Klasifikasi Penderita Penyakit Diabetes Berdasarkan Decision Tree Menggunakan Algoritma C4.5

Bagas Aulifia Riski Putra Wahyu[1], Achmad Fayi Farozi[2], Caesario Putra Mahendra[3]
Rinci Kembang Hapsari[*4]
[1,2,3,4]Jurusan Teknik Informatika, Fakultas Teknik Elektro dan Teknologi Informasi, Institut Teknologi Adhi Tama Surabaya
Email: [1]bagaswahyu891@gmail.com, [2]achmad.fy12@gmail.com, [3]csrio1130@gmail.com
Email Penulis Korespondensi: [4*]rincikembang@itats.ac.id

Abstract. *Diabetes is a metabolic disease characterized by high blood sugar levels (hyperglycemia) caused by a lack of insulin or the ineffectiveness of insulin in regulating glucose metabolism. In addition there are other factors that cause diabetes such as heredity, weight, age, blood pressure and so on. It is estimated that the death rate caused by diabetes will continue to increase every year. Treatment of diabetes can be done by controlling blood sugar levels, eating a healthy diet, exercising regularly, and if necessary, carrying out early checks to reduce the risk of developing diabetes. Therefore it is necessary to have an early diagnosis which is expected to reduce diabetes and reduce complications of diabetes in the future. One thing that can be done is to apply the method contained in data mining, namely utilizing the classification method using the C4.5 algorithm which can produce more accuracy. Classification can be used as early treatment of this disease. Algorithm C4.5 is an algorithm that is used to form a decision tree. From the test results, it produces a fairly large accuracy, namely 85% Precision of 92%, and Recall of 85%.*
Keywords: *C4.5 Algorithm; Data Mining; Diabetes mellitus; Decision Tree; Classification.*

Abstrak. *Diabetes adalah penyakit metabolik yang ditandai dengan tingginya kadar gula darah (hiperglikemia) yang disebabkan oleh kekurangan insulin atau tidak efektifnya insulin dalam mengatur metabolisme glukosa. Selain itu terdapat faktor-faktor lain menjadi penyebab terjadinya diabetes diantaranya seperti faktor keturunan, berat badan, usia, tekanan darah dan lain sebagainya. Angka kematian yang disebabkan oleh penyakit diabetes diperkirakan akan terus meningkat angka kasus kematiannya pada setiap tahunnya. Penanganan diabetes dapat dilakukan dengan pengontrolan kadar gula darah, diet sehat, olahraga teratur, dan jika perlu lakukan pengecekan dini untuk mengurangi resiko terkena penyakut diabetes. Oleh sebab itu perlu adanya diagnosis sejak dini yang diharapkan dapat menurunkan penyakit diabetes dan merendahkan komplikasi penyakit diabetes di waktu yang akan datang. Salah satu yang bisa dilakukan adalah dengan menerapkan Metode yang yang terdapat pada dalam data mining yaitu memanfaatkan metode klasifikasi menggunakan algoritma C4.5 yang dapat menghasilkan akurasi yang lebih. Klasifikasi bisa digunakan sebagai penanganan dini dari penyakit ini. Algoritma C4.5 merupakan algoritma yang digunakan sebagi pembentuk pohon keputusan (Decision Tree). Dari hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 85 % Precision sebesar 92%, dan Recall sebesar 85%.*
Kata Kunci: *Algoritma C4.5; Data Mining; Diabetes Melitus; Decision Tree; Klasifikasi.*

1. Pendahuluan

Diabetes Melitus adalah salah satu sindrom yang ditandai dengan kelainan pada metabolik yang terganggu serta pada kenaikan konsentrasi gula darah yang abnormal yang disebabkan oleh defisiensi insulin, ataupun sensitivitas insulin yang rendah dari jaringan, maupun keduanya(Kadhm et al. 2018). Diabetes tidak hanya menyebabkan kematian *premature* yang terjadi di seluruh dunia (Widyasari 2017). Penyakit ini juga menjadi penyebab utama terjadinya kebutaan penyakit jantung dan gagal ginjal (Lathifah 2017). Organisasi *International Diabetes Federation* (IDF) memperkirakan sedikitnya terdapat 463 juta orang pada usia 20-79 tahun di dunia menderita diabetes pada tahun 2019 atau setara dengan angka prevalensi sebesar 9,3% dari total penduduk usia yang sama. Berdasarkan jenis kelamin, IDF memperkirakan prevalensi diabetes di tahun 2019 yaitu meningkat 9% pada

perempuan dan 9,65% pada laki-laki. Prevalensi diabetes diperkirakan meningkat seiring penambahan umur penduduk menjadi 19,9% atau 111,2 juta orang pada umur 65-79 tahun. Angka penderita penyakit diabetes ini diprediksi terus meningkat hingga mencapai 578 juta di tahun 2030 dan 700 juta di tahun 2045 (Diabetes Mellitus 2020).

