Penerapan Algoritma Decision Tree untuk Klasifikasi Diagnosa Pasien

Zaenol Fatah[1] , Risma Nur Kholishah[2]
[1,2,] universitas Ibrahimy, Situbondo
Email: rismanurkholisah5@gmail.com

Abstrak

Penelitian ini memanfaatkan algoritma Decision Tree sebagai metode analisis tipe C4.5 untuk mengklasifikasikan kondisi kesehatan pasien berdasarkan data diagnosis dari Sri Clinics di Kota Batam. Data dianalisis menggunakan RapidMiner dengan atribut seperti suhu tubuh, tekanan darah, detak jantung, jenis kelamin, usia, dan gejala utama. Hasil penelitian menunjukkan bahwa suhu tubuh merupakan faktor paling dominan dalam klasifikasi, di mana pasien dengan suhu ≤ 37,05°C dikategorikan sehat, sedangkan suhu > 37,05°C menunjukkan penyakit ringan atau berat. Model Decision Tree menghasilkan akurasi 96,67%, membuktikan bahwa metode ini efektif dan akurat sebagai alat bantu keputusan medis untuk menunjang proses penentuan penyakit dengan efisiensi dan ketepatan tinggi

Kata Kunci: Decision Tree C4.5, Data Mining, Diagnosis Pasien, Klasifikasi

Abstrak

This study utilized the Decision Tree algorithm, a C4.5 type analysis method, to classify patient health conditions based on diagnostic data from Sri Clinics in Batam City. The data was analyzed using RapidMiner with attributes such as body temperature, blood pressure, heart rate, gender, age, and primary symptoms. The results showed that body temperature was the most dominant factor in the classification, with patients with a temperature below 37.05°C categorized as healthy, while temperatures above 37.05°C indicated mild or severe disease. The Decision Tree model achieved 96.67% accuracy, demonstrating its effectiveness and accuracy as a medical decision-making tool to support the process of determining disease with high efficiency and accuracy.

Keywords: *Decision Tree C4.5, Data Mining, Patient Diagnosis, Classification*

PENDAHULUAN

Kemajuan teknologi informasi dan analisis data telah membawa perubahan signifikan dalam dunia kesehatan, khususnya pada proses diagnosis penyakit dan pengambilan keputusan klinis. Pemanfaatan *machine learning* menjadi salah satu pendekatan populer untuk mengolah data medis dalam jumlah besar dan menghasilkan pola prediktif yang dapat membantu tenaga medis. Salah satu algoritma yang sering digunakan dalam konteks ini adalah *Decision Tree* (pohon keputusan), karena kemampuannya dalam menghasilkan model klasifikasi yang mudah dipahami dan diinterpretasikan(Purnomo, H.; Pambudi, R. E.; Irawan, R., 2025)

Algoritma *Decision Tree* bekerja dengan membagi dataset menjadi beberapa cabang berdasarkan atribut tertentu hingga mencapai keputusan akhir berbentuk kelas target. Keunggulan utama algoritma ini adalah interpretabilitasnya yang tinggi serta kemampuannya dalam menangani data kategorikal maupun numerik. Dalam bidang medis, algoritma ini telah diterapkan pada berbagai kasus diagnosis seperti penyakit jantung, kanker payudara, pneumonia, dan penyakit ginjal kronis.

Penelitian oleh Hendri et al.menunjukkan bahwa *Decision Tree*

mampu mengklasifikasikan data rekam medis pasien dengan tingkat akurasi mencapai 91,55%. Studi lain oleh PubMed mengembangkan model prediksi mortalitas pasien COVID-19 dengan pendekatan CART, C5.0, dan CHAID, yang mampu mengidentifikasi faktor-faktor risiko utama seperti usia, kadar CRP, dan fungsi ginjal. Hasil serupa ditemukan oleh(Mienye & Jere, 2024) yang menerapkan *Decision Tree* untuk memprediksi tingkat mortalitas pasien COVID-19 di Indonesia dengan hasil akurasi yang kompetitif dibandingkan metode lainnya.

