Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

Pemodelan Klasifikasi Untuk Menentukan Penyakit Diabetes dengan Faktor Penyebab Menggunakan Decision Tree C4.5 Pada Wanita

Nining Nur Habibah*, Alwis Nazir, Iwan Iskandar, Fadhilah Syafria, Lola Oktavia, Ihda Syurfi

Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1,*]11950121722@students.uiun-suska.ac.id, [2]alwis.nazir@uin-suska.ac.id, [3]iwan.iskandar@uin-suska.ac.id,
[4]fadhilah.syafria@uin-suska.ac.id, [5]lola.oktavia@uiun-suska.ac.id, [6]ihdasyurfi11@gmail.com
Email Penulis Korespodensi: 11950121722@students.uin-suska.ac.id
Submitted: **25/05/2023**; Accepted: **30/06/2023**; Published: **30/06/2023**

Abstrak−Diabetes erat kaitannya pankreas, dimana pankreas memproduksi hormon alamiah insulin, akan tetapi fungsinya bermasalah yang menyebabkan naiknya tensi gula dalam darah di tubuh. Naiknya tensi darah dapat membuat fungsi organ didalam merusak fungsi organ ditubuh seseorang seperti ginjal, jantung, dan otak. Dimana membuat seseorang mempunyai riawayat penyakit diabetes. Diabetes yang menyerang orang dewasa dapat dicegah melalui olahraga dan pola makan yang teratur dan sehat . Menurut Organisasi International Diabetes Federation (IDF) memperkirakan sedikitnya terdapat 19,5 jutaan penduduk negara Indonesia antara usia 20 hingga 79 tahun menderita diabetes tahun 2021. China berada di posisi pertama penderita diabetes sebanyak 140,9 juta jiwa. India di urutan selanjutnya dengan jumlah pengidap diabetes sebesar 74,2 juta jiwa. Oleh karena itu, diagnosis dini sangat penting karena bertujuan untuk mengurangi diabetes dan komplikasi diabetes dikemudian hari. Perlunya dilakukan pendataan pasien penderita diabetes yang diharapkan mampu untuk dilakukannya pencegahan. Oleh sebab itu menerapkan teknik klasifikasi dengan data mining dengan algoritma C4.5. Dimana klasifikasi tersebut dapat mencapai ketelitian yang unggul. Algoritma C4.5 umumnya digunakan dalam menentukan node dari pohon keputusan. Berdasarkan hasil pengujian, akurasinya 76,67 persen, presisi 72 persen, dan recall 41,67 persen.

Kata Kunci: Algoritma Decision Tree; Diabetes; Data Mining ; Klasifikasi; Pohon Keputusan;

Abstract−Diabetes is closely related to the pancreas, where the pancreas produces the natural hormone insulin, but its function is problematic which causes an increase in blood sugar levels in the body. Rising blood pressure can affect organ function in damaging the function of organs in a person's body such as the kidneys, heart and brain. Where makes a person have a history of diabetes. Diabetes that attacks adults can be prevented through exercise and a regular and healthy diet. According to the International Diabetes Federation (IDF) organization, it is estimated that at least 19.5 million Indonesian people between the ages of 20 and 79 will suffer from diabetes in 2021. China is in first place with diabetes with 140.9 million people. India is next in line with the number of people with diabetes of 74.2 million people. Therefore, early diagnosis is very important because it aims to reduce diabetes and diabetes complications in the future. It is necessary to collect data on patients with diabetes who are expected to be able to do prevention. Therefore applying classification techniques with data mining with the C4.5 algorithm. Where the classification can achieve better accuracy. Algorithm C4.5 is generally used in determining the nodes of a decision tree. Based on the test results, the accuracy is 76.67 percent, the precision is 72 percent, and the recall is 41.67 percent.

Keywords: Decision Tree Algorithm; Diabetes; Decision Tree; Classification; Data Mining

1. PENDAHULUAN

Penyakit organ dalam yang termasuk kronis adalah penyakit diabetes. Penyakit naiknya kadar gula darah atau diabetes berhubungan erat dengan pankreas, dimana mempunyai fungsi untuk insulin, namun fungsinya membuat hormon alamiah dari insulin. Naiknya diabetes dalam tubuh manusia melemahkan kerja dari organ dalam tubuh. Yang menyebabkan penderitanya mengalami diabetes. Pencegahan yang dapat dilakukan adalah melalui olahraga dan pola makan yang teratur dan sehat [1].

International Diabetes Federation (IDF) memperkirakan setidaknya 19,5 jutaan orang di negara Indonesia yang usia antara 20 dan 79 tahun akan menderita penyakit ini pada tahun 2021. China berada di posisi pertama penderita diabetes sebanyak 140,9 juta jiwa. India di urutan selanjutnya dengan jumlah pengidap diabetes sebesar 74,2 juta. Tujuan dilakukannya penelitian adalah untuk mengerti apakah seseorang menderita penyakit diabetes dengan meneliti beberapa faktor penyebab penyakit diabetes. Beberapa faktor penyebab tersebut adalah pregnancies, glucose, insulin dan lain sebagainya. Diagnosis dini sangat penting karena bertujuan untuk mengurangi diabetes dan komplikasi diabetes dikemudian hari. Perlunya dilakukan pendataan pasien penderita diabetes yang diharapkan mampu untuk dilakukannya pencegahan. Dataset penelitian ini berasal dari kaggle yang isi datanya diambil dari Rs Pima India yang berisi data penyakit diabetes pada wanita [1].

