasi mengenai penentuan penyakit diabetes.Namun, masih terdapat beberapa keterbatasan yang perlu diperbaiki, seperti jumlah atribut yang terbatas dalam proses klasifikasi dan tingkat recall yang belum optimal. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan metode lain guna meningkatkan akurasi dan prediksi. Selain itu, penggunaan dataset yang lebih luas serta penambahan fitur yang lebih relevan, seperti riwayat keluarga dan pola konsumsi makanan, dapat membantu memperbaiki model klasifikasi sehingga hasilnya lebih akurat dalam mendeteksi penderita diabetes.

DAFTAR PUSTAKA

[1] W. D. Prasetya And B. Sujatmiko, "Rancang Bangun Aplikasi Dengan Perbandingan Metode K-Nearest Neighbor (Knn) Dan Naive Bayes Dalam Klasifikasi Penderita Penyakit Diabetes," *J. Informatics Comput. Sci.*, Vol. 3, No. 04, Pp. 515–525, 2022, Doi: 10.26740/Jinacs.V3n04.P515-525.

[2] I. M. Karo Karo And H. Hendriyana, "Klasifikasi Penderita Diabetes Menggunakan Algoritma Machine Learning Dan Z-Score," *J. Teknol. Terpadu*, Vol. 8, No. 2, Pp. 94–99, 2022, Doi: 10.54914/Jtt.V8i2.564.

[3] B. A. R. P. Wahyu, A. F. Farozi, C. P. Mahendra, And R. K. Hapsari, "Klasifikasi Penderita Penyakit Diabetes Berdasarkan Decision Tree," *Integer J. Inf. Technol.*, Vol. 8, No. 1, Pp. 80–89, 2023.

[4] J. M. H. Y. Al-Afghoni *Et Al.*, "Klasifikasi Jenis Benih Kacang Menggunakan Smote Dan Decision Tree C4 . 5," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 9, No. 1, Pp. 462–469, 2025.

[5] I. Iddrus And D. W. Sari, "Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja," *J. Adv. Res. Inform.*, Vol. 1, No. 02, Pp. 1–7, 2023, Doi: 10.24929/Jars.V1i02.2684.

[6] D. Septhya, K. Rahayu, S. Rabbani, And V. Fitria, "Implementation Of Decision Tree Algorithm And Support Vector Machine For Lung Cancer Classification Implementasi Algoritma Decision Tree Dan Support Vector Machine Untuk Klasifikasi Penyakit Kanker Paru," Vol. 3, No. April, Pp. 15–19, 2023.

[7] A. Supriyadi, "Perbandingan Algoritma Naive Bayes Dan Decision Tree (C4 . 5) Dalam Klasifikasi Dosen Berprestasi," Vol. 7, No. 1, Pp. 39–49, 2023.

[8] B. N. S. Naisah Marito Putry, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *Evolusi J. Sains Dan Manaj.*, Vol. 10, No. 1, 2022, Doi: 10.31294/Evolusi.V10i1.12514.

[9] N. Ningsih, A. Aprianto, And A. Angeline, "Pendekatan Data Science Untuk Deteksi Dini Diabetes Menggunakan Naive Bayes Classifier," *J. Inf. Syst. Hosp. Technol.*, Vol. 5, No. 1, Pp. 26–31, 2023, Doi: 10.37823/Insight.V5i1.300.

[10] C. A. Rahayu, "Prediksi Penderita Diabetes Menggunakan Metode Naive Bayes," *J. Inform. Dan Tek. Elektro Terap.*, Vol. 11, No. 3, 2023, Doi: 10.23960/Jitet.V11i3.3055.

[11] P. Riliandhita, Iqbal Maulana, "Klasifikasi Penentuan Status Kemiskinan Penduduk Kelurahan," Vol. 8, No. 2, Pp. 1791–1796, 2024.

[12] M. E. S. Mutiara Novita Angie, Dita Azizah, Annisa Hertyanti, "Analisis Faktor Penentu Dalam Ulasan Dan Rekomendasi Terhadap Keputusan Konsumen Di Media Sosial Pada Marketplace Shopee Menggunakan Algoritma C4.5," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 9, No. 1, Pp. 1355–1361, 2025.

[13] D. Avianto And A. P. Wibowo, "Pembentukan Pohon Keputusan Untuk Penerima Bantuan

Beras Miskin Menggunakan Algoritma Decision Tree C4 . 5 Decision Tree Formation For Subsidized Rice Aid Recipients Using The C4 . 5 Algorithm," Vol. 9, No. 1, Pp. 59–68, 2024.

[14] F. Husna, H. Rahman, And J. Juhari, "Implementasi Data Mining Menggunakan Algoritma C4.5 Pada Klasifikasi Penjualan Hijab," *J. Ris. Mhs. Mat.*, Vol. 2, No. 2, Pp. 40–46, 2022, Doi: 10.18860/Jrmm.V2i2.14891.

[15] I. Nawawi And Z. Fatah, "Penerapan Decision Trees Dalam Mendeteksi Pola Tidur Sehat Berdasarkan Kebiasaan Gaya Hidup," Vol. 2, No. 4, Pp. 34–41, 2024.

[16] P. E. Putra, M. A. Amrullah, Y. H. Fauzi, And R. F. Saputri, "Penerapan Algoritma C4 . 5 Untuk Klasifikasi Tingkat Korban Banjir Di Indonesia," *J. Komput. Antart.*, Vol. 3, No. C, Pp. 1–7, 2025.

[17] M. D. K. Rizky Ade Safitri, Yoseph Tajul Arifin, "Penerapan Algoritma C4.5 Untuk Klasifikasi Kepuasan Layanan Aplikasi Depok Single Window Rizky," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 8, No. 3, Pp. 4094–4102, 2024.

[18] I. A. Aldiyansyah, Ade Irma Purnamasari, "Perbandingan Tingkat Akurasi Algoritma Decision Tree Dan Random Forest Dalam Mengklasifikasi Penerima Bantuan Sosial Bpnt Di Desa Slangit," *Jati (Jurnal Mhs. Tek. Inform.*, Vol. 8, No. 1, Pp. 127–132, 2024.

[19] A. W. Wicaksono And T. Setiadi, "Penerapan Klasifikasi Decision Tree (C4.5) Untuk Memprediksi Kelulusan Siswa Sekolah Dasar Di Kecamatan Juai," *Format J. Ilm. Tek. Inform.*, Vol. 12, No. 2, P. 151, 2023, Doi: 10.22441/Format.2023.V12.I2.008.