JIKA (Jurnal Informatika) Universitas Muhammadiyah Tangerang *P ISSN : 2549-0710*
Tangerang, Januari 2024, Vol 8 No.1, pp 74-81 *E ISSN : 2722-2713*

PREDIKSI PENYAKIT DIABETES DENGAN MENGGUNAKAN ALGORITMA C4.5

Maryanah Safitri[1)], Ardian Dwi Praba[2)]

[1] Informatika, Universitas Nusa Mandiri, Jl. Jatiwaringin Raya No.2 Jakarta Timur
[2] Sistem Informasi, Universitas Nusa Mandiri, Jl. Jatiwaringin Raya No.2 Jakarta Timur
Co Responden Email: ardian.ddw@nusamandiri.ac.id

Article history
Received 07 Nov 2023
Revised 10 Dec 2023
Accepted 20 Jan 2024
Available online 27 Jan 2024

Keywords
Data mining,
Decision tree,
Algoritma,
Dataset,
Diabetes

Abstract

Diabetes is a growing global health challenge, demanding accurate predictions for early intervention and effective management. In this study, we implemented the C4.5 algorithm, a proven machine learning approach, to project the risk of diabetes in specific populations. We also conducted a comprehensive analysis of the most significant clinical attributes, providing detailed insights into the disease profile and influencing factors. The results of this research are expected to serve as a foundation for the development of more effective prevention strategies and personalized treatment recommendations for individuals vulnerable to diabetes. By utilizing a comprehensive clinical dataset, we successfully developed a predictive model capable of identifying key risk factors with an accuracy rate of 96.16%. Thus, this research makes a significant contribution to global efforts aimed at reducing the prevalence of diabetes and enhancing the quality of life for patients.

Riwayat
Diterima 07 Nov 2023
Revisi 10 Des 2023
Disetujui 20 Jan 2024
Terbit 27 Jan 2024

Kata Kunci
Data mining,
Pohon Keputusan,
Algoritma,
Dataset,
Diabetes

Abstrak

Penyakit diabetes merupakan tantangan kesehatan global yang terus meningkat, memerlukan prediksi yang tepat untuk intervensi dini dan manajemen yang efektif. Dalam studi ini, kami menggunakan algoritma C4.5, suatu metode pembelajaran mesin yang terbukti, untuk memproyeksikan risiko diabetes pada kelompok khusus. Analisis mendalam terhadap atribut klinis yang paling penting juga kami lakukan, memberikan wawasan yang lebih rinci tentang profil penyakit dan faktor-faktor yang mempengaruhi. Harapannya, Hasil riset ini dapat menjadi panduan untuk mengembangkan strategi pencegahan yang lebih efektif dan rekomendasi pengobatan individual untuk individu yang menderita diabetes. Dengan menggunakan kumpulan data klinis yang komprehensif, kami berhasil membuat model prediksi yang mampu mengidentifikasi faktor risiko utama dengan tingkat akurasi mencapai 96.16%. Oleh karena itu, riset ini memberikan masukan yang sangat baik dalam upaya global untuk mengurangi prevalensi diabetes dan meningkatkan kualitas hidup pasien.

PENDAHULUAN

Penyakit gula dasrah adalah sekelompok penyakit metabolik yang ditandai dengan tingginya kadar gula darah pada seseorang yang terkena, dan bertahan dalam jangka waktu lama. Kondisi ini terjadi ketika badan tidak mampu lagi menghasilkan hormon insulin dalam nilai yang cukup sehingga menyebabkan peningkatan kadar gula dalam darah.(Muhamad Ichsan Gunawan, 2020). Insulin, hormon alami yang dibuat oleh sel beta pankreas, yang bertanggung jawab atas penggunaan gula oleh tubuh sebagai sumber energi, diproduksi oleh pankreas, yang menyebabkan diabetes mellitus adalah gangguan metabolisme. Jumlah penderita diabetes terus meningkat setiap tahunnya. Jumlah penderita diabetes di Indonesia berkisar 10 juta pada tahun 2015. Diabetes merupakan masalah kesehatan global dengan perkiraan jumlah penderita mencapai sekitar 120 juta di seluruh dunia. Jumlah penderita diabetes diperkirakan akan terus meningkat jika pengetahuan masyarakat umum mengenai faktor pemicu penyakit diabetes tidak memadai (Sanni Ucha Putri, 2021). Di samping itu, penyakit diabetes juga memberikan beban ekonomi yang signifikan