Knowledge Discovery in Database (KDD) merupakan nama sering dipakai untuk menyebut data mining, dimana proses yang dipakai untuk mengumpulkan, memakai data dalam keterkaitan pola ukuran yang mencakup luas. Teknik ini dipakai dalam proses mengambil penentu kedepannya dengan fakta keputusan sebelumnya yang telah lalu. Metode data mining biasa digunakan untuk menentukan pola data yang luas. Salah satu teknik analisis data mining yaitu klasifikasi [2].

Selama klasifikasi, objek data dievaluasi untuk menetapkannya ke kategori tertentu dari jumlah kategori yang tersedia. Pengklasifikasi membuat model berdasarkan data pelatihan yang ada dan kemudian menggunakan model tersebut untuk mengklasifikasikan data baru dengan melatih/mempelajari fungsi tujuan yang meletakan tiap

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

atribut (fitur) ke sekumpulan kelas yang sesuai. Salah satu algoritma klasifikasi yang dipakai dalam penelitian ini adalah pohon keputusan [3]

Decision Tree adalah algoritme yang termasuk dalam pembelajaran supervised machine atau salah satu pendekatan dalam pembelajaran mesin untuk mengembangkan model yang dapat mempelajari pola atau hubungan dalam data input-output yang ada, sehingga dapat melakukan prediksi atau klasifikasi yang akurat pada data baru yang belum dilihat sebelumnya untuk memecahkan masalah klasifikasi yang bertujuan untuk mengimplementasikan dengan mudah aturan model prediksi secara spesifik. Decision tree mempunyai root node serta internode dalam melakukan proses prediksi dan klasifikasi. Dengan penelitian yang menggunakan metode Decision Tree ini untuk mengklasifikasikan penyakit diabetes, memungkinkan untuk mengetahui apakah faktor penyebab diabetes yang dialami seseorang seperti faktor banyaknya mengalami kehamilan, tekanan darah, kadar massa tubuh, usia, dan lain sebagainya merupakan indikator yang membuat seseorang mengalami diabetes. Dengan mengetahui apa saja faktor penyebab penyakit diabetes diharapkan bisa menurunkan faktor penyakit diabetes dimasa yang akan datang dan dapat dilakukannya penanganan yang sesuai. Hasil dari penggolongan penyakit diabetes tersebut kemudian diproses melalui data mining [5].

Algoritma C4.5 merupakan cara yang merekomendasikan yang dipakai untuk mengembangkan data secara segmentasi sesuai dengan kekuatan prediksinya. Biasa untuk membangun pohon keputusan dari dataset pelatihan diberikan. “Keuntungan dari algoritma ini adalah memungkinkan menggunakan lebih sedikit standar secara signifikan.” [4].

Di metode klasifikasi, Algoritma Decision Tree banyak digunakan oleh peneliti untuk melakukan pengelompokan penyakit diabetes. Salah satunya yang dilakukan oleh Bagas Aulifia Riski Putra Wahyu et al dalam penelitian yang membahas tentang diagnosa dini terhadap penderita diabetes untuk mengurangi adanya resiko penyakit diabetes [5]. Selanjutnya penelitian yang dilakukan oleh Noviandi dalam memprediksi penyakit diabetes untuk mendeteksi penyakit diabetes. Di penelitian ini memakai algoritma C4.5 dalam membuat prediksi diagnostik tentang apakah pasien menderita diabetes, dengan mempertimbangkan beberapa faktor lainnya [6]

Penelitian yang dilakukan oleh Sanni Ucha Putri et al dalam judul penelitian penerapan data mining untuk memprediksi penyakit diabetes melalui algoritma C4.5. Tujuannya untuk membangun model prediktif memakai algoritma data mining C4.5 yaitu membangun pohon keputusan dan pengujian dengan rapidminner supaya dapat menangkal penyakit yang dilaksanakan secara cepat. Penelitian ini memiliki berbagai ciri yakni beban massa tubuh, umur, tensi darah, detak jantung dan kadar glukosa darah. Hasil penelitian ini menjadi patokan supaya mengetahui, berdasarkan karakteristik yang diberikan, apakah seseorang berisiko terkena diabetes atau tidak [4]

Penelitian selanjutnya yang dilakukan oleh I Made Agus Oka Gunawan et al dalam penelitian yang berjudul klasifikasi penyakit jantung menggunakan algoritma decision tree series C4.5 dengan rapidminer. Pendalamannya menggunakan c4.5 sebagai mengklasifikasikan data penyakit jantung. Algoritma C4.5 dari rangkaian pohon keputusan diproses menggunakan alat Rapidminer versi 9.10. Langkah-langkah Preprocessing, Set Roles memodelkan algoritma set pohon keputusan C4.5 dari data pelatihan, menerapkan model ke pengujian data, dan mengujinya untuk menghitung keakuratan model selama pengujian data [7].