Data mining adalah sebuah metode untuk melakukan akuisisi pengetahuan. Dengan data mining, informasi-informasi implisit dan berharga dari sebuah data dapat diekstrak. Menurut (Oktanisa and Supianto 2018) data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Menurut(Reza Noviansyah et al. 2018), pengertian data mining adalah analisa yang dilakukan secara otomatis pada data besar dan kompleks dengan tujuan untuk mendapatkan pola penting yang keberadaannya biasanya tidak disadari. Berdasakan penjelasan dari beberapa ahli dapat disimpulakan data mining yaitu suatu prosedur menciptakan ikatan dimana memiliki arti, pola, dan kecondongan dengan mengamati kelompok data besar yang berada dalam storage dengan menggunakan teknik identifikasi pola. Data mining mempunyai lima peran utama yaitu estimasi, prediksi, klasifikasi, klaster dan asosiasi. Peran data mining yang sering digunakan adalah klasifikasi dan klaster karena dapat digunakan untuk atribut yang banyak. Klaster merupakan proses pengelompokan data ke dalam beberapa kelompok sehingga data dalam kelompok tersebut memiliki tingkat kemiripan karakteristik antara data yang satu dengan yang lainnya. Sedangkan klasifikasi adalah bentuk analisis data untuk mengekstrak model yang akan digunakan untuk memprediksi label kelas.

Klasifikasi merupakan sebuah proses untuk menciptakan fungsi atau model menjelaskan kelas pada data atau konsep guna untuk memprediksi kelas dari sebuah objek yanglabelnya belum didapatkan. Klasifikasi termasuk dalam tipe supervised learning yang artinya dibutuhkan data pelatihan untuk membangun suatu model klasifikasinya. Terdapat lima kategori klasifikasi yaitu berbasis stastistik, berbasis jarak, berbasis pohon keputusan, berbasis jaringan saraf, dan berbasis aturan. Teknik klasifikasi dapat dimanfaatkan untuk meramal atau memprediksi sebuah permasalahan atau fenomena yang terjadi disekitar, contohnya klasifikasi ini dapat digunakan untuk memprediksi seseorang yang terjangkit penyakit diabetes dan tidak terjangkit. Terdapat beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi, diantaranya algoritma C4.5, algorima K-Means, algoritma Priori, algoritma Naïve Bayes, algoritma Support Vector Machines.

Berdasarkan pemaparan yang diatas diperlukan adanya pendeteksian sejak dini penyakit diabetes. Pendeteksi sejak dini diharapkan dapat menurunkan resiko komplikasi pada pasien diabetes diwaktu mendatang, guna menganalisa pasien terkena penyakit diabetes sejak dini. Salah satu yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan Algoritma C4.5. Algoritma klasifikasi yang digunakan adalah Decission Tree Algoritma C4.5.

Algoritma C4.5 merupakan pengembangan dari algoritma ID3, (Junaedi, Nuswantari, and Yasin 2019) Ross Quinlan adalah sosok yang mengembangkan Algoritma C4.5, Algoritma C4.5 digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target (Sutoyo 2018). Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule*. Algoritma C4.5 merupakan teknik klasifikasi yang banyak digunakan oleh peneliti mudah digunakan dan diinterpretasikan memliki beberapa kelebihan antara lain, mudah dimengerti, mudah diimplementasikan, membutuhkan sedikit waktu, mampu menangani data numerik dan kategorik dan dapat mengolah dataset yang besar dan rumit. Sebuah pohon keputusan atau *decission tree* adalah hasil dari perhitungan algoritma C4.5 (Yunus et al. 2021).

