Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

ANALISIS RISIKO KESEHATAN TERHADAP PENYAKIT DIABETES MENGGUNAKAN ALGORITMA C4.5

[1]Wida Prima Mustika*, [2]Reggy Helva Rezal, [3]Andi Diah Kuswanto, [4]Febri Windianti Ekatami, [5]Indah Gita Cahyani, [6]Ilham Agustian,[7]Syahputra Ibrahim

Program Studi Sistem Informasi, Fakultas Teknik dan Informatika,
Universitas Bina Sarana Informatika
Jl. Kramat Raya No.98, Senen, Jakarta Pusat, 10450, Indonesia

Correpondent Author: wida.wpm@bsi.ac.id

e-mail:[1]wida.wpm@bsi.ac.id, [2]reggyhelva28@gmail.com , [3]andi.ahk@bsi.ac.id, [4]pebriiwindiantii@gmail.com, [5]indahgitaindah@gmail.com, [6]ilhamagus59620@gmail.com, [7]syahputraibrahim6@gmail.com

Abstrak

India merupakan negara dengan jumlah penderita diabetes tertinggi kedua di dunia setelah Tiongkok dengan lebih dari 77 juta kasus pada tahun 2021. Penelitian ini bertujuan untuk menganalisis risiko terkena diabetes dengan menggunakan algoritma C4.5 berdasarkan data kesehatan masyarakat di India. Dataset yang digunakan dalam penelitian ini diperoleh dari *Kaggle* dengan judul "*Pima Indians Diabetes Dataset*" yang diperbarui pada September 2025. Pengolahan data dilakukan menggunakan aplikasi RapidMiner dengan metode data mining berdasarkan algoritma C4.5 untuk mengklasifikasikan data berdasarkan variabel seperti usia, jumlah kehamilan, indeks massa tubuh (BMI), tekanan darah, kadar gula darah, dan kadar insulin. Hasil penelitian menunjukkan bahwa algoritma C4.5 mampu mengidentifikasi pola risiko diabetes dengan tingkat akurasi yang cukup baik sehingga dapat digunakan sebagai alat bantu dalam pengambilan keputusan untuk mendeteksi risiko diabetes secara dini.

Kata Kunci: Diabetes, India, Data Mining, Algoritma C4.5, RapidMiner, Klasifikasi

Abstract

India is the country with the second highest number of diabetes cases in the world after Tiongkok with more than 77 millions cases in 2021. This study aims to analyze the risk of developing diabetes using the C4.5 algorithm based on public health data in India. The dataset used in this study was obtained from Kaggle under the title "Pima Indians Diabetes Dataset" which was updated in September 2025. Data processing was performed using the RapidMiner application with a data mining method based on the C4.5 algorithm to classify data based on variables such as age, number of pregnancies, body mass index (BMI), blood pressure, blood sugar levels, and insulin levels. The results of the study show that the C4.5 algorithm is capable of identifying diabetes risk patterns with a fairly good level of accuracy so that it can be used as a tool to assist in decision making for early detection of diabetes risk.

Keywords: *Diabetes, India, Data Mining, C4.5 Algorithm, RapidMiner, Classification.*

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

1 Pendahuluan

Diabetes melitus merupakan salah satu penyakit tidak menular dengan tingkat penyebaran yang terus meningkat secara global. *Hiperglikemia* adalah tanda dari sekelompok kondisi yang berbeda yang dikenal sebagai diabetes melitus (DM). *World Health Organization (WHO)* (2016) menjelaskan bahwa gejala DM yang tampaknya tidak berbahaya, seperti banyak minum, nafsu makan yang meningkat, frekuensi berkemih yang berlebihan, kelelahan, dan kesemutan, seringkali tidak terdiagnosis [1]. Selain gejala, diabetes melitus juga dapat diklasifikasikan menjadi dua kategori yaitu diabetes melitus tipe 1 dan diabetes melitus tipe 2. Selain itu, standar diagnosis biokimia termasuk pengukuran glukosa darah oral dan penggunaan hemoglobin A1c (HbA1c) [2]. Tingginya prevalensi diabetes dapat dilihat pada beberapa negara dengan jumlah penderita signifikan.

Analisis menunjukkan bahwa India menjadi negara kedua setelah Tiongkok yang memiliki jumlah penderita diabetes absolut tertinggi, dengan lebih dari 77 juta orang di India mengalami diabetes pada tahun 2021 [3]. Salah satu penyebab terjadinya diabetes adalah karena makanan yang dikonsumsi kurang sehat, sehingga menyebabkan peningkatan berat badan. Selain itu, kesadaran masyarakat dinilai masih rendah dan alat untuk mendeteksi diabetes masih belum optimal. Kondisi ini menunjukkan bahwa sebagian masyarakat kurang menyadari bahaya dari penyakit diabetes [4].