Selanjutnya penelitian yang dilakukan oleh Musa Yusa et al dalam judul penelitian evaluasi kinerja algoritme klasifikasi pohon keputusan ID3, C4.5, dan CART menggunakan kumpulan data penerimaan kembali diabetes. Pada penelitian ini, model algoritma klasifikasi mempunyai perbedaan nilai yang bergantung pada record data. Dataset yang dipakai yaitu kumpulan informasi terkait proses penerimaan kembali diabetes. Karena data yang digunakan masih kurang nilainya, maka pada penelitian ini dilakukan langkah preprocessing data. Setelah tahap pra-pemrosesan data, diperoleh kumpulan data berjumlah 47 atribut dan 49.735 data. Studi ini juga menguji kinerja teknik klasifikasi menggunakan algoritma pohon keputusan yang berbeda pada kumpulan data. Algoritma klasifikasi yang dipakai yaitu ID3, C4.5 dan CART. Metode pencarian yang dipakai yaitu 10-fold cross-validation. Dimana penelitian ini mendapatkan bahwa model klasifikasi C4.5 mempunyai kinerja terbaik. Nilai daya yang dihasilkan merupakan daya presisi sebesar 54,13 waktu eksekusi 6 detik [8].

Pada penelitian sebelumnya sama sama menggunakan metode c4.5 dengan 7 atribut. Sedangkan pada penelitian ini, dikembangkan suatu sistem klasifikasi dengan menggunakan 9 atribut. Penelitian ini menggunakan metode pohon keputusan untuk mengembangkan sistem klasifikasi yang mengklasifikasikan apakah seseorang menderita diabetes berdasarkan beberapa faktor pengalaman lainnya. Dan diharapkan hasil penelitian ini bermanfaat untuk peramalan seseorang menderita penyakit diabetes.

2. METODOLOGI PENELITIAN

2.1 Data Mining

Sering disebut Knowledge Discovery in Database (KDD), data mining melibatkan pengumpulan dan penggunaan data historis untuk menemukan hubungan, pola, atau hubungan dalam kumpulan data yang besar. Penambangan informasi umumnya dipakai sebagai meningkatkan ketetapan waktu masa depan melalui data yang dikumpulkan di masa lalu. Penambangan data adalah disiplin yang mempelajari metode untuk mengekstraksi informasi atau menemukan pola dalam data dalam jumlah besar. Salah satu teknik analisis data mining adalah klasifikasi [2],[3],[9].

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

2.2 Klasifikasi

Klasifikasi adalah kegiatan yang melibatkan temuan sekumpulan pola (fungsi) yang mendeskripsikan serta dapat dibedakan rancangan dari data, sehingga pola bisa dipakai dalam prediksikan yang kelasnya tidak dimengerti. Salah satu algoritma klasifikasi yang digunakan dalam penelitian ini adalah pohon keputusan [10],[11],[12].

2.3 Decision Tree

Decision tree yang sering dikenal Pohon keputusan adalah cara dalam pengklasifikasian serta prediksi yang mumpuni dalam keakuratannya. Teknik ini mengganti bukti yang luar dan menmbuat pohon keputusan sesuai aturan. Aturannya mudah dipahami dalam bahasa alami. Mereka juga dapat diekspresikan dalam format basis data seperti Structure Query Language (SQL) yang mencari notasi bahan lain. Pohon keputusan merupakan wujud yang dipakai dalam membagikan himpunan data besar menjadi himpunan terkecil melalui penerapan berbagai ketentuan yang sesuai. Dimana turut untuk rekognisi lapisan. Nodul yang tidak termasuk dalam pangkalan node yang berisi root dan internal node dari kondisi uji atribut sejumlah record dengan properti yang berbeda. Simpul akar dan simpul dalam ditandai dengan oval dan simpul daun ditandai dengan persegi panjang[13],[14].

2.4 Algortima C4.5

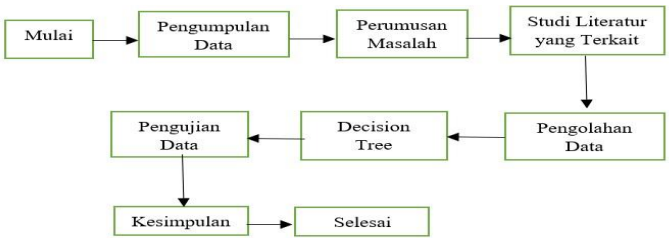
Algoritma C4.5 adalah algoritma yang sangat populer. Banyak peneliti di dunia menjelaskan hal ini oleh Xindong Wu dan Vipin Kumar dalam bukunya The Top Ten Algorithms in Data Mining. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh J. Rose Quinlan. Secara umum, algoritma C4.5 harus digunakan. Pohon keputusan terlihat seperti ini [15], [16],[17],[18],[19],[20]. Rumusnya adalah

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

A Pilih atribut sebagai pengguna root
B. Buat cabang untuk setiap nilai
C. dalam kasus industri
D. Ulangi proses untuk setiap cabang. Semua kasus cabang memiliki kelas masalah yang sama.