pada sistem kesehatan di berbagai negara. Menurut data dari Organisasi Kesehatan Dunia (WHO), prevalensi diabetes telah mengalami peningkatan yang signifikan. Diabetes didiagnosis ketika kadar glukosa darah seseorang melebihi batas normal (Nurlina, 2019). Penambahan kadar gula darah dan gangguan metabolisme lemak, karbohidrat, dan juga protein menyebabkan diabetes. Kurangnya insulin adalah penyebab utamanya. Hal ini mungkin disebabkan oleh kurangnya produksi insulin oleh sel beta Langerhans di pankreas, atau kurangnya respon tubuh terhadap insulin. (Ramadhan, 2019). Selama beberapa dekade ke depan, diperkirakan jumlah penderita diabetes akan terus meningkat secara signifikan. Oleh karena itu, penting untuk mengembangkan metode yang efektif untuk memprediksi risiko diabetes, sehingga tindakan pencegahan dan pengobatan dapat dilakukan tepat waktu.

Penambangan data adalah bahasa yang dipakai untuk menggambarkan proses menemukan informasi yang ada dalam basis data. Proses ini bertujuan untuk mengekstrak informasi yang penting dari data dalam jumlah besar yang memiliki pola yang beragam. (Deny Jollyta, 2020). Data mining adalah proses semi-otomatis yang mengandalkan teknik statistik, kecerdasan buatan, matematika, dan pembelajaran mesin untuk mengekstrak dan mencari informasi pengetahuan bernilai yang tersembunyi dalam kumpulan data besar. Kegiatan data mining merupakan bagian integral dari proses penemuan pengetahuan dalam database , yang melibatkan beberapa tahap seperti pemfilteran data, pra-pemrosesan, transformasi, penambangan data, dan evaluasi hasil. KDD juga dikenal sebagai proses penemuan pengetahuan dalam basis data (Zai, 2022).

Klasifikasi adalah salah satu dari lima fungsi utama dalam penambangan data. Sebagai bagian dari pembelajaran terbimbing, klasifikasi adalah proses prediksi terhadap objek yang belum memiliki kelas atau label. Metode Pohon keputusan (decision tree (DT)) adalah salah satu jenis classifier yang dapat dijelaskan sebagai partisi rekursif dari dataset. (Handayani, 2020). Algoritme ini memiliki kemudahan untuk dipahami, fleksibilitas, serta kesan yang mengagumkan karena dapat direpresentasikan dalam bentuk pohon keputusan. (Hana, 2020)

Dalam bidang ilmu data, algoritma C4.5 merupakan instrumen yang efekti untuk melakukan analisis data dan membuat prediksi berdasarkan kumpulan data tertentu. Algoritma ini memanfaatkan sebuah model keputusan berbentuk pohon untuk menghasilkan prediksi (Rokhman, 2021). Algoritma ini tergolong keluarga algoritma pembelajaran mesin. Algoritma ini dapat membuat model klasifikasi berbentuk pohon keputusan, yang dapat menyusun peraturan prediksi berbasis sejumlah atribut klinis.

Metode Pohon keputusan adalah representasi seperti algoritma dimana setiap node mewakili atribut yang akan diuji. Setiap cabang mencerminkan hasil tes, dan simpul daun mencirikan kumpulan kelas tertentu. Bagian atas dari pohon keputusan disebut root, dan biasanya merupakan atribut yang mempunyai dampak terbesar pada kelas tertentu. Umumnya, pohon keputusan menggunakan pendekatan pencarian top-down untuk menemukan solusi. (Nasrulloh, 2021).

Dalam struktur algoritma C4.5, data biasanya ditunjukkan berbentuk tabel yang terdiri dari atribut dan catatan. Dalam pembentukan pohon, kriteria tertentu digunakan. Sebagai contoh, cuaca, angin, dan suhu adalah faktor yang dipertimbangkan untuk menentukan apakah seseorang akan bermain tenis. Salah satu atribut ini, atribut hasil, berfungsi sebagai penentu solusi untuk setiap item data. ID3, C4.5, dan CART adalah beberapa metode yang dapat digunakan untuk pembentukan sebuah pohon keputuasn.

