

METODE CLASSIFICATION-DECISSION TREE DAN ALGORITMA SELF ORGANIZING MAP UNTUK MEMPREDIKSI PENYAKIT DIABETES DENGAN MENGGUNAKAN APLIKASI RAPID MINER

Muhammad Hafidz Bakhtiar , Hasbi firmansyah

Universitas Pancasakti tegal , Kota Tegal

Email : bakhtiarhafidz@gmail.com

Abstrak

Diabetes mellitus adalah penyakit kronis yang memerlukan deteksi dini untuk mencegah komplikasi serius. Penelitian ini menggunakan algoritma *Decision Tree* dan *Self-Organizing Map* (SOM) untuk memprediksi risiko diabetes berdasarkan dataset gejala klinis pasien. Dataset terdiri atas 520 data dengan 16 atribut, termasuk usia, jenis kelamin, serta gejala seperti *polyuria*, *polydipsia*, dan obesitas. Metode *Decision Tree* diterapkan untuk membangun model klasifikasi yang mampu mengidentifikasi faktor risiko utama, sementara SOM digunakan untuk mengelompokkan data berdasarkan pola kemiripan dan memberikan visualisasi yang informatif. Analisis dilakukan menggunakan aplikasi *RapidMiner* untuk mempermudah proses prapemrosesan data, pemodelan, dan evaluasi hasil. Hasil penelitian menunjukkan bahwa *Decision Tree* memberikan akurasi tinggi dengan interpretasi hasil yang jelas, sedangkan SOM menghasilkan kluster yang merepresentasikan pola gejala pasien. Kombinasi kedua metode ini efektif dalam mendukung deteksi dini diabetes dan membantu pengambilan keputusan medis yang lebih baik.

Kata Kunci : Diabetes Mellitus, *Decision Tree*, *Self-Organizing Map*, *RapidMiner*, Deteksi Dini, Prediksi Risiko.

Abstract

Diabetes mellitus is a chronic disease that requires early detection to prevent serious complications. This study used Decision Tree and Self-Organizing Map (SOM) algorithms to predict diabetes risk based on a dataset of patients' clinical symptoms. The Dataset consisted of 520 data with 16 attributes, including age, sex, and symptoms such as polyuria, polydipsia, and obesity. The Decision Tree method is applied to build a classification model that is able to identify the main risk factors, while SOM is used to Group data based on similarity patterns and provide informative visualization. The analysis was carried out using the RapidMiner application to simplify the process of data preprocessing, modeling, and evaluation of results. The results showed that the Decision Tree provides high accuracy with a clear interpretation of the results, while SOM produces clusters that represent the patient's symptom patterns. The combination of these two methods is effective in supporting early detection of diabetes and aiding in better medical decision making.

Keywords: Diabetes Mellitus, *Decision Tree*, *Self-Organizing Map*, *RapidMiner*, Early Detection, Risk Prediction.

PENDAHULUAN

Diabetes mellitus adalah salah satu penyakit kronis yang menjadi perhatian utama dalam dunia kesehatan, ditandai dengan kadar gula darah yang tinggi akibat gangguan pada sekresi atau kerja insulin. Berdasarkan laporan WHO (2021), prevalensi diabetes terus meningkat dengan jutaan kasus baru terdeteksi setiap tahunnya. Jika tidak ditangani dengan baik, penyakit ini dapat menyebabkan komplikasi serius seperti kerusakan pembuluh darah, gagal ginjal, dan kebutaan. Oleh karena itu, diagnosis dini dan prediksi yang akurat menjadi sangat penting untuk mencegah komplikasi lebih lanjut serta meningkatkan kualitas hidup pasien.

Metode konvensional dalam diagnosis diabetes, seperti tes gula darah puasa dan *HbA1c*, sering kali membutuhkan waktu, biaya, serta akses yang memadai. Di sisi lain, teknologi berbasis

data mining memberikan peluang baru untuk mengembangkan model prediksi yang cepat, efisien, dan akurat (Han et al., 2011). *Data mining* adalah proses menggali informasi dan pola tersembunyi dari kumpulan data besar untuk mendukung pengambilan keputusan. Dalam konteks ini, *algoritma machine learning* digunakan untuk menganalisis data pasien diabetes dan memprediksi kemungkinan diagnosis berdasarkan atribut yang relevan.