2.5 Tahapan Penelitian

Berikut ini adalah alur dari tahapan penelitian yang disajikan dalam gambar 1 .



Gambar 1. Tahapan Penelitian

Tahapan dalam penelitian ini yaitu :

1. Pengumpulan Data
   Dataset didapatkan dari website kaggle dengan data berasal dari Rumah Sakit Pima Indian. Dataset ini berjumlah sebanyak 324 data. Dataset ini memiliki 8 atribut dan 1 label dengan tipe data polinomial dan binomial.
2. Perumusan masalah
   Bagaimana mengklasifikasikan data penyakit gula darah pada wanita ?
3. Studi Literatur yang Terkait
4. Pengolahan Data
   Pada tahapan ini dataset penyakit diabetes yang bertipe data numerik akan ditransformasikan menjadi data bertipe kategori.
5. Decision Tree
   Tahapan ini untuk mengklasifikasikan data uji/data training dari dataset diabetes, dimana dilakukan perhitungan *entropy* (S) sebagai parameter yang berfungsi sebagai informasi nilai tiap atribut. Rumus untuk menentukan entropy adalah:

   $$Entropy(S) = \sum_{i=1}^{n} -\text{pi} \times \log_2 pi \quad (2)$$

   Keterangan:
   S: menyatakan himpunan pada kasus
   N: menyatakan jumlah pada pasrtisi atribut A
   |Si|: menyatakan peluang dihasilkan melalui (ya/tidak) dibagi total kasus.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

a) Menaksir nilai gain(S,A) dipakai untuk mengukur efektivitas masing – masing atribut pada node untuk mengklasifikasikan data dan nilai tertinggi akan menjadi akar pohon utama, rumus :

$$Gain(S,A) = Entropy(S) = \sum_{i=1}^{n} \frac{|Si|}{|S|} \times Entropy(S) \quad (3)$$

Dengan
S: menyatakan himpunan kasus
A: menyatakan atribut
N: menyatakan jumlah partisi atribbut pada A
|Si|: menyatakan jumlah kasus pada partisi ke i
|S|: menyatakan jumlah kasus dalam

b) Mengulang proses langkah kedua dengan menentukan entropy sampai semua record telah terpartisi
c) Partisi akan berhenti jika :
   i. Semua record bersampul n mendapat kelas yang sama
   ii. Pada record yang terpartisi tidak ada atribut
   iii. Pada cabang kosong tidak ada record

6. Pengujian Data
Pada tahap ini dataset penyakit diabetes akan diuji menggunakan 30 dataset dari data testing dan 324 dataset dari data training yang dilakukan dengan perhitungan secara manual untuk menghasilkan node pohon keputusan.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Sumber informasi dalam penelitian ini berasal dari website dataset Kaggle pada data Rumah Sakit Pima Indian. Data tersebut mencakup dari beberapa atribut, diantaranya: Pregnancies, Glucose, Bloodpresure, Skinthickness, Insulin, BMI, Diabetes Pedigree Function, Age, dan Outcome dengan jumlah data yang didapat adalah 768 record data. Data yang digunakan dalam penelitian ini adalah data penyakit diabetes yang diderita pada wanita di India. Algoritma Decision Tree dan perangkat lunak Rapidminer digunakan untuk membuat model aturan dari kumpulan data yang penulis kumpulkan. Tabel 1 merupakan dataset penelitian yang digunakan.

Tabel 1. Dataset Penelitian

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	414	33.6	0.627	50	1
1	85	66	29	539	26.6	0.351	31	0
8	183	64	24	18	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	145	76	34	560	44.2	0.63	31	1
5	116	74	33	120	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	90	42	95	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	30	95	30.5	0.232	54	1
4	110	92	45	95	37.6	0.191	30	0
10	168	74	24	120	38	0.537	34	1
10	139	80	33	100	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	90	33	100	30	0.484	32	1
0	151	90	46	560	42.1	0.371	21	1
7	107	74	33	100	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	33	152	35.4	0.388	50	0
7	196	90	33	152	39.8	0.451	41	1
9	119	80	35	152	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	33	152	39.4	0.257	43	1

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

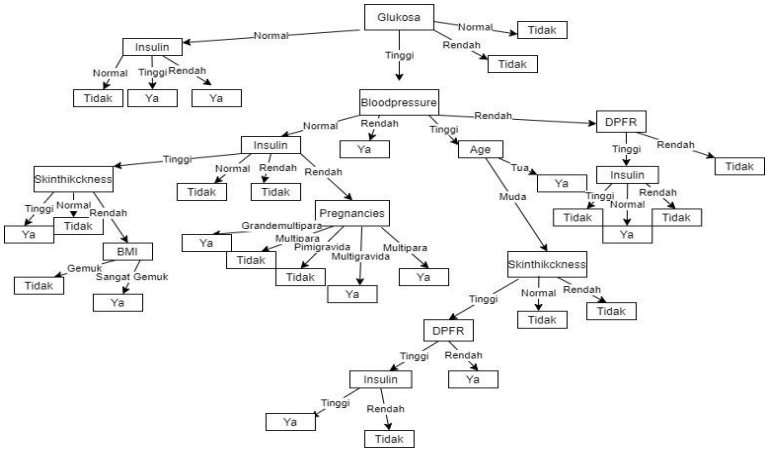
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	97	66	15	140	23.2	0.487	22	0
13	145	82	19	110	22.2	0.245	57	0