Algoritma C4.5 ini telah dilakukan penelitian sebelumnya untuk memprediksi kategori indeks prestasi mahasiswa dilihat dari hasil dari penelitian ini diperoleh C4.5 memiliki akurasi sebesar 61,54% (Alverina, Chrismanto, and Santosa 2018). Penelitian lain dilakukan klasifikasi terhadap dataset penderita penyakit diabetes dengan menggunakan metode KNN yang berjudul "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes" yang menghasilkan tingkat akurasi tertinggi 39% pada K=3, presisi tertinggi 65% pada K=3 dan K=5, recall tertinggi 36%

pada K=3, dan F-Measure tertinggi pada K=3 (Argina 2020). Merujuk pada penjelasan ini maka penelitian ini akan memanfaatkan algoritma decision tree C4.5 untuk melakukan klasifikasi penyakit diabetes.

2. Tinjauan Pustaka

2.1. *Decisioin Tree*

Pohon adalah struktur data yang terdiri dari simpul dan tepi (edge). Ada tiga jenis simpul pada pohon: simpul akar (root/node), simpul bercabang/internal (branch/internal node), dan simpul daun (leaf/node). Pohon keputusan adalah penyederhanaan teknik klasifikasi untuk jumlah kelas yang terbatas, dengan simpul internal dan simpul akar diberi label sebagai nama atribut, edge diberi label sebagai kemungkinan nilai atribut, dan simpul daun diberi label sebagai kelas yang berbeda (Eska 2016).

Decision tree adalah salah satu teknik pembelajaran mesin (*machine learning*) yang menggunakan aturan klasifikasi struktur sekuensial hierarkis dengan cara mempartisi dataset training secara rekursif (Simanjuntak, Barus, and Anita 2021).

Decision tree adalah struktur flowchart yang berbentuk seperti pohon, dimana setiap node bagian dalam menandakan pengujian suatu atribut, dengan cabang yang dihasilkan menunjukkan hasil pengujian, dan node daun mewakili distribusi kelas (Zulma and Chamidah 2021)

2.2. *Confussion Matrix*

Confussion Matrix digunakan untuk menghitung nilai akurasi. Pengukuran kinerja menggunakan confusion matrix memiliki empat istilah sebagai gambaran dari hasil klasifikasi (Hadianto, Novitasari, and Rahmawati 2019). Adapun keempat istilah tersebut yaitu :

1. False Positive (FP), yaitu data negatif tapi terprediksi sebagai data positif.
2. False Negative (FN), yaitu data positif yang terprediksi sebagai data negatif.
3. True Positive (TP), yaitu data positif yang terprediksi benar.
4. True Negative (TN), yaitu data negatif yang terprediksi dengan benar.

Untuk menghitung akurasi digunakan rumus sebagai berikut:

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

Sensitivitas atau Recall adalah rasio prediksi benar positif dipadukan dengan keseluruhan data yang benar positif atau mengukur proporsi positif asli yang diramal secara benar sebagai positif(Hapsari et al. 2020). Dalam sensitivitas berkaitan dengan kecakapan pengujian untuk mengenali hasil yang positif dari sejumlah data yang seharusnya positif. Untuk menghitung sensitivitas atau recall menggunakan persamaan dibawah ini:

$$Sensitifitas = \frac{TP}{TP + FN} \quad (4)$$

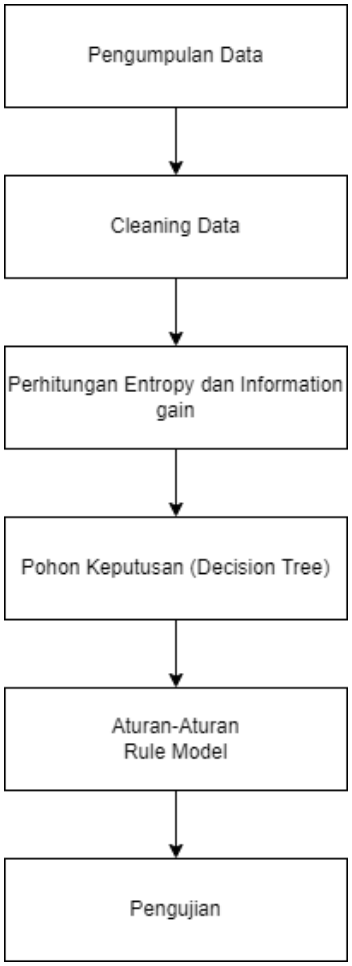
Sedangkan precision adalah rasio ramalan benar positif dipadukan dengan semua hasil yang diprediksi positif. Precision menggambarkan matrik untuk menghitung kemampuan sistem dalam menghasilkan data yang penting. Precison pada data mining adalah hasil jumlah data yang true positive dibagi dengan jumlah data yang dikenali sebagai positif. Untuk menghitung precison menggunakan persamaan dibawah ini:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