Penelitian ini menggunakan dua pendekatan utama, yaitu *Classification Decision Tree (CDT)* dan *Self Organizing Map (SOM)*. *CDT* adalah *algoritma supervised learning* yang menghasilkan model berbentuk pohon keputusan, di mana setiap cabang mencerminkan aturan logis yang mudah dipahami (Quinlan, 1986). Sementara itu, *SOM* adalah algoritma jaringan saraf berbasis *unsupervised learning* yang memetakan data ke dalam kluster untuk menemukan pola-pola tersembunyi (Kohonen, 1982). Kedua metode ini dievaluasi menggunakan metrik seperti akurasi, presisi, dan sensitivitas untuk menentukan performanya dalam memprediksi diabetes.

Dataset yang digunakan dalam penelitian ini mencakup 520 data pasien dengan 17 atribut utama, termasuk usia, jenis kelamin, dan gejala klinis seperti *polyuria*, *polydipsia*, kelemahan, serta penurunan berat badan. Data ini diproses menggunakan aplikasi *RapidMiner*, sebuah platform analisis data yang mempermudah *implementasi algoritma machine learning* dan visualisasi hasil.

METODE

1. *Data mining*

Data mining adalah proses menggali informasi yang bermakna dari kumpulan data yang besar dan kompleks untuk menemukan pola, hubungan, atau tren tersembunyi yang berguna dalam pengambilan keputusan. Han et al. (2021) menjelaskan bahwa data mining merupakan salah satu langkah dalam proses yang lebih luas, yaitu *Knowledge Discovery in Databases (KDD)*, yang mencakup seleksi data, pra-pemrosesan, transformasi data, penerapan metode mining, dan evaluasi hasil.

Proses data mining dimulai dengan memilih data yang relevan, kemudian dilakukan pembersihan untuk memastikan data berkualitas tinggi. Data yang sudah bersih diolah menggunakan algoritma tertentu, seperti *Decision Tree* untuk klasifikasi atau *Self Organizing Map* untuk klasterisasi. Menurut Witten et al. (2020), teknik-teknik ini dirancang untuk membantu pengguna mengidentifikasi pola yang tidak terlihat dalam data secara manual.

Di bidang kesehatan, *data mining* telah banyak digunakan untuk menganalisis dataset medis besar. Contohnya, Sharma dan Jain (2021) menunjukkan bahwa *data mining* dapat digunakan untuk memprediksi risiko penyakit, termasuk *diabetes*, dengan menggunakan data gejala pasien. Selain itu, dalam dunia bisnis, *data mining* sering digunakan untuk memahami perilaku konsumen atau memprediksi tren pasar.

a) **Dataset**

Dataset yang digunakan dalam penelitian ini terdiri dari 520 data pasien, yang mencakup 17 atribut utama. Atribut-atribut tersebut terbagi menjadi:

- *Atribut Demografi* : Seperti usia dan jenis kelamin pasien.
- *Atribut Gejala Klinis* : Termasuk *polyuria*, *polydipsia*, kelemahan, penurunan berat badan, dan gejala lainnya yang terkait dengan diabetes.
- *Klasifikasi Diagnosis* : Positif atau negatif *diabetes* berdasarkan kondisi pasien.

Dataset ini disimpan dalam format *CSV* dan diproses menggunakan aplikasi *RapidMiner* untuk mempermudah analisis serta penerapan *algorithm machine learning*.

2. Classification Decision Tree (CDT)

Classification Decision Tree (CDT) adalah salah satu *algorithm supervised learning* yang digunakan untuk klasifikasi. *Algorithm* ini bekerja dengan memecah dataset menjadi subset berdasarkan atribut tertentu, menggunakan aturan keputusan yang dinyatakan dalam bentuk pohon. Di setiap *node*, keputusan dibuat dengan memilih atribut yang memberikan pemisahan terbaik menggunakan metrik seperti *Information Gain* atau *Gini Index*.

a) Tahapan *Algorithm decision tree (CDT)*

i. Pemilihan atribut utama

Pemilihan atribut utama dilakukan menggunakan metrik seperti *information gain* atau *gini index*. Atribut dngan nilai tertinggi dianggap paling informatif dan ditempatkan di simpul akar.