Permasalahan tersebut menunjukkan bahwa diperlukan suatu komputasi yang mampu membantu mendeteksi diabetes secara dini. Teknik data mining memungkinkan analisis data kesehatan dalam jumlah besar untuk menemukan pola tersembunyi yang dapat digunakan untuk memprediksi penyakit, memungkinkan sistem informasi untuk melakukan proses deteksi penyakit diabetes melitus pada manusia [5]. Algoritma C4.5 adalah salah satu algoritma data mining yang bersifat prediktif dan digunakan untuk klasifikasi atau pengelompokan [6]. Dengan demikian, penggunaan algoritma C4.5 dalam penelitian sebelumnya menunjukkan potensi yang signifikan dalam memprediksi penyakit diabetes berdasarkan faktor-faktor risiko seperti usia, kehamilan, indeks massa tubuh (BMI), tekanan darah, kadar gula darah, dan insulin, sehingga relevan untuk diterapkan dalam penelitian [7].

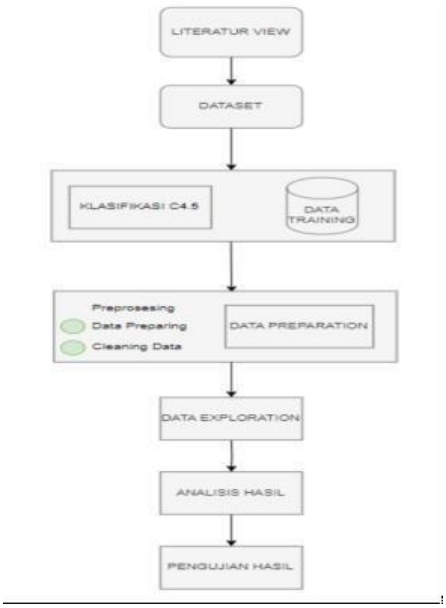
2 Metode Penelitian *(or Research Method)*

Penelitian ini menggunakan pendekatan data mining dengan menerapkan algoritma C4.5 untuk menganalisis risiko diabetes berdasarkan data kesehatan masyarakat di India. Model prediksi algoritma C4.5 menggunakan struktur hirarki berupa pohon. Setiap pohon terdiri dari cabang-cabang dan setiap cabang mewakili atribut tertentu yang harus dipenuhi agar bisa berpindah ke cabang berikutnya hingga mencapai akhir [8]. Metode ini dipilih karena mampu mengklasifikasikan data secara efisien dan membentuk pola keputusan yang mudah dipahami. Metode ini dipilih karena mampu mengklasifikasikan data dengan baik dan menghasilkan pola keputusan yang mudah dipahami. Penelitian berkonsentrasi pada bagaimana data dikumpulkan, diproses, dan dianalisis untuk menentukan pola yang dapat digunakan untuk memprediksi risiko terkena diabetes. Penelitian ini mencakup pertimbangan literatur, pengumpulan set data pelatihan, persiapan data, eksplorasi data, analisis hasil, dan pengujian hasil. Untuk memastikan bahwa model yang dibangun mampu memberikan analisis yang tepat dan dapat diandalkan sebagai dasar untuk deteksi dini diabetes, setiap tahap berperan penting

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183



Gambar 2.1. Tahap Penelitian.

Secara umum, langkah-langkah penelitian ini terlihat pada Gambar 2.1 Alur Tahap Penelitian, yang menampilkan proses penelitian mulai dari tahap awal hingga penilaian model.

A. Literatur View

Literatur view mencakup pengumpulan informasi dari berbagai sumber seperti jurnal dan artikel penelitian yang relevan. Dataset yang digunakan berasal dari *Kaggle* dengan judul *"Pima Indians Diabetes Dataset"*. Penelitian ini meliputi dasar teori mengenai diabetes melitus beserta faktor-faktor risikonya, konsep data mining, serta penerapan algoritma C4.5 dalam klasifikasi data kesehatan.

B. Penentuan Data Training Set

Dataset *Kaggle "Pima Indians Diabetes"* digunakan untuk mengumpulkan data kesehatan masyarakat India dengan sembilan fitur gula darah, tekanan darah, insulin, kehamilan, usia, riwayat keluarga, dan BMI. Data dipilih dengan hati-hati agar representatif dan berkualitas, sehingga dapat digunakan dengan baik dalam proses pelatihan dan pengujian model untuk mempelajari pola risiko diabetes dengan benar.