Dengan menerapkan algoritma Decision Tree pada data klinis yang relevan, tujuan dari penelitian ini adalah mengembangkan model prediktif yang akurat untuk meramalkan risiko diabetes dan mengenali faktor-faktor yang paling signifikan dalam perkembangan penyakit ini. Harapannya, penelitian ini dapat memberikan petunjuk berharga bagi praktisi kesehatan untuk memahami dan mengelola diabetes dengan lebih efektif, serta berkontribusi pada upaya pencegahan dan mengurangi beban penyakit diabetes secara global.

Metode *mechine learning* adalah sebuah metode yang digunakan untuk mengidentifikasi keterkaitan krusial, pola, serta hubungan dalam kumpulan data melalui analisis data. Terdapat serangkaian enam

teknik utama yang sering dipakai dalam praktik data mining, di antaranya adalah prediksi atau peramalan, deskripsi atau visualisasi data, klasifikasi, estimasi, asosiasi, dan pengelompokan atau (Hana, 2020).

a. Deskripsi
Proses ini bertujuan untuk mengenali pola yang timbul berulang kali dalam kumpulan data, dan kemudian menggunakan pola-pola tersebut sebagai landasan untuk merumuskan kriteria serta alur yang gampang untuk dipahami..

b. Ramalan
Motode klasifikasi memberikan data ke kelas berdasarkan prediksi tingkah laku atau value masa depan .

c. Pengelompokan
Pengelompokan adalah suatu metode untuk mengenali ciri-ciri tertentu dari suatu data dan mengelompokkan nilai tersebut ke dalam kelas-kelas yang sesuai.

d. Estimasi
Estimasi adalah proses memperkirakan jumlah, ukuran, atau nilai suatu entitas berdasarkan informasi yang tersedia, dengan tujuan memberikan perkiraan yang seakurat mungkin dalam suatu konteks tertentu.

e. Clustering
Klustering adalah proses pengelompokan atau klasterisasi data, di mana data dengan karakteristik yang mirip dikelompokkan bersama dalam satu kelompok, sementara data yang memiliki perbedaan dikelompokkan secara terpisah...

f. Asosiasi
Ketika menggunakan teknik asosiasi, kita mencari karakteristik yang muncul dalam kondisi khusus atau membuat aturan asosiatif antara kombinasi item.

Klasifikasi merujuk pada proses mengenali atribut-atribut khusus dari data dan mengelompokkannya ke dalam kelas-kelas yang sesuai berdasarkan karakteristik individu tersebut. Dalam proses klasifikasi, dilakukan identifikasi karakteristik suatu objek, dan objek yang memiliki karakteristik serupa ditempatkan ke dalam ruang-ruang yang telah ditetapkan sebelumnya.. (Ikhsan Romli, 2020).

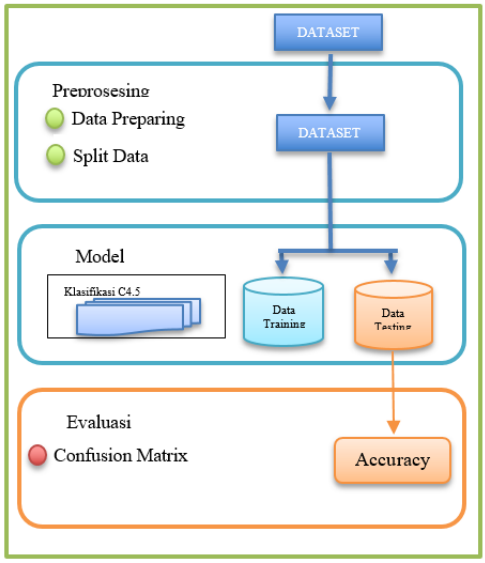
Dalam pelatihan, digunakan data baru yang disebut sebagai data pengujian. Hasil dari proses ini adalah probabilitas tertentu yang terkait dengan data pengujian. Dalam mengklasifikasikan dataset, sangat penting bahwa setiap entri data memiliki label atau atribut yang dituju. Tujuan utama dari proses klasifikasi adalah memprediksi kelas objek untuk setiap masalah dalam dataset. Sebuah tugas klasifikasi biasanya dimulai dengan kumpulan data di mana kelasnya sudah diketahui. Salah satu jenis tugas klasifikasi yang sangat mudah adalah pengelompokan biner. (Chan, 2018).

Proses klasifikasi melibatkan pengolahan nilai yang telah ada sebelumnya, yang sering disebut sebagai data mentah.