Sejumlah penelitian juga membandingkan kinerja *Decision Tree* dengan algoritma klasifikasi lain seperti *Naïve Bayes*, *K-Nearest Neighbor (KNN)*, dan *Random Forest*. Hasilnya menunjukkan bahwa *Decision Tree* masih menjadi pilihan utama dalam konteks interpretabilitas hasil dan efisiensi komputasi, meskipun terkadang perlu dioptimalkan dengan metode seperti *Particle Swarm Optimization (PSO)* untuk meningkatkan akurasi (Sari et al., 2025)

Meskipun demikian, penerapan algoritma *Decision Tree* di Indonesia masih terbatas pada skala penelitian dan belum banyak diintegrasikan ke dalam sistem pendukung keputusan medis berbasis digital secara luas. Padahal, penerapan model ini berpotensi membantu dokter dan rumah sakit dalam proses diagnosis dini, deteksi penyakit kronis, hingga prediksi risiko kematian pasien secara lebih akurat dan efisien (Jainudin & Abdullah, 2025).

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk mengimplementasikan dan menganalisis kinerja algoritma *Decision Tree* dalam klasifikasi diagnosis pasien. Tujuan utama penelitian ini adalah untuk menilai sejauh mana algoritma *Decision Tree* dapat memberikan hasil prediksi yang akurat berdasarkan data medis, serta mengkaji potensi penerapannya sebagai bagian dari sistem pendukung keputusan di bidang kesehatan digital.

METODE

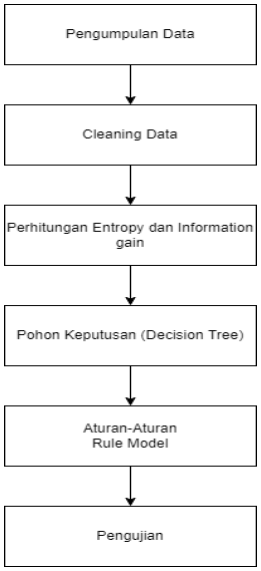
Data Mining merupakan proses menggabungkan teknik statistik, matematika, kecerdasan buatan (AI), dan pembelajaran mesin untuk mengekstrak informasi yang relevan dan berguna dari data dalam jumlah besar(Srirahayu & Pribadie, 2023). Data mining didefisikan sebagai suatu metode pengambilan pengetahuan yang dibangun dengan menggunakan data-data historis yang sudah terkumpul sekian lama(Cahyono, E. T.(2020)).

Berdasarkan fungsinya, data mining dapat dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, pengklusteran1 dan asosiasi (Wahidin, A. J., & Syukrilla, W.A. (2023)). Penggunaan algoritma data mining dilakukan untuk menggali data yang agar memudahkan identifikasi informasi(Baharuddin et al., 2019) Namun semakin besar data yang diolah maka semakin besar juga waktu pemrosesannya.

Algotirma C4.5 merupakan algoritma yang umum digunakan untuk pengambilan keputusan. C4.5akan mencari solusi permasalahan dengan menjadikan kriteria sebagai node yang saling berhubungan membentuk seperti struktur pohon(Khotimah & Istiawan, 2018).

Model prediksi dengan algoritma C4.5 mengacu pada proses pengambilan keputusan yang direpresentasikan dalam bentuk struktur hierarkis berupa pohon keputusan. Pada pohon tersebut terdapat cabang-cabang yang masing-masing menggambarkan atribut tertentu yang harus dipenuhi sebelum menuju cabang berikutnya hingga mencapai simpul akhir. Dalam algoritma C4.5, data diolah dalam bentuk tabel yang terdiri atas sejumlah atribut dan record sebagai dasar pembentukan pohon keputusan.Atribut

digunakan sebagai parameter yang dibuat sebagai kriteria dalam pembuatan pohon, dan record sebagai penentun nilai keputusan pada pohon(Anugerah & Laut, 2020).



Gambar 1.Tahap Penelitian

Seluruh rangkaian proses dalam penelitian ini disajikan pada diagram alur pada Gambar 1. Tahapan penelitian dimulai dari proses pengumpulan data, dilanjutkan dengan pembersihan data, perhitungan nilai entropy dan information gain, pembentukan model pohon keputusan (Decision Tree), penyusunan rule atau aturan keputusan, serta diakhiri dengan tahap validasi dan pengujian model.

1. **Pengumpulan data**
   Penelitian ini menggunakan data diagnosis pasien dari platfrom Kaggle. Dataset terdiri dari atribut suhu tubuh, tekanan darah, detak jantung, jenis kelamin, usia, dan gejala utama. Data ini digunakan untuk klasifikasi diagnosis, yaitu kategori sehat, penyakit ringan, dan penyakit berat.
2. **Cleansing data**
   Pembersihan data (*data cleansing*) untuk menghapus nilai kosong (*missing value*) dan data ekstrem (*outlier*) guna memastikan integritas dataset sebelum proses pelatihan model(Rehman & Belhaouari, 2021)
3. **Perhitungan Entropy dan Information Gain**
   Pada tahap ini dilakukan perhitungan menggunakan algoritma C4.5 untuk menentukan atribut terbaik yang akan dijadikan node pada pohon keputusan.