C. Data Preparation

Sebelum digunakan algoritma C4.5, tahap persiapan data dilakukan untuk memastikan kualitas dataset. *"Dataset Diabetes Pima Indians"* diproses dengan mengubah format, menormalisasi data numerik, menghapus nilai kosong, dan mengencoding atribut kategorikal. Tujuan dari langkah ini adalah untuk memastikan bahwa dataset menjadi teratur, terorganisir, dan siap digunakan untuk membangun model klasifikasi diabetes.

D. Data Exploration

Dataset *"Pima Indians Diabetes Dataset"* memiliki sembilan metrik kesehatan, yaitu usia, riwayat keluarga, BMI, gula darah, tekanan darah, insulin, dan kehamilan. Untuk memahami distribusi

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

dan variasi masing-masing atribut, tahap eksplorasi data dilakukan dengan menggunakan analisis statistik deskriptif, yang mencakup mean, nilai minimum, maksimum, dan standar deviasi. Proses ini membantu menemukan pola, hubungan antar variabel, dan outlier, serta menemukan atribut yang berdampak pada risiko diabetes sebelum memulai proses klasifikasi dengan algoritma C4.5.

$$\overline{x} = \frac{\sum x_i}{n}$$

Keterangan :

$\bar{X}$ = Nilai rata-rata

Xi = Data ke-i

n = Jumlah data

Rumus ini digunakan untuk menghitung rata-rata dari setiap atribut seperti kadar glukosa darah, tekanan darah, dan indeks massa tubuh (BMI).

Selanjutnya, untuk mengukur tingkat penyebaran data terhadap rata-ratanya digunakan rumus standar deviasi (SD) sebagai berikut:

$$\sqrt{\frac{\sum (X_i - \overline{X})^2}{n - 1}}$$

Keterangan:

SD = Standar Deviasi

Xi = Nilai data ke-i

$\bar{X}$ = Nilai rata-rata dari seluruh data

n = Jumlah data

$\Sigma (Xi - \bar{X})^2$ = Jumlah kuadrat selisih antara setiap nilai data dengan nilai rata-ratanya

Selain itu, untuk mengetahui hubungan antar atribut yang berpotensi mempengaruhi risiko diabetes, dilakukan analisis korelasi menggunakan rumus korelasi pearson berikut:

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

$$r_{xy} = \frac{n\sum x_i\, y_i - (\sum x_i)(\sum y_i)}{\sqrt{\left(n\sum x_i^2 - (\sum x_i)^2\right)\left(n\sum y_i^2 - (\sum y_i)^2\right)}}$$

Keterangan :

$\mathbf{r_{xy}}$ = Korelasi antara x dan y

n = Banyaknya sampel

$\mathbf{x_i}$ = Nilai x ke-i

$\mathbf{y_i}$ = Nilai y ke-i

Melalui tahap eksplorasi data, peneliti dapat mengetahui bahwa atribut seperti kadar glukosa darah, BMI, dan usia memiliki korelasi yang cukup tinggi terhadap status diabetes. Informasi ini menjadi dasar penting dalam menentukan atribut yang digunakan pada tahap pembentukan model klasifikasi dengan algoritma C4.5.

E. Analisis Hasil

Setelah pra-pemrosesan selesai, algoritma C4.5 digunakan untuk membuat pohon keputusan yang memetakan pola risiko diabetes. Aturan dari pohon keputusan dianalisis untuk mengetahui faktor risiko utama diabetes. Dengan menggunakan confusion matrix, metrik ketepatan, ketepatan, dan recall digunakan untuk mengevaluasi kinerja model. Hasilnya menunjukkan bahwa algoritma C4.5 berhasil dalam mengklasifikasikan risiko diabetes dan menunjukkan komponen yang paling penting.

F. Pengujian Hasil

Pengujian hasil adalah tahap akhir penelitian untuk mengetahui seberapa baik kemampuan model untuk memproses data baru. Dengan menggunakan data uji yang berbeda, model C4.5 diuji dan dinilai menggunakan *confusion matrix*, *ROC curve*, dan nilai *AUC*. Nilai AUC yang mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang sangat baik untuk membedakan data yang berisiko diabetes dari data yang tidak berisiko.

3 Hasil dan Pembahasan (or Results and Analysis)

Pengolahan data dilakukan melalui beberapa tahap, mulai dari memasukkan dataset, membersihkan data, eksplorasi, hingga membangun model klasifikasi menggunakan algoritma C4.5 di RapidMiner. RapidMiner digunakan karena antarmuka visualnya mempermudah pembuatan model tanpa menulis kode.