Proses klasifikasi sangat penting untuk memprediksi kelas-kelas objek yang tidak berlabel dengan menemukan metode dan fungsi yang dapat menjelaskan atau memisahkan konsep dan kelas dalam data. Model dapat menggunakan pohon keputusan, aturan “jika-maka”, atau rumus. Metode klasifikasi ini sangat populer karena pohon keputusan mudah dipahami orang. (Siska Febriani, 2021). Selain itu, ada banyak cara yang dapat digunakan untuk menyelesaikan permasalahan terkait pengelompokan dan pembelajaran terawasi.

METODE PENELITIAN

Dalam penelitian ini, kami akan menerapkan pendekatan eksperimental untuk menyelesaikan studi kami. Kami akan menggunakan algoritma klasifikasi C4.5 untuk menganalisis dataset. Tahap pertama adalah menyiapkan dataset. Kemudian kita bagi menjadi dataset latih dan dataset uji, dengan perbandingan 80% dataset latih dan 20% dataset uji. Gunakan data pelatihan untuk membentuk pola dan model, dan data uji untuk menguji model. Algoritma yang digunakan adalah model pengelompokan yang menggunakan algoritma C4.5. Kemudian, dilakukan evaluasi dengan menggunakan matriks kebingungan dan menghasilkan nilai akurasi. Dalam konteks klasifikasi, akurasi adalah metrik evaluasi yang digunakan untuk menilai sejauh mana algoritma klasifikasi dapat mengidentifikasi kelas target dengan baik dari semua kasus yang diamati. Akurasi dinyatakan dalam bentuk persentase dan dihitung sebagai rasio perkiraan yang benar dibandingkan dengan jumlah prediksi. Informasi lengkap tentang langkah-langkah penelitian dapat ditemukan dalam gambar di bawah ini:



Gambar 1. Kerangka Penelitian

A. Sumber Data

Riset ini mengambil data dari internet yaitu dataset UCI Machine Learning Repository yang terletak di alamat atau link https: //t.ly/Il2cK. Dataset yang kami gunakan adalah dataset prediksi risiko diabetes tahap awal yang disebut diabetes_data_upload.csv. Penelitian ini menggunakan 17 field dan total data sebanyak 520 item. Data ini mencakup informasi tentang orang yang mengalami gejala yang dapat memicu penyakit diabetes. Dataset ini diperoleh melalui respons kuesioner secara langsung dari orang-orang yang sudah positif didiagnosis menderita diabetes atau yang belum menderita diabetes namun punya banyak gejala. Data ini dikumpulkan dari responden yang mengisi kuesioner di Rumah Sakit Diabetes Sylhet, Bangladesh.

B. Split Data Otomatis

Pada tahapan ini, 520 dataset diabetes akan dipceah dalam dua kemlompok: dataset pelatihan dan dataset pengujian. Persentase pelatihan akan digunakan untuk membentuk pola atau model, sedangkan dataset pengujian akan digunakan untuk menguji model tersebut..

C. Model Yang Diusulkan

Dalam Artikel ini kami menggunakan algoritma klasifikasi data mining dengan algoritma *decision tree* yang terbukti efektif. Dalam proses pembuatan pohon keputusan, algoritma C4.5 telah terbukti mampu menghasilkan langkah-langkah yang jelas dan terstruktur:

1. Menciptakan akar adalah tahap pertama dalam pembentukan pohon keputusan. Setelah itu, data akan dibagi berdasarkan karakteristik yang relevan untuk membentuk daun.
2. Langkah berikutnya adalah pemangkasan pohon keputusan yang telah terbentuk. Pemangkasan ini mencakup identifikasi dan pemotongan cabang yang tidak digunakan dari pohon yang telah terbentuk. Tujuan pemangkasan ini selain meminimalisir ukuran pohon adalah untuk meminimalisir kesalahan dalam memprediksi kasus baru dari proses divide and conquer. Pemangkasan dilakukan dengan dua metode: pre-pruning dan post-pruning.
3. Setelah itu, langkah berikutnya adalah merumuskan aturan keputusan dari pohon yang telah terbentuk. Aturan-aturan ini didapatkan dengan mengikuti jejak dari akar hingga ke daun dalam pohon keputusan.