3.Metodologi Penelitian

Data Mining merupakan proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk meringkas informasi menjadi bermanfaat (Pramadani, Sunandar, and ... 2019). Data Mining didefinisikan sebagai proses penemuan pola pada data (Han, Kamber, and Pei 2012). Berdasarkan fungsinya, data mining dapat dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi (Ginantra et al. 2021). Penggunaan algoritma data mining dilakukan untuk menggali data yang agar memudahkan identifikasi informasi(Baharuddin, Azis, and Hasanuddin 2019). Namun semakin besar data yang diolah maka semakin besar juga waktu pemrosesannya.

Algotirma C4.5 merupakan algoritma yang umum digunakan untuk pengambilan keputusan. C4.5akan mencari solusi permasalahan dengan menjadikan kriteria sebagai node yang saling berhubungan membentuk seperti struktur pohon(Khotimah and Istiawan 2018). Model prediksi algorima C4.5 mengacu pada suatu keputusan menggunakan struktur hirarki atau pohon. Setiap pohon memiliki cabang, setiap cabang mewakili suatu atribut yang harus dipenuhi untuk menuju cabang selanjutnya hingga berakhir. Konsep data dalam algoritma C4.5 adalah data dinyatakan dalam bentuk tabel yang terdiri dari atribut dan record. Atribut digunakan sebagai parameter yang dibuat sebagai kriteria dalam pembuatan pohon, dan record sebagai penentun nilai keputusan pada pohon (Iqbal, Usino, and Triono 2020).



Gambar 1. Tahap Penelitian

Seluruh tahapan yang dilakukan dalam penelitian ini digambarkan dalam diagram alur yang dapat dilihat pada Gambar 1. Tahap Penelitian. Dimana peroses penelitian ini diawali dengan pengumpulan data, *cleaning data*, perhitungan entropy dan information gain, pohon kepustusan (*Decision tree*), aturan-aturan *rule* model dan diakhiri dengan validasi dan pengujian.

3.1. Pengumpulan Data

Penelitian ini menggunakan data Indian Liver Patient Dataset yang didapat dari website "https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset". Dataset ini berisi data yang dikumpulkan dari para pasien yang ada di timur laut Andhra Pradesh, India.

Dataset yang digunakan terdiri dari 768 data dengan 9 atribut. Dalam dataset terbagi menjadi dua kelompok, yaitu pasein diabetes terdiri dari 268 pasien dan non diabetes terdiri dari 500 pasien.

3.2. *Cleaning Data*

Dilakukan praproses data seperti membersihkan data, menghapus data dan atribut yang tidak relevan, menghilangkan outlier, dan sebagainya. Berdasarkan dataset diabetes yang digunakan terdapat beberapa data yang memiliki nilai (*value*) tidak lengkap. Sehingga dalam penelitian ini dilakukan *cleaning data,* dengan menghapus data yang dengan nilai (*value*) yang tidak lengkap.

3.3. Perhitungan *Entropy* dan *Information Gain*

Pada tahap ini dilakukan perhitungan menggunakan algoritma C4.5. Dimana akan dilakukan penghitung *entropy*(S) sebagai parameter yang berfungsi untuk informasi keberagaman setiap nilai atribut. Kategori atau kriteria terhadap atribut keputusan dalam sebuah dataset. untuk menentukan entropy dengan rumus

$$Entropy(S) = \sum_{i=0}^{n} -p_i \times \log_2 p_i \quad (1)$$

Keterangan:

S menyatakan himpunan kasus
n menyatakan jumlah partisi atribut A
|Si| menyatakan probabilitas yang didapat dari jumlah (ya/tidak) dibagi total kasus.

(a) Hitung nilai *gain*(S,A) adalah untuk mengukur efektivitas masing masing atribut pada *node* tertentu untuk mengklasifikasikan data, nilai terbesar akan menjadi akar pohon utama dengan rumus :

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

Dengan,

S menyatakan himpunan kasus
A menyatakan Atribut
n menyatakan jumlah partisi atribut A
|Si| menyatakan jumlah kasus pada Partisi ke-i
$|S|$ menyatakan jumlah kasus dalam

(b) Mengulangi langkah kedua yaitu menentukan *entropy* sampai semua *record* data terpartisi.
(c) Partisi akan berhenti apabila :
   i. Semua *record* yang ada pada simpul N mendapat kelas yang sama.
   ii. Tidak adanya atribut pada *record* yang dipartisi.
   iii. Tidak adanya *record* pada cabang yang kosong

3.4. Pohon Keputusan (*Decision Tree)*

Hasil dari proses perhitungan untuk mencari nilai entropy untuk masing-masing atribut. Nilai entropy tersebut akan proses untuk mendapat nilai gain yang nantinya setelah semua proses prosedur yang dilakukan, hasil akhir dari proses tersebut adalah pohon kepustusan (*decision tree*).