Untuk mendapatkan hasil perhitungan menggunakan algoritma C4.5 decision tree, sebelumnya perlu dilakukan transformasi data untuk memudahkan dalam proses perhitungan sesuai dengan gambar pada tabel 1. Kriteria kadar gula darah akan bernilai normal jika kriteria antara 70 - 100mgdl, bernilai rendah jika kriteria dibawah 70mgdl, dan bernilai tinggi jika kriteria bernilai lebih dari 100mgdl. Bloodpressure seseorang dikatakan normal jika kriteria bernilai 120/80 mmHg, bernilai tinggi jika melebihi dari 80 mmHg. Normal skinthickness seseorang akan dikatakan jika bernilai 4.90 – 21mm. Normal insulin berkisar antara 140-199 mgdl, bernilai tinggi jika lebih dari 200mgdl. Normal bmi seseorang berkisar antara 18,5-25. Normal dpfr seseorang bernilai jika kurang dari 80, dan bernilai tinggi jika lebih dari 90. Batas usia muda adalah dari umur 10-44 tahun, umur pertengahan antara 45-59 tahun, usia lanjut 60-74 tahun, usia lanjut tua 75-90 tahun, dan umur sangat tua lebih dari 90 tahun. Normal dari Hasil mengubah data numerik menjadi data kategorikal, ditunjukkan pada Tabel 2.

Tabel 2. Mengubah data menjadi kategori

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Grandemultipara	Tinggi	Rendah	Tinggi	Tinggi	kegemukan	Tinggi	Tua	Positif
Pimigravida	Normal	Rendah	Tinggi	Tinggi	Normal	Tinggi	Muda	Negatif
Grandemultipara	Tinggi	Rendah	Tinggi	Rendah	Normal	Tinggi	Muda	Positif
Pimigravida	Normal	Rendah	Tinggi	Rendah	Gemuk	Tinggi	Muda	Negatif
Nulipara	Tinggi	Rendah	Tinggi	Tinggi	sangat gemuk	Rendah	Muda	Positif
Grandemultipara	Tinggi	Rendah	Tinggi	Rendah	kegemukan	Tinggi	Muda	Negatif
Multipara	Normal	Rendah	Tinggi	Rendah	Gemuk	Tinggi	Muda	Positif
Grandemultipara	Tinggi	Tinggi	Tinggi	Rendah	Gemuk	Tinggi	Muda	Negatif
Multigravida	Tinggi	Rendah	Tinggi	Tinggi	Gemuk	Tinggi	Tua	Positif
Grandemultipara	Tinggi	Tinggi	Tinggi	Rendah	sangat gemuk	Tinggi	Tua	Positif
Multipara	Tinggi	Tinggi	Tinggi	Rendah	sangat gemuk	Tinggi	Muda	Negatif
Grandemultipara	Tinggi	Rendah	Tinggi	Rendah	sangat gemuk	Tinggi	Muda	Positif
Grandemultipara	Tinggi	Normal	Tinggi	Rendah	Normal	Tinggi	Tua	Negatif
Pimigravida	Tinggi	Rendah	Tinggi	Tinggi	Normal	Tinggi	Tua	Positif
Grandemultipara	Tinggi	Rendah	Rendah	Normal	kegemukan	Tinggi	Tua	Positif
Grandemultipara	Normal	Tinggi	Tinggi	Rendah	Normal	Tinggi	Muda	Positif
Nulipara	Tinggi	Tinggi	Tinggi	Tinggi	sangat gemuk	Tinggi	Muda	Positif
Grandemultipara	Tinggi	Rendah	Tinggi	Rendah	Normal	Tinggi	Muda	Positif
Pimigravida	Tinggi	Rendah	Tinggi	Rendah	Normal	Tinggi	Muda	Negatif
Pimigravida	Tinggi	Rendah	Tinggi	Rendah	Normal	Tinggi	Muda	Positif
Multipara	Tinggi	Tinggi	Tinggi	Tinggi	Gemuk	Tinggi	Muda	Negatif
Grandemultipara	Normal	Tinggi	Tinggi	Normal	Normal	Tinggi	Tua	Negatif
Grandemultipara	Tinggi	Tinggi	Tinggi	Normal	Normal	Tinggi	Tua	Positif
Grandemultipara	Tinggi	Tinggi	Tinggi	Normal	Normal	Tinggi	Muda	Positif
Grandemultipara	Tinggi	Tinggi	Tinggi	Normal	Normal	Tinggi	Tua	Positif
Grandemultipara	Tinggi	Rendah	Tinggi	Rendah	Normal	Tinggi	Tua	Positif
Grandemultipara	Tinggi	Rendah	Tinggi	Normal	Normal	Tinggi	Tua	Positif
Pimigravida	Normal	Rendah	Rendah	Normal	Normal	Tinggi	Muda	Negatif
Grandemultipara	Tinggi	Tinggi	Rendah	Rendah	Normal	Tinggi	Tua	Negatif
Grandemultipara	Tinggi	Tinggi	Tinggi	Normal	kegemukan	Tinggi	Muda	Negatif