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

   (a) Perhitungan Entropy
   Entropy digunakan untuk mengukur tingkat ketidakpastian dan keberagaman pada setiap atribut. Rumus entropy adalah:

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

   (b) Perhitungan Information Gain
   Gain adalah nilai yang digunakan untuk mengukur efektivitas atribut dalam mengklasifikasikan data. Atribut dengan nilai gain tertinggi akan menjadi akar pohon.
   (c) Pengulangan Perhitungan
   Proses perhitungan entropy dan gain diulang secara rekursif pada setiap node hingga seluruh data terpartisi sempurna atau tidak ada lagi atribut yang dapat dipartisi.
4. **Pohon Keputusan (Decision Tree)**
   Hasil dari tahap perhitungan entropy dan gain menghasilkan struktur pohon keputusan. Pohon keputusan ini memuat hubungan antara atribut seperti suhu tubuh, detak jantung, tekanan darah, dan gejala utama untuk menentukan kategori diagnosis pasien. Struktur pohon inilah yang menjadi dasar terbentuknya rule model.
5. **Pengujian**
   Tahap akhir adalah evaluasi performa algoritma Decision Tree C4.5. Pengujian dilakukan menggunakan RapidMiner dengan pembagian dataset

sebesar 75% untuk data training dan 25% untuk data testing.

Evaluasi model dilakukan menggunakan confusion matrix untuk memperoleh nilai:

- Accuracy
- Precision
- Recall

Nilai-nilai tersebut digunakan untuk menilai kinerja model klasifikasi dalam memprediksi tingkat keparahan kesehatan pasien.

Pengolahan data

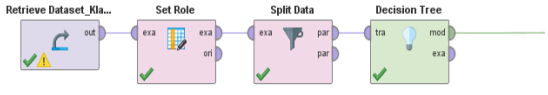
Dataset yang digunakan dalam penelitian ini berasal dari platform Kaggle, sehingga termasuk dalam kategori dataset publik karena dapat diakses secara bebas tanpa izin khusus dan dapat digunakan untuk keperluan analisis maupun penelitian. Dataset ini terdiri dari 24 record dengan 8 atribut, yaitu usia, jenis kelamin, suhu tubuh, detak jantung, tekanan darah, gejala utama, riwayat penyakit, dan diagnosis sebagai label target. Seluruh data tersebut kemudian diimpor ke RapidMiner untuk diproses lebih lanjut pada tahap pengolahan data dan pemodelan klasifikasi.

Tabel 1. Data Set

Usia	Jenis Kelamin	Suhu Tubuh	Detak Jantung	Tekanan Darah	Gejala Utama	Riwayat Penyakit	Diagnosis
14	Perempuan	37	100	Rendah	Nyeri Dada	Tidak Ada	Sehat
59	Perempuan	38,9	101	Normal	Lemas	Tidak Ada	Penyakit Ringan
52	Laki-laki	37,2	114	Rendah	Lemas	Tidak Ada	Penyakit Ringan
56	Laki-laki	38,3	103	Tinggi	Lemas	Tidak Ada	Penyakit Ringan
71	Laki-laki	38,3	126	Tinggi	Nyeri Dada	Ada	Penyakit Berat
64	Laki-laki	38,5	126	Rendah	Lemas	Tidak Ada	Penyakit Ringan
48	Laki-laki	38	116	Rendah	Batuk	Ada	Penyakit Ringan
46	Perempuan	38	67	Normal	Lemas	Ada	Penyakit Ringan
75	Laki-laki	37,1	87	Rendah	Batuk	Tidak Ada	Penyakit Ringan
36	Perempuan	37,2	116	Rendah	Sesak Nafas	Tidak Ada	Sehat
47	Perempuan	36,3	81	Tinggi	Sesak Nafas	Ada	Sehat
83	Perempuan	37,5	112	Normal	Demam	Ada	Penyakit Ringan
13	Perempuan	40,3	61	Tinggi	Lemas	Ada	Penyakit Ringan
49	Perempuan	37,4	117	Tinggi	Sesak Nafas	Ada	Sehat
74	Laki-laki	36,4	114	Rendah	Nyeri Dada	Tidak Ada	Sehat
61	Laki-laki	36,1	122	Normal	Demam	Ada	Sehat
41	Perempuan	38,4	124	Normal	Batuk	Ada	Penyakit Ringan
21	Perempuan	39,4	108	Tinggi	Lemas	Ada	Penyakit Ringan
17	Perempuan	39,3	140	Tinggi	Demam	Tidak Ada	Penyakit Berat
6	Laki-laki	39,4	116	Rendah	Lemas	Tidak Ada	Penyakit Ringan
38	Laki-laki	39,7	68	Rendah	Batuk	Ada	Penyakit Ringan
73	Perempuan	36,1	137	Rendah	Nyeri Dada	Tidak Ada	Sehat
4	Perempuan	38,9	87	Rendah	Batuk	Ada	Penyakit Ringan
9	Perempuan	37,6	116	Normal	Sesak Nafas	Ada	Sehat