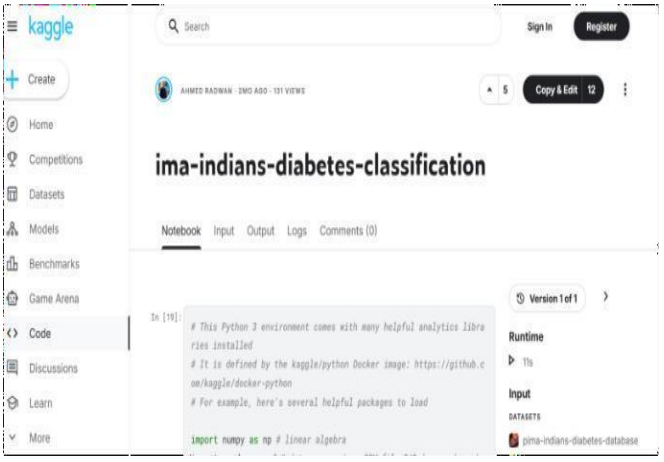
1. Pemilihan dataset

Di mana penelitian menggunakan dataset *"Pima Indians Diabetes Classification"* dari *Kaggle* yang memuat informasi kesehatan perempuan India Pima, seperti kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, faktor genetik, usia, dan status diabetes. Dataset ditampilkan dalam bentuk tabel sebelum diproses lebih lanjut.

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183



Gambar 1. Dataset Kaggle "Pima Indians diabetes Classification"
Sumber Dataset: https://www.kaggle.com/code/ahmedrdwan/ima-indians-diabetes-classification

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	Blood Pressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
2	6	148	72	35	0	33,6	0,627	50	Positive
3	1	85	66	29	0	26,6	0,351	31	Negative
4	8	183	64	0	0	23,3	0,672	32	Positive
5	1	89	66	23	94	28,1	0,167	21	Negative
6	0	137	40	35	168	43,1	2.288	33	Positive
7	5	116	74	0	0	25,6	0,201	30	Negative
8	3	78	50	32	88	31	0,248	26	Positive
9	10	115	0	0	0	35,3	0,134	29	Negative
10	2	197	70	45	543	30,5	0,158	53	Positive
11	8	125	96	0	0	0	0,232	54	Positive
12	4	110	92	0	0	37,6	0,191	30	Negative
13	10	168	74	0	0	38	0,537	34	Positive
14	10	139	80	0	0	27,1	1.441	57	Negative
15	1	189	60	23	846	30,1	0,398	59	Positive
16	5	166	72	19	175	25,8	0,587	51	Positive
17	7	100	0	0	0	30	0,484	32	Positive
18	0	118	84	47	230	45,8	0,551	31	Positive
19	7	107	74	0	0	29,6	0,254	31	Positive
20	1	103	30	38	83	43,3	0,183	33	Negative
21	1	115	70	30	96	34,6	0,529	32	Positive
22	3	126	88	41	235	39,3	0,704	27	Negative
23	8	99	84	0	0	35,4	0,388	50	Negative
24	7	196	90	0	0	39,8	0,451	41	Positive
25	9	119	80	35	0	29	0,263	29	Positive
26	11	143	94	33	146	36,6	0,254	51	Positive
27	10	125	70	26	115	31,1	0,205	41	Positive
28	7	147	76	0	0	39,4	0,257	43	Positive

Gambar 2. Atribut Dataset Diabetes

2. Data Preparation (Persiapan Data)

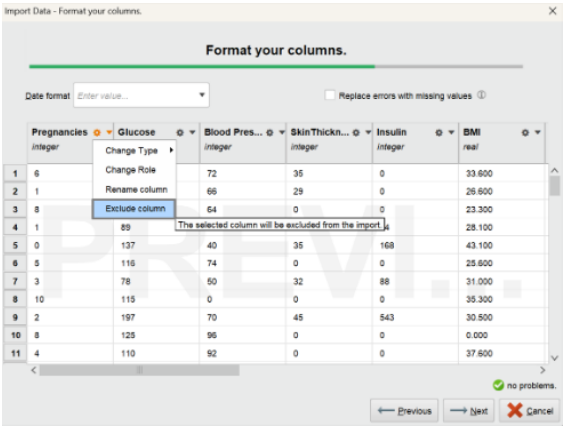
Dataset diimpor ke dalam aplikasi *RapidMiner* untuk memulai proses persiapan data. Selanjutnya, dataset yang diunduh dimasukkan ke lingkungan kerja *RapidMiner* untuk pengaturan dan analisis lebih lanjut. Proses ini memastikan data mentah dapat dikenali oleh sistem dan tersimpan dalam repository sehingga siap digunakan pada tahap berikutnya.

