

Klasifikasi Penyakit Paru-Paru Menggunakan Algoritma Decision Tree Studi Kasus Pada Data Medis

Ali Ridla^{1,*}, Alviatur Rizqiyah²^{a,b}Program Studi Sistem Informasi, Universitas Ibrahimy

Jl. KHR. Syamsul Arifin No.1-2, Sukorejo, Sumberejo, Kec. Banyuputih, Kabupaten Situbondo, Jawa Timur 68374, Indonesia

¹Eel.riedla@gmail.com, ²alviaturrizqiyah42@gmail.com

*Penulis koresponden

Diterima	Direvisi	Disetujui	Dipublikasikan
9/7/2025	15/8/2025	28/8/2025	15/9/2025

ABSTRACT

Lung diseases are one of the main causes of morbidity and mortality in various countries, hence the need for effective and accurate classification methods to support early diagnosis. The problem statement in this research is how to utilize decision tree algorithms to classify lung diseases based on medical data. The methodology used includes data preprocessing stages, selection of relevant features from patient clinical attributes (age, smoking history, symptoms, and radiology results), as well as building a decision tree model using the C4.5 algorithm. Evaluation is conducted by measuring accuracy, precision, and recall against test data. The test results show that the decision tree model has a fairly high accuracy in distinguishing categories of lung diseases, and also provides a structure that can be easily interpreted by medical personnel. The discussion shows that the transparency and interpretability of decision trees provide advantages in supporting the clinical decision-making process. In conclusion, decision tree algorithms can be a reliable and efficient approach in medical decision support systems, especially for classifying lung diseases based on patient health data.

KEYWORDS

Lung Disease Classification, Decision Tree, Rapid Miner, Chi Square.

ABSTRAK

Penyakit paru-paru menjadi salah satu penyebab utama morbiditas dan mortalitas di berbagai negara, sehingga diperlukan metode klasifikasi yang efektif dan akurat untuk mendukung diagnosa dini. Rumusan masalah dalam penelitian ini adalah bagaimana memanfaatkan algoritma decision tree untuk mengklasifikasikan penyakit paru-paru berdasarkan data medis. Metodologi yang digunakan mencakup tahap praproses data, pemilihan fitur relevan dari atribut klinis pasien (usia, riwayat merokok, gejala, dan hasil radiologi), serta pembangunan model decision tree menggunakan algoritma C4.5. Untuk menilai akurasi, presisi, dan recall data uji, evaluasi dilakukan. Hasil pengujian menunjukkan bahwa model decision tree memiliki akurasi yang cukup tinggi dalam membedakan kategori penyakit paru-paru, serta memberikan struktur yang dapat diinterpretasikan dengan mudah oleh tenaga medis. Pembahasan menunjukkan bahwa transparansi dan interpretabilitas dari decision tree memberikan keunggulan dalam

KATA KUNCI

Klasifikasi Penyakit Paru-Paru, Decision Tree, Rapid Minner, Chi Square.

mendukung proses pengambilan keputusan klinis. Kesimpulannya, algoritma decision tree dapat dijadikan pendekatan yang andal dan efisien dalam sistem pendukung keputusan medis, khususnya untuk klasifikasi penyakit paru-paru berbasis data kesehatan pasien.

This is an open access article under the CC-BY-SA license.



1 PENDAHULUAN

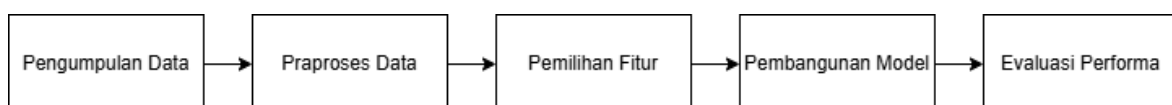
Penyakit paru-paru merupakan salah satu penyebab utama kematian global, khususnya di negara berkembang dengan tingkat polusi tinggi, prevalensi merokok yang tinggi, serta keterbatasan fasilitas diagnostik. Untuk meningkatkan kualitas layanan kesehatan dan menurunkan angka mortalitas, deteksi dini dan klasifikasi penyakit paru-paru menjadi langkah yang sangat krusial. Namun, kesamaan gejala antar jenis penyakit paru-paru kerap menyulitkan diagnosis awal, sehingga dibutuhkan pendekatan sistematis dan berbasis teknologi untuk membantu pengambilan keputusan medis.

Kemajuan teknologi kecerdasan buatan, khususnya *machine learning*, membuka peluang bagi pengembangan sistem klasifikasi berbasis data medis. Salah satu algoritma yang banyak digunakan adalah *decision tree*, karena tidak hanya mampu menghasilkan model klasifikasi dengan akurasi tinggi, tetapi juga memiliki keunggulan dalam hal interpretabilitas—menjadikannya ideal dalam konteks klinis. Penelitian oleh Dinova dan Purnama (2021) melaporkan bahwa penggunaan algoritma *decision tree* mampu menghasilkan akurasi 89,56% dalam klasifikasi penyakit paru-paru berdasarkan data medis pasien[1]. Sementara itu, Salim dan Sugeng (2017) menunjukkan bahwa algoritma *decision tree* J48 mampu mengidentifikasi karakteristik pasien diabetes mellitus dengan akurasi sebesar 88,42% berdasarkan data rekam medis elektronik, serta menghasilkan aturan klasifikasi yang dapat digunakan dalam pengambilan keputusan klinis[2].