Algoritma Pohon keputusan untuk membangun sebuah pohon keputusan pada dasarnya digambarkan sebagai berikut:

a. Dengan syarat tertentu, kita dapat menghitung total dataset berdasarkan anggota atribut hasil. Meskipun syarat tersebut pada awalnya belum ditentukan, kita dapat menentukannya pada waktunya.
b. Ketika menggunakan pohon keputusan untuk klasifikasi, kita memilih atribut sebagai node untuk membagi data.
c. Bentuk cabang untuk setiap anggota dari Node.
d. Ketika menggunakan pohon keputusan untuk klasifikasi, penting untuk memeriksa apakah tersedia node yang memiliki nilai entropy nol. Jika ada, perlu ditentukan daun yang telah terbentuk dari node tersebut. Namun, jika semua node memiliki nilai entropy nol, maka proses harus dihentikan..
e. Jika terdapat node dengan nilai entropi yang tinggi besar dari nol, ulangi proses dari awal dengan node sebagai syarat, dan lanjutkan hingga semua node memiliki nilai entropi nol..

Node adalah atribut yang mempunyai gain terbesar dari sekumpulan atribut. Gain suatu atribut dapat dihitung dengan menggunakan formula yang ditemukan dalam rumus berikut ini :

$$Gain(S, A) = Entropy(S) - \left(\sum_{i=1}^{n} \frac{A_i}{S} * Entropy(A_i) \right)$$

Keterangan: *S* : Kasus.
A : Atribut
n : Jumlah partisi atribut *A*
A_i : Jumlah kasus pada partisi ke-i.
S : Jumlah kasus.

Sementara itu, untuk menghitung nilai *Entropy* dapat dilihat pada persamaan berikut ini:

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

Keterangan: *S* : Himpunan kasus.
n : Jumlah partisi *S*
pi : Proporsi dari S_i ke *S*.

4. Confussion Matrix

Confusion Matrix adalah hasil perhitungan dari suatu pengelimppokan data mining yang ditampilkan dalam format tabel. Sering digunakan sebagai cara untuk mengukur tingkat keakuratan, confusion matrix digunakan untuk mengukur kinerja. Empat istilah menggambarkan hasil klasifikasi dalam penggunaan confusion matrix. Keempat istilah ini mencakup:

1. Kesalahan positif (FP) terjadi ketika data negatif salah diprediksi sebagai data positif.
2. Kesalahan negatif (FN) terjadi ketika data positif salah diprediksi sebagai data negatif.
3. Positif benar (TP) terjadi saat data positif diprediksi dengan benar.
4. Negatif benar (TN) terjadi saat data negatif diprediksi dengan benar.

Dalam metode klasifikasi yang sebenarnya, kita dapat mengamati bentuk Confusion Matrix secara singkat pada tabel di bawah ini.:

Tabel 1. Tabel Confussion Matrix

Clasifikasi		**Predicted Class**	
		Y	**N**
Observed Class	Kelas Y	A(True Positive)	B(False negative)
	Kelas N	C(Class Posotove)	D(True Negative)

Rumus Perhitungan untuk mendapatkan nilai akurasi seperti ini

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN} x\ 100\%$$

Algoritma klasifikasi bertujuan untuk mengembangkan model yang dapat memberikan tingkat akurasi yang terbaik. Evaluasi kinerja model dari metode klasifikasi dilakukan saat model diuji menggunakan dataset testing, karena penting bagi model untuk secara konsisten memberikan prediksi yang akurat terhadap data training. Sensitivitas atau Recall adalah sebuah ukuran yang mengukur rasio peramalan benar positif terhadap jumlah total keseluruhan data yang sebenarnya positif, sehingga dapat mengevaluasi sejauh mana kemampuan pengujian untuk megetahui hasil positif dari kumpulan data yang seharusnya positif. Sensitivitas atau Recall dapat dihitung dengan persamaan berikut ini:

$$Sensitivitas = \frac{TP}{\text{TP} + \text{FN}}$$

Namun, presisi dapat didefinisikan sebagai rasio dari hasil prediksi yang benar dan positif terhadap total hasil prediksi yang positif. Dalam dunia data mining, presisi diukur dengan membagi jumlah data asli positif dengan total jumlah data yang diidentifikasi sebagai positif. Ini menunjukkan kinerja sebuah sistem dalam menghasilkan data yang sesuai. Untuk menghitung presisi, persamaan di bawah ini digunakan:

$$Sensitivitas = \frac{TP}{\text{TP} + \text{FN}}$$

HASIL DAN PEMBAHASAN

Dataset publik yang digunakan dalam penelitian bersifat publicyang berasal dari Repositori Pembelajaran Mesin Uci, yang merupakan kumpulan data tentang prediksi risiko diabetes tahap awal. Terdapat 520 catatan dalam dataset ini. Untuk pelatihan, digunakan 110 entri data dengan 16 fitur. Dari 16 fitur tersebut, 15 fitur merupakan atribut reguler dan 1 fitur merupakan atribut khusus. Fitur-fitur yang digunakan termasuk:

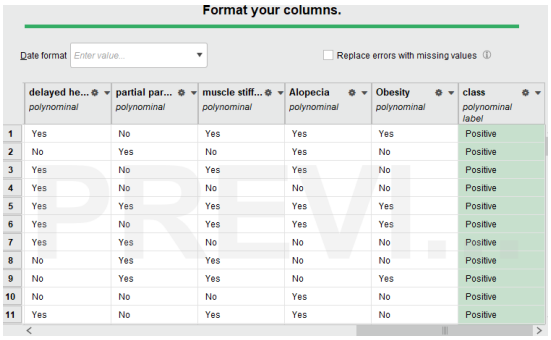
Table 2. Atribut dataset

No	**Atribut**	**Tipe**	**Nilai atribut**
1	Gender	Polynominal	Male, Female
2	Poliuria	Polynominal	Y, N

3	Polidipsia	Polynominal	Y, N
4	Sudden weight loss	Polynominal	Y, N
5	Weakness	Polynominal	Y, N
6	Polyphagia	Polynominal	Y, N
7	Genital thrush	Polynominal	Y, N
8	Visual blurring	Polynominal	Y, N
9	Itching	Polynominal	Y, N
10	Irritability	Polynominal	Y, N
11	Delayed healing	Polynominal	Y, N
12	Partial paresis	Polynominal	Y, N
13	Muscle stiffness	Polynominal	Y, N
14	Alopecia	Polynominal	Y, N
15	Obesity	Polynominal	Y, N
16	Class	Polynominal	Positive, Negative

Pengujian dataset dilaksanakan dengan memanfaatkan aplikasi Rapid Miner. Rapid Miner merupakan software yang didevelop oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg. Aplikasi ini didesain dengan antarmuka yang user-friendly untuk mempermudah pengguna atau user dalam mengoperasikan aplikasi atau software tersebut. (Srisulistiowati, 2021).

Berikut ini adalah gambaran dataset yang diakses dengan mengunakan aplikasi rapidminer sebelum dilakukan pengujian.



	delayed he... polynominal	partial par... polynominal	muscle stiff... polynominal	Alopecia polynominal	Obesity polynominal	class polynominal label
1	Yes	No	Yes	Yes	Yes	Positive
2	No	Yes	No	Yes	No	Positive
3	Yes	No	Yes	Yes	No	Positive
4	Yes	No	No	No	No	Positive
5	Yes	Yes	Yes	Yes	Yes	Positive
6	Yes	No	Yes	Yes	Yes	Positive
7	Yes	Yes	No	No	No	Positive
8	No	Yes	Yes	No	No	Positive
9	No	Yes	Yes	No	Yes	Positive
10	No	No	No	Yes	No	Positive
11	Yes	No	Yes	Yes	No	Positive

Gambar 1 . dataset di rapidminer

Pada tahapan ini, dilakukan eksperimen dan pengujian metode dengan melakukan perhitungan dan menemukan aturan-aturan yang terkandung dalam algoritma yang diajukan, yaitu Algoritma C4.5. Proses dimulai dengan mencari akar pohon pada decision tree. Untuk menentukan akar pohon, kita mencari nilai gain dari masing-masing atribut. Atribut yang memiliki nilai gain terbanyak akan menjadi akar pohon. Sebelum mencari nilai gain, langkah awal adalah menentukan nilai entropy.

Langkah awal adalah menemukan nilai total entropy. Dari total 110 data, terdapat 58 data positif dan 52 data negatif, sehingga:

Entropy (Total) = ((- 8/110)*log2(58/110)+
(-52/110)*log2(52/110))
= 0,997852777

Kemudian hitung nilai entropy dan gain dari masing-masing atribut, sebagai berikut:

Gender (Male) =((24/74)*log2(24/74)+
(-50/74)*log2(50/74))=
0,909022156

Gender (Female) =((34/36)*log2(34/36)+
(-2/36) * log2 (2/36)) =
0,309543429

Gain= 0,997852777 – ((74 / 110)*
0,909022156) – ((36 / 110) *
0,309543429) = 0,285023658

Lakukan tahapan perhitungan serupa untuk menemukan nilai entropi dan gain dari atribut-atribut lainnya.