Hasil transformasi data dibuat oleh proses komputer yang menentukan nilai entropi dari setiap atribut. Nilai entropi yang diperoleh diproses untuk mendapatkan nilai konfirmasi. Kemudian, setelah semua operasi selesai, hasil akhir dari proses tersebut adalah pohon keputusan pada Gambar 2.



Gambar 2. Pohon Keputusan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

Dalam penelitian yang dilakukan memakai data yang diujikan berjumlah 30 data pasien penderita diabetes. Setelah informasi yang diperlukan tersedia, fitur yang diperlukan dari penelitian ini ditetapkanyaitu pemeriksaan pendahuluan terhadap pasien, meliputi: Kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes, usia dan hasil. Sekaligus diprediksi apakah kondisi pasien positif ataupun negatif, yakni variabel hasil. Berikut adalah hasil yang didapat setelah dilakukannya klasifikasi berdasarkan atribut di atas dan diolah menggunakan perhitungan manual C4.5. Kalkulasi yang didapatkan pada kolom dibawah berikut.

Tabel 3. Rekapitulasi Manual Excel

Atribut		Jumlah	Positif	Negatif	Entropy	Gain
Total		30	18	12	0.97095059	0.144484344
Pregnancies	Nulipara	2	2	0	0	
	Pimigravida	6	2	4	0.91829583	
	Multigravida	1	1	0	0	
	Multipara	3	1	2	0.91829583	
	Grandemultipara	18	12	6	0.91829583	
Glucose	Normal	6	2	4	0.91829583	0.05265476
	Tinggi	24	16	8	0.91829583	
Bloodpressure	Rendah	17	12	5	0.87398105	0.075694667
	Tinggi	12	6	6	1	
	Normal	1	0	1	0	
Skinthickness	Tinggi	27	17	10	0.95095605	0.023260567
	Rendah	3	1	2	0.91829583	
Insulin	Normal	8	5	3	0.954434	
	Rendah	15	8	7	0.99679163	
	Tinggi	7	5	2	0.86312057	
Bmi	Normal	16	10	6	0.954434	0.046439345
	Gemuk	5	2	3	0.97095059	
	Kegemukan	4	2	2	1	
	Sangat Gemuk	5	4	1	0.72192809	
Dpf	Rendah	1	1	0	0	0.025116243
	Tinggi	29	17	12	0.97844933	
Age	Muda	18	9	9	1	0.046439345
	Tua	12	9	3	0.81127812	

kolom entropi pencarian 1 adalah:
Entropy keseluruhan
= ((-30/18) * IMLOG2(18/30) + (-12/30) * IMLOG2(12/30)) = 0.9709506
Entropy pada atribut pregnancies
Nulipara = ((-2/2) * IMLOG2(2/2) + (-0/2) * IMLOG2(0/2))
Pimigravida = ((-2/6) * IMLOG2(2/6) + (-4/6) * IMLOG2(4/6))
Multigravida = ((-1/1) * IMLOG2(1/1) + (-0/1) * IMLOG2(0/1))
Multipara = ((-1/3) * IMLOG2(1/3) + (-2/3) * IMLOG2(2/3))
Grandemultipara = ((-12/18) * IMLOG2(12/18) + (-6/18) * IMLOG2(6/18))
Gain total yang didapat = (0.9709506) - (((2/30) * (0) + (((6/30) * (0.91829583) + (((1/30) * (0) + (((3/30) * (0.91829583) + (((18/30) * (0.91829583)))

Entropy total pada attribut glukose
Normal = ((-2/6) * IMLOG2(2/6)) + ((-4/6) * IMLOG2(4/6))
Tinggi = ((-16/24) * IMLOG2(16/24) + (-8/24) * IMLOG2(8/24))
Gain total = (0.97095059) - (((6/30) * (0.91829583) + ((24/30) * (0.91829583)))

Entropy total pada atribut bloodpressure
Normal = ((-0/1) * IMLOG2(0/1) + (-1/1) * IMLOG2(1/1))
Rendah = ((-12/17) * IMLOG2(12/17)) + ((-5/17) * IMLOG2(5/17))
Tinggi = ((-6/12)8IMLOG2(6/12) + (-6/12) * IMLOG2(6/12))
Gain total = (0.97095059) - (((17/30) * (0.87398105) + ((12/30) * (1) + ((1/30) * (0)))

Entropy total pada attribut skinthicness
Normal = ((-17/27) * IMLOG2(17/27) + ((-10/27) * IMLOG2(10/27))
Rendah = ((-1/3) * IMLOG2(1/3) + ((-2/3) * IMLOG2(2/3))
Tinggi = ((-17/27) * IMLOG2(17/27) + ((-10/27) * IMLOG2(10/27))
Gain total = (0.023260567) - (((27/30) * (0.95095605) + ((1/30) * (0.95095605) + ((3/30) * (0.91829583)))