Setelah dataset di import, data dibersihkan dengan memilih variabel yang relevan dengan kolom exclude untuk menghilangkan atribut yang tidak diperlukan. Dalam penelitian ini, variabel yang digunakan adalah karakteristik yang langsung terkait dengan faktor risiko diabetes. Ini termasuk kehamilan, gula darah, tekanan darah, BMI, keturunan, usia, dan hasil sebagai variabel target.

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

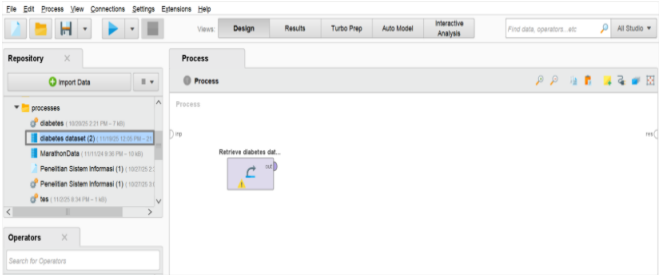
Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183



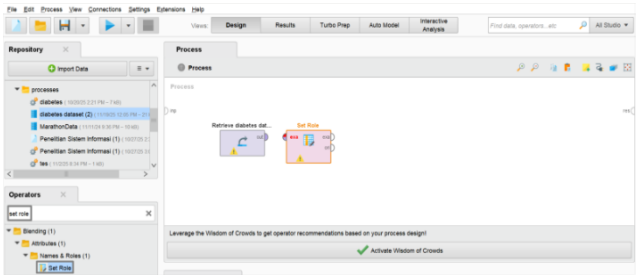
Gambar 3. Proses Penentuan Atribut

Dataset yang telah diseleksi kemudian diambil kembali dari repository menggunakan operator *Retrieve*. Operator ini berfungsi memanggil dataset sehingga dapat diproses dalam *canvas* analisis *RapidMiner*. Seluruh atribut yang telah ditentukan sebelumnya akan terbawa ke dalam proses ini dan menjadi input bagi tahapan analisis berikutnya.



Gambar 4. Operator Retrieve dari Database yang Sudah Diimpor

Penetapan peran atribut dilakukan menggunakan operator *Set Role*. Fitur ini digunakan untuk mendefinisikan fungsi masing-masing atribut di dalam proses pemodelan. Pada penelitian ini, atribut Outcome ditetapkan sebagai label, yaitu variabel yang akan diprediksi oleh algoritma klasifikasi, sedangkan atribut lain dipertahankan sebagai prediktor. Penetapan peran atribut ini penting agar *RapidMiner* dapat mengenali mana fitur masukan dan mana target yang harus diprediksi.

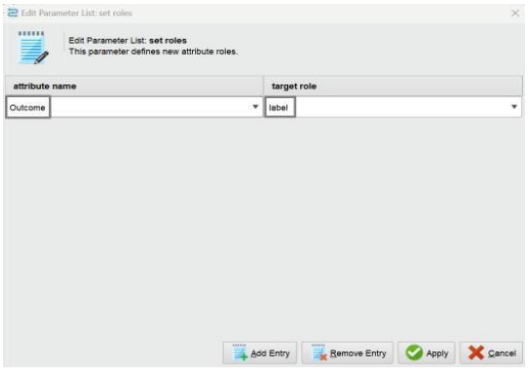


Gambar 5. Proses Operator Set Role

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

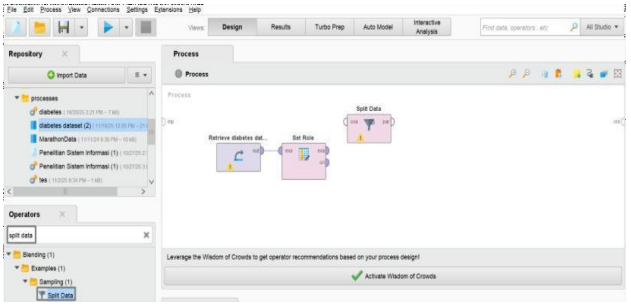
DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183



Gambar 6. Penentuan Atribut Sebagai Label

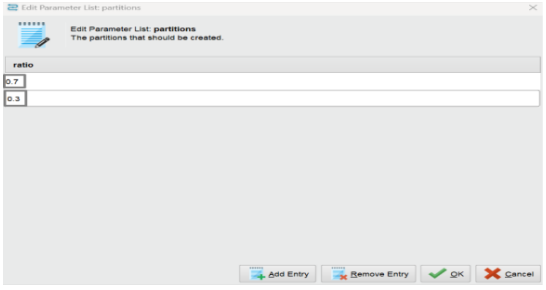
3. Data Transformation (Transformasi Data)

Pada tahap transformasi data, operator *Split Data* digunakan untuk membagi dataset menjadi beberapa bagian sesuai kebutuhan proses analisis. Operator ini memastikan bahwa setiap bagian data yang dihasilkan memiliki struktur atribut yang sama dengan dataset awal, namun berisi data yang berbeda. Dengan demikian, proses ini tidak mengubah nilai-nilai atribut, tetapi hanya memisahkan bentuk penyajian data menjadi dua bagian yang siap digunakan dalam pelatihan model dan pengujian kinerja model.