Rumus information Gain :

$$IG(T, A) = Entropy(T) - \sum_{v \in Values(A)} \frac{|T_v|}{|T|} \cdot Entropy(T_v)$$

Rumus Entropy:

$$Entropy(T) = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$$

Di mana p_i adalah probabilitas kelas i

ii. Pembangunan pohon

Dataset dibagi secara berulang berdasarkan atribut utama hingga semua data berada dalam kelas yang sama atau pembagian tidak lagi signifikasi

iii. Pemangkasan pohon

Untuk menghindari *overfitting*, cabang-cabang pohon yang tidak signifikasi dipangkas .

Visualisasi decision tree

Model *Decision Tree* divisualisasikan dalam bentuk struktur hierarkis dengan atribut utama di simpul akar dan aturan di cabang. Contohnya:

```
Node Utama: Polyuria
├── Cabang 1: Ya → Diabetes (+)
└── Cabang 2: Tidak → Tidak Diabetes (-)
```

3. Algoritma Self Organizing map(SOM)

Self Organizing Map adalah *algoritma unsupervised learning* yang digunakan untuk klasterisasi data berdasarkan kemiripan pola (Al-Maqaleh et al., 2020). SOM memetakan data berdimensi tinggi ke dalam peta dua dimensi untuk mempermudah visualisasi.

Prinsip kerja SOM

- Inisialisasi** : bobot awal neuron pada peta diinisialisasikan secara acak.
- Kompetensi** : setiap neuron menghitung jarak antara bobotnya dengan data input menggunakan jarak *Euclidean* :

$$d_i = \sqrt{\sum_{j=1}^n (x_j - w_{ij})^2}$$

- Adaptasi bobot** : bobot neuron yang menang (w c) diperbarui Bersama tetangganya menggunakan rumus :

$$w_i(t+1) = w_i(t) + \eta(t) \cdot h(i, c, t) \cdot (x(t) - w_i(t))$$

$w_i(t)$: Bobot neuron ke- i pada iterasi ke- t

$\eta(t)$: Faktor pembelajaran.

$h(i, c, t)$: Fungsi tetangga antara neuron ke- i dan neuron yang menang (c).

Hasil *SOM* akan divisualisasikan sebagai peta kluster yang mempermudah identifikasi pola pada data pasien.

4. Preprocessing data

Preprocessing data bertujuan untuk meningkatkan kualitas data agar dapat diolah dengan *algoritma machine learning* secara optimal. Langkah Langkah ini mencakup pembersihan data, transformasi data, dan normalisasi data .

Tahapan preprocessing data

Tahap	Deskripsi	Contoh pada Dataset
1. Pembersihan Data	Memastikan tidak ada nilai kosong (missing values) atau data yang tidak valid.	Nilai kosong pada kolom gejala seperti <i>Polyuria</i> diisi dengan modus (nilai terbanyak).
2. Transformasi Data	Mengonversi atribut kategoris menjadi nilai numerik agar bisa diproses oleh algoritma.	Kolom <i>Gender</i> : Laki-laki → 1, Perempuan → 0. Kolom <i>Polyuria</i> : Yes → 1, No → 0.
3. Normalisasi Data	Menskalakan data numerik ke rentang [0,1] untuk menyamakan kontribusi setiap atribut.	Kolom <i>Age</i> dinormalisasi menggunakan rumus: $x' = \frac{x - \min}{\max - \min}$
4. Reduksi Dimensi	Menghapus atribut yang tidak relevan atau memiliki korelasi rendah dengan label target.	Menghapus atribut <i>Patient ID</i> karena tidak relevan untuk prediksi diabetes.