Kesimpulan dari perhitungan tersebut adalah sebagai berikut:

1. Atribut Gender memiliki nilai gain terbanyak, yaitu 0,285023658. Hal ini menunjukkan bahwa atribut Gender dapat digunakan untuk memprediksi kelas diabetes dengan lebih akurat.
2. Atribut Age memiliki nilai gain terendah, yaitu 0,000000000. Hal ini menunjukkan bahwa atribut Age tidak dapat digunakan untuk memprediksi kelas diabetes.

Berikut ini adalah penjelasan singkat dari masing-masing atribut:

1. Atribut Gender: Atribut ini memiliki nilai gain yang tinggi karena dapat mengurangi ketidakpastian dalam memprediksi kelas diabetes. Hal ini karena pria lebih berisiko terkena diabetes daripada wanita.
2. Atribut Age: Atribut ini memiliki nilai gain yang rendah karena tidak dapat mengurangi ketidakpastian dalam

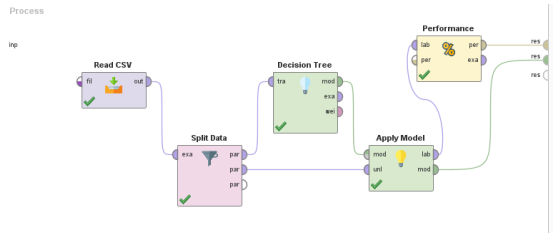
memprediksi kelas diabetes. Hal ini karena tidak ada hubungan yang jelas antara usia dan risiko diabetes.

Secara keseluruhan, hasil perhitungan tersebut menunjukkan bahwa atribut Gender adalah atribut yang paling relevan untuk memprediksi kelas diabetes. Atribut ini dapat digunakan untuk mengurangi ketidakpastian dalam memprediksi kelas diabetes, sehingga dapat meningkatkan akurasi klasifikasi.

Percobaan Algoritma C4.5

Sekarang kita akan melihat hasil dari percobaan menggunakan validasi silang pada aplikasi RapidMiner.

Berikan penjelasan lebih detail terkait dengan aplikasi aplikasi RapidMiner yang digunakan dalam penelitian ini, bagaimana proses-proses yang berlangsung step by step.



Gambar 2 . Pengujian metode

Pada pengujian ini ada beberapa langkah yang harus dikerjakan mulai dari load dataset sampai hasil akhir.

1. Fungsi Read CSV digunakan untuk mengambil dataset yang akan digunakan dalam pengujian.
2. Teknik Split merupakan metode validasi yang mengacak pembagian dataset menjadi dua kelompok, di mana sebagian berperan sebagai dataset training dan sebagian lagi sebagai dataset testing.
3. Decision Tree merupakan metode yang digunakan untuk melakukan uji pada dataset.
4. Fungsi Apply Model digunakan untuk mendapatkan prediksi pada data yang tidak terlihat atau melakukan transformasi data dengan menjalankan model preprocessing.
5. Kegunaan Performance adalah untuk mengevaluasi kinerja model.

Singkatnya, proses pembuatan model adalah sebagai berikut: model digunakan untuk menguji akurasi prediksinya dengan memasukkan data uji dari data pelatihan dengan pemodelan "Tree-Decision Tree". Selanjutnya, untuk menguji tingkat akurasi, penggunaan model dan validasi "Performance - Predictive % Performance (Clasification)" pada aplikasi RapidMiner..

Evaluasi dan Validasi Hasil

Untuk mengevaluasi kinerja metode C4.5, kami dapat menguji akurasi prediksi setelah selesainya proses pengolahan data. Tujuan dari penelitian ini adalah untuk menilai akurasi analisis data pasien diabetes mellitus dan memprediksi apakah gejala pasien dapat menunjukkan apakah mereka positif atau negatif menderita diabetes mellitus. Kita akan menggunakan matriks konfusi untuk mengevaluasi akurasi.

accuracy: 97.88%

	true Positive	true Negative	class precision
pred. Positive	317	8	97.54%
pred. Negative	3	192	98.46%
class recall	99.06%	96.00%	

Gambar 3 . Nilai Akurasi dari pengujian silang

Dari hasil penelitian di atas, dapat kita lihat bahwa class negative memiliki recall sebesar 99,06% dan presisi sebesar 97,54%, sementara class positive memiliki recall sebesar 96,00% dan presisi sebesar 98,46%.