Hasil dan Pembahasan

Pengujian sistem dilakukan dengan memanfaatkan perangkat lunak rapidMiner. Alur prosesnya, sebagaimana ditampilkan pada gambar 3, melibatkan sejumlah operator seperti retrieve data, set role, split data, dan decision tree.

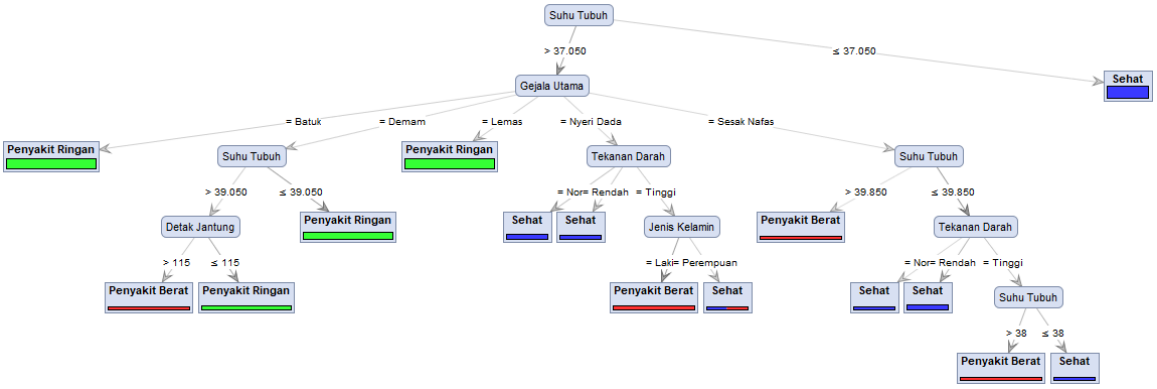


Gambar 3. RapidMiner

Gambar tersebut menampilkan

1. **Retrieve Data**
   Mengambil dataset dari sumber yang telah disiapkan. Dataset berisi atribut fitur dan satu label sebagai target klasifikasi untuk proses pemodelan.
2. **Set Role**
   Menetapkan peran setiap atribut, di mana satu kolom dijadikan **label**, sedangkan atribut lainnya digunakan sebagai variabel input bagi algoritma.
3. **Split Data**
   Membagi dataset menjadi data pelatihan dan data pengujian, misalnya 70% untuk training dan 30% untuk testing, agar model dapat dievaluasi menggunakan data yang belum pernah dilihat sebelumnya.
4. **Decision Tree**

Membangun model klasifikasi menggunakan algoritma Decision Tree. Algoritma memilih atribut terbaik sebagai dasar pemecahan node dan menghasilkan struktur pohon keputusan untuk memprediksi kelas pada data uji.



Gambar 4. Pohon Keputusan

Pohon keputusan tersebut mengklasifikasikan kondisi pasien berdasarkan suhu tubuh sebagai titik awal; jika suhu ≤ 37,05°C maka pasien dinyatakan sehat, sedangkan jika suhu > 37,05°C penilaian dilanjutkan ke gejala utama berupa batuk, demam, lemas, nyeri dada, atau sesak napas. Batuk dan lemas langsung dikategorikan sebagai penyakit ringan; pada demam, jika suhu > 39,05°C maka detak jantung menentukan hasil (>115 menjadi penyakit berat, ≤115 penyakit ringan), sementara jika suhu ≤ 39,05°C hasilnya tetap penyakit ringan. Untuk nyeri dada, tekanan darah menjadi acuan: tekanan normal/rendah mengarah ke sehat, sedangkan tekanan tinggi masih diperinci dengan jenis kelamin (laki-laki = penyakit berat, perempuan = sehat). Pada sesak napas, jika suhu > 39,85°C maka hasilnya penyakit ringan, namun jika ≤ 39,85°C maka keputusan bergantung pada tekanan darah—normal/rendah menghasilkan sehat, sedangkan tekanan tinggi masih memerlukan pengecekan suhu tambahan (>38°C penyakit berat, ≤38°C sehat). Dengan demikian, klasifikasi akhir dapat berupa sehat, penyakit ringan, atau penyakit berat tergantung kombinasi suhu tubuh, gejala utama, tekanan darah, detak jantung, dan jenis kelamin.