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

3.5. Pengujian

Pada tahapan pengujian untuk menentukan tingkat akurasi dan pohon keputusan algoritma C4.5. Dalam pengujian penelitian ini dilakukan dengan cara perhitungan secara manual untuk menghasilkan pohon keputusan. Dalam tahap pengujian dilakukan dengan split datasetnya adalah 75% dijadikan sebagai data learning dan 25% dijadikan sebagai data testing.

4. Hasil dan Pembahasan

Dataset yang digunakan ini berisi data yang dikumpulkan dari para pasien diabetes dan non diabetes yang ada di timur laut Andhra Pradesh, India. Dataset ini memiliki 9 atribut, 768 data, dan memiliki 6,921 *instance* atau isi data. *Instance* pada dataset ini memiliki nilai missing pada beberapa atribut sehingga tidak semua atribut dapat diproses secara langsung.

Tabel 1. Data Penelitian

No.	Glukosa	BloodPressure	BMI	Age	Outcome
1	148	72	33.6	50	1
2	85	66	26.6	31	0
3	183	64	23.3	32	1
4	89	66	28.1	21	0
5	137	40	43.1	33	1
6	116	74	25.6	30	0
7	78	50	31	26	1
8	197	70	30.5	53	1
9	110	92	37.6	30	0
10	168	74	38	34	1
11	139	80	27.1	57	0
12	189	60	30.1	59	1
13	166	72	25.8	51	1
14	118	84	45.8	31	1
15	107	74	29.6	31	1
16	103	30	43.3	33	0
17	115	70	34.6	32	1
18	126	88	39.3	27	0
19	99	84	35.4	50	0
20	196	90	39.8	41	1

Dilakukan pemilihan atribut untuk memfokuskan data dan atribut-atribut yang berhubungan dengan diagnosa penyakit diabetes sehingga atribut yang digunakan diantaranya adalah Glokosa, BloodPressure, Body Massa Index (BMI), dan Age. Setelah dilakukan pemilihan fitur dan *cleaning data.* Dataset yang digunakan dlam penelitian terdiri dari 768 data, dengan 249 data pasien diabetes dan 475 pasien non-diabetes. Sample data dan atribut yang digunakan ini ditunjukkan pada Tabel 1. Pada kolom *outcome* bernilai 0 dan 1, dimana 1 berarti pasien diabetes dan 0 adalah pasien non diabetes.

Tabel 2. Acuan Perubahan Data

Atribut	Nilai	kategori
Glukosa	<140	Baik
	141 – 199	Sedang
	>200	Buruk
BloodPressure	<80	Normal
	81 – 89	Prahipertensi
	90 – 99	Hipertensi 1
	100 – 119	Hipertensi 2
	>120	Krisis
BMI	<18,5	Kurang
	18,6 - 29,9	Normal
	>30	Obesitas
Umur	21 – 59	Dewasa
	>60	Lansia

Hasil analisis data yang diperoleh setelah melewati beberapa tahapan pada prosedur percobaan, maka untuk mendapat hasil perhitungan menggunakan algoritma decision tree C4.5 dari data perlu melakukan

transformasi data dengan tujuan untuk pemilihan atribut untuk memfokuskan data dan atribut-atribut yang digunakan hanya yang berhubungan dengan diabetes seperti yang terdapat pada Tabel 1.

Data Penelitan, atribut-atribut yang digunakan diataranya adalah Glokosa, BloodPressure, Body Massa Index (BMI), dan Age. Pengubahan data ini dilakukan berdasarkan acuan diagnosis dari beberapa dokter yaitu dr. Meva Nareza dr. Tjin Willy, dan dr. Karlina Lestari. Acuan ini ditunjukan pada Tabel 2. Acuan Perubahan Data.