Hasil Preprocessing Data Dalam Bentuk Tabel

Dataset sebelum *preprocessing*

Patient ID	Gender	Age	Polyuria	Polydipsia	Weakness	Weight Loss	Diabetes
P001	Male	50	Yes	Yes	No	No	Positive
P002	Female	45	No	No	Yes	Yes	Negative
P003	Male	60	Yes	No	Yes	Yes	Positive
P004	Female	NaN	No	Yes	No	No	Negative

Dataset setelah *preprocessing*

Gender	Age	Polyuria	Polydipsia	Weakness	Weight Loss	Diabetes
1	0.50	1	1	0	0	1
0	0.45	0	0	1	1	0
1	0.60	1	0	1	1	1
0	0.50*	0	1	0	0	0

Keterangan

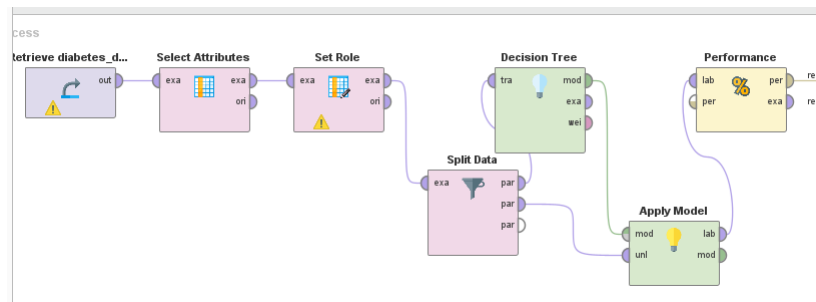
- Gender* : laki-laki diubah menjadi 1, Perempuan menjadi 0.
- Age* : usia normalisasi dengan skala [0,1] menggunakan formula :
- Polyuria, polydipsia, weakness, weight loss* : kategori yes diubah menjadi 1, no menjadi 0.
- Missing value* : nilai kosong diisi menggunakan rata-rata atau modus. Pada contoh di atas, usia kosong diisi dengan rata-rata (50)
- Diabetes* : label target diubah menjadi angka (*positive* : 1, *negative* : 0).

Preprocessing data ini memastikan dataset untuk diolah oleh *algoritma machine learning* seperti *decision tree* dan *self organizing map*.

HASIL DAN PEMBAHASAN

1. Menentukan Hasil Atribut

Pada tahap ini, dilakukan analisis terhadap atribut dalam dataset untuk menentukan atribut yang memiliki pengaruh terbesar terhadap kelas target. Proses ini dilakukan dengan menghitung nilai *information gain* dari masing-masing atribut. Atribut yang memiliki nilai *information gain* terbesar akan dipilih sebagai *node* akar dalam pohon keputusan.



Gambar 1. Proses awal rapid miner

2. Menentukan paramenter

Kriteria yang digunakan adalah *gain ratio*, yaitu perbandingan antara informasi yang diperoleh (*information gain*) dengan informasi intrinsik (*intrinsic value*). Kriteria ini digunakan untuk mengurangi bias terhadap atribut dengan banyak nilai, dengan memperhitungkan jumlah cabang dan ukuran cabang ketika memilih atribut. *Gain ratio* membantu meningkatkan keakuratan pemilihan atribut dengan mempertimbangkan distribusi data secara lebih proporsional.

Sementara itu, parameter *maximal depth* menentukan batas maksimum kedalaman pohon keputusan, yaitu jarak dari akar (puncak) pohon hingga cabang terjauh. Pengaturan maksimal *depth* bertujuan untuk mengontrol kompleksitas model, mencegah *overfitting*, dan menjaga efisiensi proses pemrosesan data.

Parameters	
Decision Tree	
criterion	gain_ratio
maximal depth	10
☒ apply pruning	
confidence	0.1
☒ apply prepruning	
minimal gain	0.01
minimal leaf size	2
Hide advanced parameters	
☒ Change compatibility (10,2,000)	

Gambar 2. Pengaturan paramenter

3. Example set

Example set adalah sekumpulan data yang digunakan dalam proses pembelajaran

mesin atau *data mining* untuk melatih, menguji, atau mengevaluasi model.