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

Entropy total pada attribut insulin
Normal = ((-17/27) * IMLOG2(17/27) + ((-10/27) * IMLOG2(10/27))
Rendah = ((-1/3) * IMLOG2(1/3) + ((-2/3) * IMLOG2(2/3))
Tinggi = ((-17/27) * IMLOG2(17/27) + ((-10/27) * IMLOG2(10/27))
Gain total = (0.023260567) - (((27/30) * (0.95095605) + ((1/30) * (0.95095605) + ((3/30) * (0.91829583))

Entropy total pada attribut bmi
ormal = ((-10/16) * IMLOG2(10/16) + ((-6/16) * IMLOG2(6/16))
Gemuk = ((-2/5) * IMLOG2(2/5) + ((-2/5) * IMLOG2(2/5))
Kegemukan = ((-2/4) * IMLOG2(2/4) + ((-2/4) * IMLOG2(2/4))
Sangat gemuk = ((-4/5) * IMLOG2(4/5) + ((-1/5) * IMLOG2(1/5))
Gain total = (0.023260567) - (((16/30) * (0.954434) + ((5/30) * (0.97095059) + ((4/30) * (1) + (5/30) * (0.72192809)))

Entropy total pada attribut dpfr
Rendah = ((-1/1) * IMLOG2(1/1) + ((-0/1) * IMLOG2(0/1))
Tinggi = ((-17/29) * IMLOG2(17/29) + ((-12/29) * IMLOG2(12/29))
Gain total = (0.023260567) - (((1/30) * (0) + ((29/30) * (0.025116243)))

Entropy total pada attribut age
Rendah = ((-9/18) * IMLOG2(9/18) + ((-9/18) * IMLOG2(9/18))
Tinggi = ((-9/12) * IMLOG2(9/12) + ((-3/12) * IMLOG2(9/12))
Gain total = (0.023260567) - (((9/30) * (1) + ((12/30) * (0.81127812)))

Data yang telah terklasifikasi kemudian diproses oleh RapidMiner dengan menggunakan efisiensi untuk validasi dan reliabilitas data dalam pencarian akurasi data. Data akurat diolah dengan RapidMiner untuk mengetahui hasil diabetes berdasarkan pengolahan data dengan RapidMiner. Algoritma C4.5 terlebih dulu memilih node untuk dijadikan node dalam mencari total masalah dengan entropi persamaan umum. . Kasus ditentukan oleh riwayat menderita mengalami sakit, beban massa tubuh serta peningkatan glukosa. Perhitungan verifikasi kemudian dikerjakan pada setiap atribut.

Pengecekan hasil kalkulasi langsung dengan perangkat lunak RapidMiner dikerjakan lewat serangkaian pengerjaan hingga didapatkan rekap pengolahan informasi yang dilakukan berjumlah yaitu 76,67%. Dari pohon keputusan yang dihasilkan, nilai prediksi dapat ditentukan dengan menyesuaikan aturan ke kumpulan data, di mana hasil akhir membandingkan nilai yang dihasilkan dan nilai prediksi untuk mendapatkan nilai matriks kebingungan. Berdasarkan pencocokan nilai hasil dengan nilai dalam kumpulan data yang memiliki nilai yang sama dengan hasil prediksi, dengan hasil prediksi positif sebanyak 18 data, prediksi true negatif sebanyak 7, dan false negatif bernilai 5 seperti yang ditunjukkan pada gambar 3 di bawah ini.

	true ya	true tidak	class precision
pred. ya	18	7	72.00%
pred. tidak	0	5	100.00%
class recall	100.00%	41.67%	

Gambar 3. Confussion Matrix

Berdasarkan hasil matriks konfusi pada gambar 3, dilakukan perhitungan matriks konfusi untuk menentukan skor presisi, akurasi, dan recall. Hasilnya ditunjukkan pada Gambar 3. Power vector dengan presisi 76,67%, presisi 72%, dan recall 41,67%. Nilai-nilai disini adalah indikator yang mengukur kinerja algoritma pohon keputusan C4.5. Tingkat akurasi yang tinggi ini dapat dicapai dengan menambahkan atribut baru atau menghilangkan atribut minor, memberikan kesempatan untuk memperoleh tingkat akurasi yang lebih tinggi. Pada penelitian klasifikasi diagnosis diabetes menurut metode decision tree yaitu 76,67, dengan nilai akurasi 72%. Berdasarkan penjelasan pada gambar, didapatkan bahwa metode pohon keputusan C4.5 bernilai baik di klasifikasi data dalam data mining.