Glukosa sesorang akan dikategorikan baik jika nilainya kurang dari 140, dikategori sedang jika nilainya diantara 141-199 dan dikategorikan buruk jika nilainya dilebih dari 200. Blood Prerssure sesorang akan dikategorikan normal jika nilainya kurang dari 140, dikategori prahipertensi jika nilainya diantara 81-89, dikategorikan hipertensi 1 jika nilainya diantara 90-99, dikategorikan hipertensi 2 jika nilainya diantara 100-119, dan dikategorikan krisis jika nilainya dilebih dari 120. *Body Massa Index* (BMI) sesorang akan dikategorikan baik jika nilainya kurang dari 18,5, dikategori sedang jika nilainya diantara 18,6-29,9 dan dikategorikan buruk jika nilainya dilebih dari 30. Umur sesorang akan dikategorikan dewasa jika nilainya diantara dari 21-59 dan dikategorikan lansia jika nilainya dilebih dari 60. Hasil dari perubahan data numerik menjadi data kategori dari 768 data diperoleh data seleksi yang dapat dilihat pada Tabel 3.

Tabel 3. Pengubahan Data Menjadi Kategori

No.	Glukosa	BloodPressure	BMI	Age
1	Sedang	Normal	Obesitas	Dewasa
2	Baik	Normal	Normal	Dewasa
3	Sedang	Normal	Normal	Dewasa
4	Baik	Normal	Normal	Dewasa
5	Baik	Normal	Obesitas	Dewasa
6	Baik	Normal	Normal	Dewasa
7	Baik	Normal	Obesitas	Dewasa
8	Baik	Normal	Obesitas	Dewasa
9	Sedang	Normal	Obesitas	Dewasa
10	Baik	Hipertensi1	Kurang	Dewasa
11	Baik	Hipertensi1	Obesitas	Dewasa
12	Sedang	Normal	Obesitas	Dewasa
13	Baik	Normal	Normal	Dewasa
14	Sedang	Normal	Obesitas	Dewasa
15	Sedang	Normal	Normal	Dewasa
16	Baik	Normal	Obesitas	Dewasa
17	Baik	Prahipertensi	Obesitas	Dewasa
18	Baik	Normal	Normal	Dewasa
19	Baik	Normal	Obesitas	Dewasa
20	Baik	Normal	Obesitas	Dewasa

Hasil dari perubahan data dilakukan proses perhitungan untuk mencari nilai entropy untuk masing-masing atribut. Dari nilai entropy yang diperoleh akan proses untuk mendapat nilai gain yang nantinya setelah semua proses prosedur yang dilakukan, hasil akhir dari proses tersebut adalah pohon kepustusan (*decision tree*) pada Gambar 2.

Berdasarkan Gambar 2 Pohon Keputusan menghasilkan rule sebagai berikut:

a) Jika Glukosa adalah Sedang maka label = Diabetes
b) Jika Glukosa adalah Buruk maka label = Tidak Diabetes
c) Jika Glukosa adalah Baik, dan BMI Kurang maka label = Tidak Diabetes
d) Jika Glukosa adalah Baik, BMI Normal dan BloodPressure Normal maka label = Tidak Diabetes
e) Jika Glukosa adalah Baik, BMI Normal dan BloodPressure Prahipertensi maka label = Tidak Diabetes
f) Jika Glukosa adalah Baik, BMI Normal dan BloodPressure Hipertensi1 maka label = Tidak Diabetes
g) Jika Glukosa adalah Baik, BMI Normal dan BloodPressure Hipertensi2 maka label = Tidak Diabetes
h) Jika Glukosa adalah Baik, BMI Normal dan BloodPressure Krisis maka label = Tidak Diabetesan

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

Gambar 2. Pohon Keputusan

Dari pohon keputusan yang diperoleh, nilai prediksi dapat ditentukan dari mencocokan rule dengan dataset yang dimana hasil akhir akan membandingkan nilai outcome dan nilai prediksi sehingga didapatkanya nilai Confussion Matrix. Berdasarkan pencocokan nilai outcome pada nilai dataset yang memiliki nilai sama dengan hasil prediksi, maka diperoleh true positive jumlah prediksi positif yang benar terdapat 11 data, true negative jumlah prediksi negatif yang benar sebanyak 6 data, false positive jumlah prediksi positif yang salah sebanyak 1 data dan false negative yang salah sebanyak 2 data.