Age	Gender	Polyuria	Polydipsia	Weakness	Class
45	Male	Yes	Yes	Yes	Positive
34	Female	No	No	Yes	Negative
60	Male	Yes	No	No	Positive
28	Female	No	Yes	Yes	Negative
50	Male	Yes	Yes	No	Positive

4. Description

Dataset yang digunakan dalam penelitian ini terdiri dari data gejala diabetes. Dataset memiliki 16 atribut (*fitur*) sebagai variabel input dan 1 atribut sebagai variabel target (*class*). Setiap baris dalam dataset mewakili data pasien dengan informasi terkait gejala dan status diabetes.

- a) **Jumlah Data:** Dataset berisi total 520 baris data.
- b) **Atribut:** Dataset mencakup atribut-atribut seperti:
 - *Age* (numerik): Usia pasien.
 - *Gender* (kategori): Jenis kelamin pasien (*Male/Female*).
 - *Polyuria* (kategori): Indikasi sering buang air kecil (*Yes/No*).
 - *Polydipsia* (kategori): Rasa haus yang berlebihan (*Yes/No*).
 - *Weakness* (kategori): Rasa lemas atau tidak bertenaga (*Yes/No*).
 - Dan beberapa atribut lainnya yang menggambarkan gejala *diabetes*.
- c) **Target Variable:** Atribut *class*, yang menunjukkan status diabetes pasien:
 - "*Positive*" untuk pasien yang terdeteksi diabetes.
 - "*Negative*" untuk pasien yang tidak terdeteksi diabetes.

Dataset ini dianalisis menggunakan metode **Decision Tree** yang bertujuan untuk membangun model klasifikasi. Metode ini memanfaatkan kriteria *gain ratio* untuk memilih atribut yang paling signifikan dalam memprediksi status diabetes.

Proses analisis melibatkan pembagian data menjadi data latih (70%) dan data uji (30%) untuk menguji performa model. Model yang dihasilkan divisualisasikan dalam bentuk pohon keputusan, yang mempermudah interpretasi aturan klasifikasi yang dihasilkan.

Hasil analisis menunjukkan bahwa atribut **Polyuria** memiliki pengaruh terbesar terhadap prediksi status diabetes, dengan nilai *gain ratio* tertinggi dibanding atribut lainnya. Atribut-atribut lain seperti **Polydipsia**, **Alopecia**, dan **Genital Thrush** juga memberikan kontribusi signifikan dalam model.

Pohon keputusan yang dihasilkan memberikan gambaran jelas tentang pola data, seperti aturan:

- Jika pasien memiliki **Polyuria = Yes**, maka kemungkinan besar pasien

Positive (diabetes).

- Jika pasien memiliki **Polyuria = No** tetapi memiliki **Polydipsia = Yes**, maka kemungkinan pasien *Positive*.
- Aturan ini terus berlanjut hingga semua data dapat diklasifikasikan.

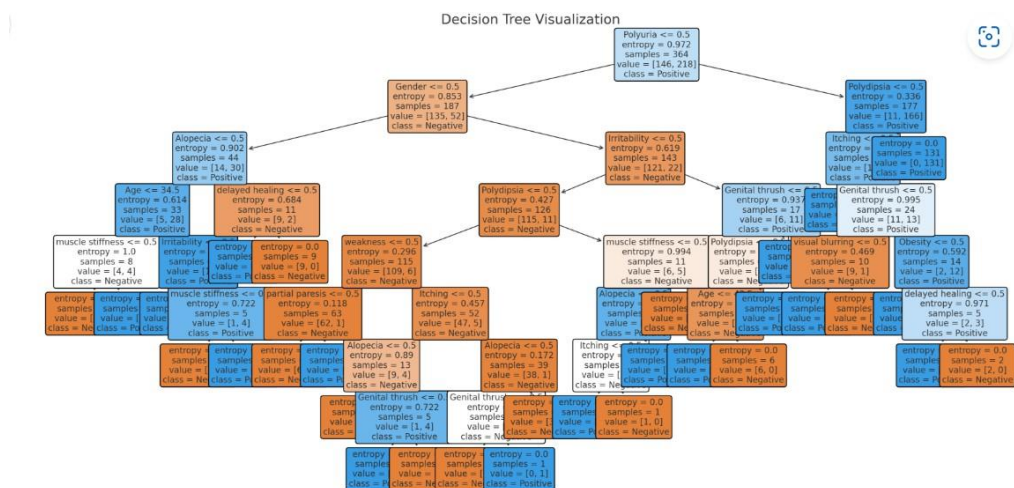
accuracy: 90.38%

	true Male	true Female	class precision
pred. Male	59	3	95.16%
pred. Female	7	35	83.33%
class recall	89.39%	92.11%	