Gambar 7. Proses Operator Split Data

Pembagian dilakukan dengan menetapkan persentase tertentu, contohnya 70% digunakan sebagai data latih dan 30% sebagai data uji. Data latih berfungsi untuk membangun dan menyesuaikan parameter model, sedangkan data uji digunakan untuk mengevaluasi kemampuan model dalam memprediksi data yang belum pernah dilihat sebelumnya. Dengan transformasi ini, sistem mendapatkan dua kelompok data yang sudah siap digunakan dalam tahap berikutnya, yaitu untuk pelatihan model klasifikasi dan pengujian akurasi model agar performa model menjadi optimal.



Gambar 8. Pembagian Data Latih dan Data Uji

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

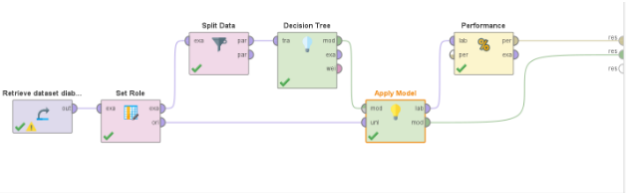
4. Modeling (Pembangunan Model)

Langkah pemodelan merupakan fase penting di mana model prediktif dibuat berdasarkan data yang sudah disiapkan. *Decision Tree* akan belajar mengenai pola dan hubungan dalam data latih, sehingga mampu membuat model yang dapat digunakan untuk memprediksi atau mengklasifikasikan data baru atau data uji secara akurat. Pada tahap modeling, *Apply Model* digunakan untuk menerapkan model yang sudah dibuat sebelumnya ke dataset baru, biasanya data uji. Setelah model diterapkan, sistem menghasilkan dataset yang diperkaya dengan atribut prediksi, yaitu nilai kelas yang diprediksi oleh model. Operator ini tidak mengubah struktur dataset asli, hanya menambahkan kolom untuk hasil prediksi dan, dalam beberapa kasus, nilai keyakinan atau probabilitas.

Gambar 9. Proses Operator Apply Model

5. Evaluation (Evaluasi Model)

Pada tahap evaluasi, operator kinerja digunakan untuk menghitung metrik seperti akurasi, precision, recall, dan F-measure yang didasarkan pada hasil prediksi *Apply Model*. Dengan menggunakan metrik ini, peneliti dapat menilai ketepatan model, kemampuan untuk mengenali kelas positif, dan keseimbangan antara akurasi dan recall. Dengan demikian, tahap evaluasi memberikan gambaran lengkap tentang efektivitas model dan menentukan apakah model masih layak digunakan atau memerlukan perbaikan.



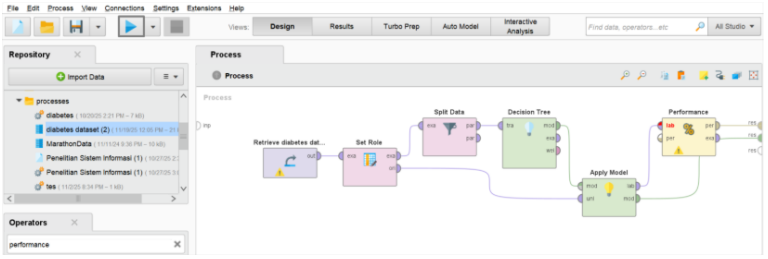
Gambar 10. Evaluasi Model Performance

Proses terakhir dalam penelitian adalah menjalankan seluruh tahapan analisis di *RapidMiner*. Untuk menjalankan proses tersebut. Proses ini dilakukan secara bertahap untuk mendapatkan model klasifikasi yang mampu memprediksi faktor-faktor risiko penyakit diabetes. Hasil akhir dari analisis ini berupa nilai-nilai evaluasi seperti *accuracy*, *precision*, dan *recall*, yang digunakan sebagai dasar dalam menilai kualitas model yang telah dibuat.

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183



Gambar 11. Menjalankan Proses Analisis

Variabel data penelitian yang digunakan pada penelitian ini dapat dilihat pada tabel 1 yakni sebagai berikut.