Tabel 4. ***Confusion Matrix***

OUTCOME	PREDIKSI	
	Diabetes	Tidak Diabetes
Diabetes	11	1
Tidak Diabetes	2	6

Berdasarkan hasil diperoleh pada Tabel 4 Confussion Matrix, Confussion Matrix ini dilakukan proses perhitungan untuk mengetahui nilai akurasi, nilai precision, dan nilai recoll yang hasilnya dapat dilihat pada Tabel 5 Performace Vector dengan akurasi 85%, precision 92%, dan recall 85%. Nilai-nilai pada ini yang akan menjadi indikator pengukuran kinerja algoritma decision tree C4.5. Tingginya nilai akurasi ini dapat dilakukan dengan penambahan atribut baru atau menghapus atribut yang tidak penting, Sehingga adanya peluang untuk mendapatkan nilai akurasi yang lebih tinggi. Pada penelitian klasifikasi diagnosis penyakit dibetes melitus menggunakan metode Naïve Bayes yang memiliki akurasi tertinggi sebesar 80% dan untuk nilai precision yaitu 0.86 [16]. Berdasarkan pada penjelansan diatas dapat disimpulkan metode decision tree C4.5 lebih baik dibandingkan algoritma Naïve Bayes untuk klasifikasi data dalam data mining.

Tabel 5. ***Performace Vector***

Variabel	Rumus	Hasil (%)
Akurasi	(TP + TN / Jumlah Data)*100	85
Precision	(TP / (TP + FP))*100	91,68
Recall	(TP / (TP + FN))*100	84,61

5. Kesimpulan

Hasil klasifikasi dari penderita penyakit diabetes dengan atribut yang terdapat dalam dataset diabetes yaitu Glukosa, BMI, BloodPressure, Umur dapat dijadikan sebagai data untuk klasifikasi penderita penyakit diabetes. Penelitian ini menggunakan Algoritma C4.5 untuk pengkalsifikasian seseorang terkena penyakit diabetes atau tidak. Dari total data 768 dengan tahapan pengumpulan data, pengolahan data dan pengujian dengan diterapkannya algoritma C4.5. Untuk perhitungan manual terdapat beberapa tahapan pada algoritma C4.5 yakni mencari nilai entropy, kemudian setelah mencari nilai entropy selanjutnya mencari nilai gain, setelah nilai gain di dapatkan mencari nilai gain tertinggi untuk dijadikan node akar. Dilakukan perhitungan berulangkali sampai hasilnya telah memiliki keputusan semuanya. Perhitungan menggunakan algoritma C4.5 membentuk decision tree yang memiliki 8 rule diharapkan menjadi suatu informasi tentang penyakit diabetes. Evaluasi dari penelitian ini diukur dengan akurasi, precision, dan recall didapatkan akurasi 85%, precision 92%, dan recall 85%. Kurangnya nilai akurasi akurasi decision tree hal ini dapat dilakukan penambahan atribut baru atau menghapus atribut yang tidak penting. Saran yang dapat diberikan dari hasil penelilitian ini adalah untuk menggunakan dataset dari rumah sakit yang berada di sekitar Indonesia. Semakin banyak data akan semakin akurat, dan menggunakan algoritma yang berbeda dengan tujuan untuk mengetahui perbandingannya.

Referensi

Alverina, Dea, Antonius Rachmat Chrismanto, and R. Gunawan Santosa. 2018. "Perbandingan Algoritma C4.5 Dan CART Dalam Memprediksi Kategori Indeks Prestasi Mahasiswa." *Jurnal Teknologi dan Sistem Komputer* 6(2): 76–83.

Argina, Andi Maulida. 2020. "Penerapan Metode Klasifikasi K-Nearest Neigbor Pada Dataset Penderita Penyakit Diabetes." *Indonesian Journal of Data and Science* 1(2): 29–33.

Baharuddin, Mus Mulyadi, Huzain Azis, and Tasrif Hasanuddin. 2019. "Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca." *ILKOM Jurnal Ilmiah* 11(3): 269–74.

"Diabetes Mellitus." 2020. *Pusat Data Dan Informasi Kementrian Kesehatan Republik Indonesia*. https://pusdatin.kemkes.go.id/article/view/20111800001/diabetes-melitus.html.

Eska, Juna. 2016. "Penerapan Data Mining Untuk Prekdiksi Penjualan Wallpaper Menggunakan Algoritma C4.5 STMIK Royal Ksiaran." *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)* 2: 9–13.

Ginantra, Ni Luh Wiwik Sri Rahayu et al. 2021. *Data-Mining-Algoritma-Dan-Implementasi-1638852186*.