SIMPULAN (PENUTUP)

Berdasarkan hasil studi mengenai penggunaan algoritma Decision Tree (C4.5) dalam menganalisis data diagnosis pasien, sejumlah poin penting dapat disimpulkan sebagai berikut:

1. Algoritma Decision Tree C4.5 mampu mengklasifikasikan kondisi pasien secara akurat berdasarkan atribut-atribut seperti suhu tubuh, gejala utama, tekanan darah, detak jantung, jenis kelamin, dan usia. Model yang dihasilkan dapat menggambarkan pola hubungan antara faktor-faktor tersebut terhadap tingkat keparahan penyakit.
2. Hasil pengujian menunjukkan bahwa akurasi model mencapai 96,67%, yang berarti model memiliki kemampuan yang sangat baik dalam memprediksi kondisi pasien, baik dalam kategori sehat, penyakit ringan, maupun penyakit berat.
3. Atribut suhu tubuh menjadi faktor paling dominan dalam menentukan hasil klasifikasi. Nilai suhu tubuh ≤ 37,05°C umumnya mengindikasikan

kondisi sehat, sedangkan suhu tubuh di atas 37,05°C menunjukkan adanya gejala penyakit dengan tingkat keparahan yang berbeda-beda tergantung pada atribut pendukung lainnya.

4. Faktor gejala utama seperti demam, batuk, lemas, nyeri dada, dan sesak napas turut memengaruhi hasil diagnosis. Kombinasi antara suhu tubuh tinggi dan detak jantung lebih dari 115 bpm sering kali menunjukkan kondisi penyakit berat.
5. Meskipun akurasi model tinggi, analisis menunjukkan adanya kemungkinan overfitting, di mana model terlalu menyesuaikan diri terhadap data pelatihan. Oleh karena itu, pengujian lanjutan dengan dataset yang lebih besar dan beragam diperlukan agar model dapat memberikan hasil prediksi yang lebih stabil dan general.

DAFTAR PUSTAKA

Anugerah, P. T., & Laut, S. (2020). *SISTEM PENDUKUNG KEPUTUSAN MENENTUKAN HASIL BUDIDAYA UDANG VANAME DENGAN METODE ALGORITMA C4 . 5*. *14*(1), 28–39.

Baharuddin, M. M., Hasanuddin, T., & Azis, H. (2019). *ANALISIS PERFORMA METODE K-NEAREST NEIGHBOR UNTUK*. *11*(28), 269–274.

Cahyono, E. T. (n.d.). *DATA MINING : SOLUSI PENGEMBANGAN PENGETAHUAN BERDASARKAN BASIS*.

Jainudin, K., & Abdullah, A. (2025). *Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Metode*. *8*(3), 232–240.

Khotimah, N., & Istiawan, D. (2018). *Perbandingan Algoritma C4 . 5 , Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang*. 41–50.

Mienye, I. D., & Jere, N. (2024). A Survey of Decision Trees : Concepts , Algorithms , and Applications. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2024.3416838

Purnomo, H., Pambudi, R. E., & Irawan, R. (n.d.). *Penerapan Decision Tree untuk Klasifikasi Penyakit Berdasarkan Data Rekam Medis*. *7*(1), 1–8.

Rehman, A., & Belhaouari, S. B. (2021). Unsupervised outlier detection in multidimensional data. *Journal of Big Data*. https://doi.org/10.1186/s40537-021-00469-z

Sari, I. P., Elvitaria, L., & Sari, I. P. (2025). *Classification of mushroom types based on digital image processing using convolutional neural network*. *13*(4), 379–388.

Srirahayu, A., & Pribadie, L. S. (2023). *Review Paper Data Mining Klasifikasi Data Mining*. *14*(April).

Wahidin, A. J., & Syukrilla, W. A. (n.d.). *Data mining*.