No	Atribut	Value
1.	Kehamilan	Rendah, Sedang, Resiko Tinggi.
2.	Glukosa	Pra diabetes, Normal, Tinggi.
3.	Tekanan Darah	Rendah, Normal, Resiko Tinggi.
4.	BMI	Normal, Kelebihan Berat Badan, Obesitas.
5.	Keturunan	Rendah, Sedang, Agak tinggi, Tinggi
6.	Usia	Remaja, Dewasa, Lansia.
7.	Outcome	Positive dan Negatif.

Tabel 1. Variabel Data Penelitian

Dari 200 data, 140 data latih (70%) dan 60 data uji (30%). Menentukan fitur yang paling penting untuk memisahkan data berdasarkan kelasnya adalah langkah pertama dalam pembangunan pohon keputusan C4.5. Dalam proses ini, perhitungan Gain Informasi digunakan; ini adalah perhitungan yang menghitung seberapa besar pengurangan ketidakpastian (entropy) yang terjadi setelah data dibagi menurut atribut tertentu. Informasi Gain memiliki nilai yang lebih tinggi, yang berarti atribut tersebut lebih penting dan akan menjadi titik awal dalam pohon keputusan.

Berikut ini rumus untuk menghitung nilai Information Gain:

$$Gain(S,A) = Enthropy(S) - \sum_{i=1}^{n} - \frac{|Si|}{|S|} * Enthopy(Si)$$

Keterangan:

- **S** = himpunan data (dataset keseluruhan)
- **A** = atribut yang diuji
- **n** = jumlah partisi data berdasarkan nilai atribut A
- $\mathbf{S_i}$ = subset dari S yang memiliki nilai atribut A ke-i

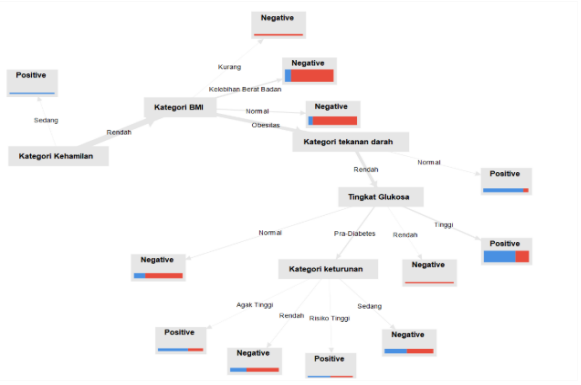
DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

- $|S_i|$ = jumlah data dalam subset ke-i
- $|S|$ = jumlah total data dalam himpunan S
- **Entropy(S)** = tingkat ketidakpastian atau impurity dari keseluruhan dataset
- **Entropy(S_i)** = tingkat ketidakpastian pada subset data ke-i

Setelah perhitungan pendapatan informasi, fitur dengan nilai tertinggi digunakan sebagai dasar dari keputusan. Metode ini menghasilkan struktur pohon yang menunjukkan hubungan antara fitur yang menentukan risiko diabetes. Setiap cabang pohon menunjukkan kondisi yang menghasilkan klasifikasi diabetes positif atau negatif.



Gambar 12. Hasil Klasifikasi Menggunakan *Decision Tree*

Setelah struktur pohon keputusan terbentuk, langkah selanjutnya adalah melakukan pengujian lebih lanjut untuk mengetahui seberapa efektif model tersebut dalam melakukan klasifikasi. Maka, tahap berikutnya adalah menghitung tingkat akurasi model sebagai ukuran kemampuan algoritma C4.5 dalam klasifikasi. Pohon keputusan dibangun dari atribut dengan nilai gain tertinggi. Implementasi dan pengujian Dataset dibuat dengan menggunakan *RapidMiner*.

Berikut merupakan analisis hasil pengujian dari 60 data uji dengan 140 data latih dengan setiap data memiliki 6 variabel dataset gejala dan 1 variabel outcome.

accuracy: 76.00%

	true Positive	true Negative	class precision
pred. Positive	49	22	69.01%
pred. Negative	26	103	79.84%
class recall	65.33%	82.40%	

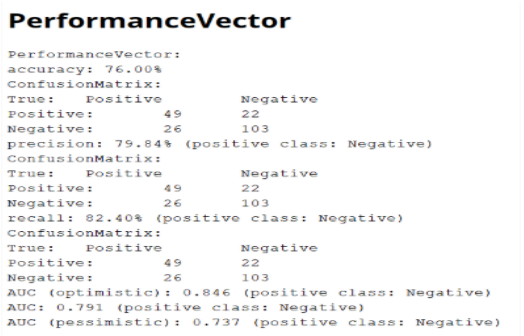
Gambar 13. Hasil Akurasi

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

Jurnal Manajemen Informatika Jayakarta
Volume 6, Nomor 2, Apri 2026: halaman 89-101
P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