Hadianto, Nur, Hafifah Bella Novitasari, and Ami Rahmawati. 2019. "Klasifikasi Peminjaman Nasabah Bank Menggunakan Metode Neural Network." *Jurnal Pilar Nusa Mandiri* 15(2): 163–70.

Han, Jiwei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Data Mining Concepts and Techniques*. Thirrd. Morgan Kauffman Publishers.

Hapsari, R K, M I Utoyo, R Rulaningtyas, and H Suprajitno. 2020. "Iris Segmentation Using Hough Transform Method and Fuzzy C-Means Method." *Journal of Physics: Conference Series* 1477: 022037. https://iopscience.iop.org/article/10.1088/1742-6596/1477/2/022037.

Iqbal, Muchamad, Wendi Usino, and Triono Triono. 2020. "Sistem Pendukung Keputusan Menentukan Hasil Budidaya Udang Vaname Dengan Metode Algoritma C4.5 (Pt Anugerah Sumber Laut Jaya)." *Jurnal Tekno Insentif* 14(1): 28–39.

Junaedi, Ifan, Ndaru Nuswantari, and Verdi Yasin. 2019. "Perancangan Dan Implementasi Algoritma C4 . 5 Untuk Data Mining." *Journal of Information System, Informatics and Computing* 3(1): 29–44. http://journal.stmikjayakarta.ac.id/index.php/jisicom/article/view/203%0Ahttp://journal.stmikjayakarta.ac.id/index.php/jisicom/article/download/203/158.

Kadhm, Mustafa S, Doaa N Mhawi, Ikhlas Watan Ghindawi, and Duaa Enteesha Mhawi. 2018. "An Accurate Diabetes Prediction System Based on K-Means Clustering and Proposed Classification Approach." *International Journal of Applied Engineering Research* 13(6): 4038–41. http://www.ripublication.com.

Khotimah, Nur, and Deden Istiawan. 2018. "Perbandingan Algoritma C4.5, Naïve Bayes Dan K-Nearest Neighbour Untuk Prediksi Lahan Kritis Di Kabupaten Pemalang." *Urecol* 7(1): 41–50.

Lathifah, Nur Lailatul. 2017. "Hubungan Durasi Penyakit Dan Kadar Gula Darah Dengan Keluhan Subyektif Penderita Diabetes Melitus." *Jurnal Berkala Epidemiologi* 5(2): 231–39. https://e-journal.unair.ac.id/JBE/article/view/4781.

Oktanisa, Irvi, and Ahmad Afif Supianto. 2018. "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing." *Jurnal Teknologi Informasi dan Ilmu Komputer* 5(5): 567.

Pramadani, E, H Sunandar, and ... 2019. "Implementasi Data Mining Penjualan Koran Dengan Metode C4. 5 (Studi Kasus: Pt. Media Massa Cahaya Pembaharuan)." *Informasi dan ...* 6: 11–15. https://www.ejurnal.stmik-budidarma.ac.id/index.php/inti/article/view/1012%0Ahttps://www.ejurnal.stmik-budidarma.ac.id/index.php/inti/article/download/1012/871.

Reza Noviansyah, M et al. 2018. "Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya)." *Jurnal Coding, Sistem Komputer Untan* 06(2): 48–56.

Simanjuntak, Krisna Ferdinan Leo, Annita Carolina Br Barus, and Anita. 2021. "Implementasi Metode Decision Tree Dan Algoritma C4.5 Untuk Klasifikasi Kepribadian Masyarakat." *JOISIE Journal Of Information System And Informatics Engineering* 5(1): 51–59.

Sutoyo, Imam. 2018. "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik." *Jurnal Pilar Nusa Mandiri* 14(2): 217.

Widyasari, Nina. 2017. "HUBUNGAN KARAKTERISTIK RESPONDEN DENGAN RISIKO DIABETES MELITUS DAN DISLIPIDEMIA KELURAHAN TANAH KALIKEDINDING." *Jurnal Berkala Epidemiologi* 5(April 2017): 130–41.

Yunus, Muhamad, Hanandriya Ramadhan, Dimas Rizki Aji, and Agus Yulianto. 2021. "Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP)." *Paradigma - Jurnal Komputer dan Informatika* 23(2).

Zulma, G D M, and N Chamidah. 2021. "Perbandingan Metode Klasifikasi Naive Bayes, Decision Tree Dan K-Nearest Neighbor Pada Data Log Firewall." *Senamika* (April): 679–88. https://conference.upnvj.ac.id/index.php/senamika/article/view/1396.