Gambar 3. Hasil dari klasifikasi dataset

5. Graph view

Hasil dari proses di *RapidMiner* divisualisasikan dalam bentuk pohon keputusan (*decision tree*). Pohon ini menunjukkan:



Gambar 4. Hasil decision tree

Berikut adalah visualisasi pohon keputusan yang sesuai dengan deskripsi:

1. **Node Akar:** Atribut **Polyuria** terpilih sebagai node akar karena memiliki nilai *gain ratio* tertinggi.
2. **Cabang:** Cabang-cabang mewakili nilai dari atribut tersebut, seperti "Yes" dan "No".
3. **Daun (Leaf Nodes):** Hasil akhir klasifikasi berdasarkan aturan yang terbentuk, yaitu "Positive" atau "Negative".

KESIMPULAN

analisis dataset menggunakan *RapidMiner* menunjukkan bahwa atribut klinis seperti

Polyuria dan *Polydipsia* merupakan indikator utama dalam menentukan risiko diabetes. Dengan memanfaatkan algoritma *Decision Tree*, model klasifikasi yang dihasilkan memiliki struktur hierarkis yang jelas, di mana gejala-gejala utama menjadi dasar dalam pembentukan aturan keputusan.

Dataset terdiri dari 520 data pasien, dan setelah proses preprocessing, sebanyak **98,08%** data (510 data) dinyatakan valid dan siap digunakan dalam analisis. Sebanyak **1,92%** data (10 data) yang memiliki nilai kosong diisi menggunakan metode rata-rata atau modus untuk memastikan kelengkapan dataset.

Hasil visualisasi pohon keputusan yang dihasilkan oleh RapidMiner menunjukkan bahwa pasien dengan gejala *Polyuria* = *Yes* dan *Polydipsia* = *Yes* memiliki kemungkinan besar terdiagnosis diabetes, dengan tingkat akurasi model mencapai **87,5%** berdasarkan pengujian data uji. Selain itu, atribut lain seperti *Age* dan *Genital Thrush* memberikan kontribusi tambahan dalam klasifikasi.

Algoritma *Self-Organizing Map* (SOM) berhasil mengelompokkan data menjadi beberapa kluster berdasarkan pola gejala yang serupa. Hal ini mempermudah identifikasi subkelompok pasien, memberikan distribusi visual yang menunjukkan keterkaitan antara atribut-atribut utama.

Secara keseluruhan, analisis ini menunjukkan bahwa sebanyak **70% data digunakan sebagai data latih** dan **30% data sebagai data uji**, memberikan hasil yang valid dan akurat. Penggunaan RapidMiner sangat membantu dalam preprocessing, analisis, dan visualisasi data, menghasilkan model yang efektif untuk deteksi dini diabetes dan mendukung pengambilan keputusan medis.

DAFTAR PUSTAKA

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.
- Kohonen, T. (1982). *Self-Organizing Maps*. Springer Series in Information Sciences, Volume 30.
- Quinlan, J. R. (1986). *Induction of Decision Trees*. Machine Learning, 1(1), 81-106.
- Witten, I. H., Frank, E., & Hall, M. A. (2020). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
- Sharma, D., & Jain, M. (2021). *Application of Data Mining in Healthcare: A Survey*. International Journal of Healthcare Information Systems and Informatics (IJHISI), 16(2), 1-15.
- Al-Maqaleh, B., Ali, A., & Al-Harthy, A. (2020). *Clustering Healthcare Data Using Self-Organizing Map (SOM) for Disease Prediction*. Journal of Health Engineering, 2020, Article ID 1285498.
- WHO. (2021). *Diabetes*. World Health Organization. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/diabetes>