4. KESIMPULAN

Hasil klasifikasi penderita diabetes dengan ciri-ciri yang terdapat pada database diabetes yaitu. kehamilan, glukosa, tekanan darah, ketebalan kulit, BMI insulin, riwayat keluarga diabetes, usia, dapat digunakan sebagai informasi dalam klasifikasi diabetes. Hasil yang didapat dengan menggunakan data uji sebanyak 30 dataset yang diujikan melalui software rapidminer menunjukkan bahwa melalui data tersebut kebanyakan pasien penderita dinyatakan positif diabetes. Penelitian ini menggunakan algoritma C4.5 untuk mengklasifikasikan seseorang yang menderita diabetes atau tidak. Dari total 324 data, dibuat tahapan akuisisi data, pengolahan data dan pengujian dengan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 4, Nomor 4, Juni 2023
e-ISSN 2685-998X
DOI 10.30865/json.v4i4.6202

implementasi algoritma C4.5. Untuk perhitungan manual, algoritma C4.5 melibatkan beberapa langkah yaitu menghitung nilai entropy, kemudian mencari nilai entropy, mencari nilai gain, dan setelah mendapatkan nilai gain mencari nilai gain terbesar untuk digunakan sebagai root node. Perhitungan yang dilakukan oleh algoritma C4.5 membentuk pohon keputusan dengan 8 aturan yang diharapkan dapat memberikan informasi tentang penyakit diabetes. Analisis penelitian ini diukur dengan presisi, akurasi dan recall dan hasilnya akurasi 76,67%, akurasi 72% dan recall 41,67%. Saran penambahan fitur baru dapat diberikan dari hasil penelitian ini. Semakin banyak data, semakin akurat dan dengan algoritma yang berbeda perbandingannya dimungkinkan.

REFERENCES

[1] M. I. Mahdi, "Penderita Diabetes Indonesia Terbesar Kelima di Dunia," *2 Februari*, 2022. .
[2] I. Fida *et al.*, "I a c d k p d," pp. 1–8, 2023.
[3] D. K. Gautam *et al.*, "Микроальбуминурия Предиктор Сердечнососудистого Риска У Больных Сахарным Диабетом 1-Го И 2-Го Типов Без Осложнений," *BMC Res. Notes*, vol. 10, no. 1, pp. 1–4, 2019.
[4] S. Putri, E. Irawan, and F. Rizky, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *Januari*, vol. 2, no. 1, pp. 39–46, 2021.
[5] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.
[6] N. Noviandi, "Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit Diabetes," *Indones. Heal. Inf. Manag. J.*, vol. 6, no. 1, pp. 1–5, 2018.
[7] V. No *et al.*, "Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4 . 5 Dengan Rapidminer," vol. 5, no. 2, pp. 73–83, 2023.
[8] M. Yusa, E. Utami, and E. Luthfi. Taufiq, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes," *Infosys (Information Syst. J.*, vol. 4, no. 1, pp. 23–34, 2016.
[9] I. Mubarog, A. Setyanto, and H. Sismoro, "Sistem Klasifikasi Pada Penyakit Breast Cancer Dengan Menggunakan Metode Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 2, p. 109, 2021, doi: 10.24076/citec.2019v6i2.246.
[10] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.
[11] K. F. Irnanda, D. Hartama, and A. P. Windarto, "Analisa Klasifikasi C4.5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 327, 2021, doi: 10.30865/mib.v5i1.2763.
[12] L. N. Rani, "Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit," *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.
[13] E. P. Cynthia and E. Ismanto, "Metode Decision Tree Algoritma C.45 Dalam Mengklasifikasi Data Penjualan Bisnis Gerai Makanan Cepat Saji," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 3, no. July, p. 1, 2018, doi: 10.30645/jurasik.v3i0.60.
[14] D. Sartika and D. I. Sensuse, "Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian," *Jatisi*, vol. 1, no. 2, pp. 151–161, 2017.
[15] B. Novianti, T. Rismawan, and S. Bahri, "Implementasi Data Mining Dengan Algoritma C4.5 Untuk Penjurusan Siswa (Studi Kasus: Sma Negeri 1 Pontianak)," *J. Coding, Sist. Komput. Untan*, vol. 04, no. 3, pp. 75–84, 2016.
[16] H. Widayu, S. D. Nasution, N. Silalahi, and Mesran, "Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5," *Media Inform. Budidarma*, vol. Vol 1, No, no. 2, p. 37, 2017.
[17] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.
[18] I. Junaedi, N. Nuswantari, and V. Yasin, "Perancangan Dan Implementasi Algoritma C4 . 5 Untuk Data Mining," *J. Inf. Syst. Informatics Comput.*, vol. 3, no. 1, pp. 29–44, 2019, [Online]. Available: http://journal.stmikjayakarta.ac.id/index.php/jisicom/article/view/203%0Ahttp://journal.stmikjayakarta.ac.id/index.php/jisicom/article/download/203/158.
[19] J. Eska, "Penerapan Data Mining Untuk Prekdiksi Penjualan Wallpaper Menggunakan Algoritma C4.5 STMIK Royal Ksiaran," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 2, pp. 9–13, 2016.
[20] F. Riandari and A. Simangunsong, "Penerapan Algoritma C4.5 Untuk Mengukur Tingkat Kepuasan Mahasiswa," *Pap. Knowl. . Towar. a Media Hist. Doc.*, vol. 5, no. 2, pp. 40–51, 2019.