Setelah mencapai akurasi 76%, model menunjukkan bahwa sebagian besar prediksi sudah tepat, meskipun potensi kesalahan sekitar 24% masih ada karena kemiripan pola antar kelas. Setelah mendapatkan akurasi ini, analisis dilanjutkan dengan menggunakan matriks kekacauan untuk mengevaluasi kinerja model pada setiap kelas, baik prediksi yang benar maupun salah. Kemampuan model untuk membedakan data positif dan negatif digambarkan dengan matriks kekacauan



Gambar 14. Hasil *Performance Vector*

Dengan menggunakan algoritma C4.5 untuk mengklasifikasikan kondisi kesehatan berdasarkan karakteristik seperti glukosa, tekanan darah, BMI, keturunan, dan usia, 200 data dibagi menjadi 140 data latih dan 60 data uji. Hasil pengujian menunjukkan akurasi 76,00%, *precision* 79,84%, dan *recall* 82,40%. Hasil menunjukkan bahwa C4.5 cukup efektif dalam memprediksi kemungkinan terkena diabetes. Selain memberikan hasil klasifikasi, pohon keputusan yang dibuat juga menunjukkan bagaimana usia, kadar glukosa, dan indeks massa tubuh (BMI) mempengaruhi kemungkinan terkena diabetes.

4 Kesimpulan (or Conclusion)

Metode ini mampu memprediksi risiko diabetes dengan baik, menurut penelitian yang memanfaatkan algoritma C4.5 pada dataset Diabetes Pima Indians. Model menunjukkan akurasi 76%, *presisi* 79,84%, *recall* 82,40%, dan *AUC* 0,846 yang baik setelah tahap persiapan data, eksplorasi, pembentukan model, dan pengujian. Secara keseluruhan, algoritma C4.5 berfungsi dengan baik untuk mendeteksi risiko diabetes sejak dini dan dapat digunakan sebagai dasar untuk sistem atau aplikasi yang mendukung keputusan kesehatan.

Referensi (Reference)

[1] E. Handayani, N. Maesaroh, N. Azizah, and A. H. Mukaromah, “Sosialisasi Penyakit Diabetes Melitus Pada Kelompok Dasawisma Sendangguwo Kelurahan Gemah Kecamatan Pedurungan Kota Semarang,” *Pros. Semin. Nas. Unimus*, vol. 4, pp. 2565–2572, 2021.

[2] K. R. Widiasari, I. Made, K. Wijaya, and P. A. Suputra, “Tatalaksana Diabetes Melitus Tipe II,” *Ganesha Med. J.*, vol. 1, no. 2, pp. 114–120, 2021, [Online].

[3] W. P. Sakul and H. Andriani, “Prevalensi Global Diabetes Mellitus : Tinjauan Sistematis Negara-Negara Dengan Beban Tinggi Dan Strategi,” vol. 9, no. 1, pp. 5150–5158, 2025.

[4] A. Fauzi, “Aplikasi Sistem Pakar Dengan Metode Naive Bayes untuk Mendeteksi Penyakit Diabetes Expert System Application Using the Naive Bayes Method to Detect Diabetes,” vol. 15, no. April, pp. 17–31, 2025.

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta

P-ISSN: 2746-5985
e-ISSN: 2797-0930

DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183

[5] Chandra, David, and Hendri, "Implementasi Algoritma C4.5 Untuk Deteksi Dini Penyakit Diabetes Mellitus Pada Manusia," *Bull. Comput. Sci. Res.*, vol. 4, no. 2, pp. 241–248, 2024, doi: 10.47065/bulletincsr.v4i2.337.

[6] R. A. Siallagan and Fitriyani, "Prediksi Penyakit Diabetes Mellitus Menggunakan Algoritma C4.5," *J. Responsif Ris. Sains dan Inform.*, vol. 3, no. 1, pp. 44–52, 2021, doi: 10.51977/jti.v3i1.407.

[7] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, "Klasifikasi penyakit diabetes mellitus berdasarkan faktor-faktor penyebab diabetes menggunakan algoritma c4.5," vol. 07, pp. 1265–1270, 2022.

[8] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4 . 5," 2020.


DOI: https://doi.org/10.52362/jmijayakarta.v6i2.2183
This work is licensed under a Creative Commons Attribution 4.0 International License.
http://journal.stmikjayakarta.ac.id/index.php/JMIJayakarta