KESIMPULAN

Berdasarkan penelitian dan percobaan yang dilakukan terhadap penggunaan pohon keputusan dengan menerapkan algoritma C4.5, ditemukan rangkuman sebagai berikut:

a. Metode decision tree atau algoritma C4.5 bisa bermanfaat untuk digunakan memprediksi penyakit diabetes mellitus menjadi kelas negatif dan positif. Metode ini memiliki tingkat akurasi sebesar 91,82% dalam memprediksi penyakit diabetes mellitus.
b. Algoritma C4.5 terbukti menjadi solusi yang efektif untuk memprediksi penyakit diabetes mellitus. Selain itu, algoritma ini menunjukkan fleksibilitas dan efektivitas yang luar biasa dalam proses pengklasifikasian. Metode ini membantu dokter mengklasifikasikan diabetes mellitus berdasarkan gejala pasien.

REFERENSI

A, H. R. (2020). Penentuan Rekomendasi Petrainingan Pengembangan Diri Bagi Pegawai Negeri Sipil Menggunakan Algoritma C4.5 Dengan Principal Component Analysis. *J. TEKNO KOMPAK*.

Chan, P. P. (2018). Pengembangan Aplikasi Perhitungan Prediksi Stock Motor Menggunakan Algoritma C 4.5 Sebagai Bagian dari Sistem Pengambilan Keputusan (Studi Kasus di Saudara Motor). *INOVTEK Polbeng -Seri Inform*.

Deny Jollyta, W. R. (2020). *Konsep Data Mining Dan Penerapan.* Deepublish.

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan AlgoritmaDecision TreeC4.5. *Jurnal Sistem Komputer dan Kecerdasan Buatan*.

Haikal, M., Kusuma, R. S., Nauvanda, S. E., & Safitri, M. (2022). Perancangan User Interface dan User Experience Pada Web MB Tours and Travel Bekasi. *JIKA (Jurnal Informatika)*, *6*(3), 271–278. https://doi.org/10.31000/jika.v6i3.6777

Handayani, P. K. (2020). Penerapan Principal Component Analysis untuk Peningkatan Kinerja Algoritma Decission Tree pada IRIS Dataset. *IJTIS*.

Ikhsan Romli, A. T. (2020). Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5. *Jurnal Sains Komputer & Informatika*.

Muhamad Ichsan Gunawan, D. S. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression. J*urnal Edukasi dan Penelitian Informatika*.

Nasrulloh, A. h. (2021). Implementasi Algoritma Decision Tree untuk Klasifikasi Produk LARIS . *Jurnal Ilmiah Ilmu Komputer*.

Nurlina. (2019). Politeknik Kesehatan Makassar Jurnal Media Keperawatan: Politeknik Kesehatan Makassar. *Jurnal Media Keperawatan*.

Praba, A. D., Safitri, M., & Faridi, F. (2021). Implementasi Datatables Server-Side untuk Mempercepat Load Halaman pada Aplikasi E-Commerce. *JIKA (Jurnal Informatika)*, *5*(2), 139–144. https://doi.org/10.31000/jika.v5i2.4339

Ramadhan, M. A. (2019). Patient Empowerment Dan Self-Management Pada Pasien Diabetes Mellitus Tipe 2. *Jurnal Ilmiah Kesehatan Sandi Husada*.

Rokhman, K. A. (2021). Perbandingan Metode Support Vector Machine dan Decision Tree untuk Analisi Sentimen Review Komentar pada Aplikasi Transportasi Online. Jurnal of Information System Management.

Sanni Ucha Putri, E. I. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5 . *KESATRIA: Jurnal Penerapan Sistem Informasi*.

Siska Febriani, H. S. (2021). ANALISIS DATA HASIL DIAGNOSA UNTUK KLASIFIKASI GANGGUAN KEPRIBADIAN MENGGUNAKAN ALGORITMA C4.5 . *Jurnal Teknologi dan Sistem Informasi*.

Srisulistiowati, D. B. (2021). Sistem Informasi Prediksi Penjualan Alat Tulis Kantor dengan Metode FP-GROWTH. *Jurnal Sistem Informatika Universita Suryadarma*.

Zai, C. (2022). Implementasi Data Mining Sebagai Pengolahan Data. *Portal Data*.