Dengan latar belakang ini, penelitian ini bertujuan untuk membangun dan mengevaluasi model klasifikasi penyakit paru-paru dengan menggunakan algoritma decision tree berbasis data rekam medis elektronik. Diharapkan, model ini dapat menjadi bagian dari sistem pendukung keputusan medis yang andal, transparan, serta dapat diterapkan di fasilitas kesehatan primer.

2 METODE PENELITIAN

Metode eksperimen dan pendekatan kuantitatif digunakan dalam penelitian ini untuk membuat model klasifikasi penyakit paru-paru berbasis algoritma decision tree. Tahapan penelitian terdiri dari lima langkah utama, yaitu: pengumpulan data, praproses data, pemilihan fitur, pembangunan model, dan evaluasi performa.



2.1 Pengumpulan Data

Data dalam penelitian ini bersumber dari dataset yang tersedia di platform Kaggle <https://www.kaggle.com/datasets/andot03bsrc/dataset-predic-terkena-penyakit-paruparu>. Dataset ini dirancang untuk mendukung proses prediksi gangguan paru-paru berdasarkan atribut sosial dan klinis pasien, seperti usia, jenis kelamin, kebiasaan merokok, aktivitas olahraga, hingga riwayat komorbiditas.

Dataset ini terdiri dari lebih dari 300 entri pasien, dengan masing-masing entri mewakili satu individu. Perangkat lunak Microsoft Excel 2021 digunakan untuk mengolah format data.csv sebelum diproses menggunakan bahasa pemrograman Python, yang menggunakan pustaka *pandas*, *scikit-learn*, dan *matplotlib* untuk analisis dan pembangunan model.

Nama Atribut	Jenis Data	Deskripsi Singkat
Usia	Numerik	Umur pasien dalam tahun
Jenis_Kelamin	Kategorikal	Laki-laki / Perempuan
Merokok	Biner	Ya / Tidak (status perokok aktif)
Bekerja	Biner	Ya / Tidak (status pekerjaan)
Rumah_Tangga	Biner	Ya / Tidak (status sebagai anggota rumah tangga aktif)
Aktivitas_Begadang	Biner	Ya / Tidak (kebiasaan tidur larut malam)
Aktivitas_Olahraga	Biner	Ya / Tidak (kebiasaan berolahraga rutin)
Asuransi	Biner	Ya / Tidak (kepemilikan asuransi kesehatan)
Penyakit_Bawaan	Kategorikal	Riwayat komorbid seperti diabetes, hipertensi, dll.
Hasil	Label Kelas	Diagnosis akhir (Normal, TB, PPOK, dll.)

Penggunaan dataset publik seperti ini telah didokumentasikan dalam berbagai penelitian sebelumnya. Salah satu contohnya adalah penelitian yang dilakukan oleh Larassati et al. (2022) yang menggunakan data Kaggle untuk membangun sistem prediksi penyakit jantung berbasis *Naive Bayes*, yang memperlihatkan efektivitas pendekatan serupa dalam bidang kesehatan[3].

2.2 Praproses Data

Sebelum dilakukan pelatihan model klasifikasi, dataset melalui tahap praproses untuk memastikan kualitas dan kesiapan data. Adapun tahapan praproses yang dilakukan adalah sebagai berikut:

- Pembersihan Data (Cleaning)**
 Data diperiksa terhadap entri kosong, nilai yang tidak valid, atau duplikat. Misalnya, nilai usia yang tidak masuk akal atau kolom "Penyakit_Bawaan" yang kosong. Proses ini dilakukan menggunakan kombinasi Microsoft Excel dan pustaka *pandas* di Python.
- Transformasi Nilai Kategorikal**
 Atribut kategorikal seperti Jenis_Kelamin, Merokok, dan Aktivitas_Olahraga dikonversi ke nilai numerik menggunakan *Label Encoding*, di mana nilai "Ya" dikodekan menjadi 1 dan "Tidak" menjadi 0.
- Pengkodean Label Target**
 Atribut Hasil sebagai variabel target dikodekan dalam format biner, yaitu "Ya" = 1 (positif penyakit paru) dan "Tidak" = 0 (negatif). Jika dataset terdiri dari beberapa kategori penyakit, maka digunakan pengkodean multikelas.
- Validasi Tipe Data dan Skala**
 Tipe data setiap atribut diperiksa agar sesuai dengan kebutuhan algoritma klasifikasi. Selain itu, jika digunakan algoritma yang sensitif terhadap skala, seperti KNN atau SVM, maka

atribut numerik seperti Usia dinormalisasi menggunakan Min-Max Scaling atau Standardization.

Seluruh tahapan ini dilakukan sebelum pelatihan model menggunakan pustaka *scikit-learn*, guna memastikan bahwa data yang digunakan telah bersih, terstruktur, dan layak untuk diproses lebih lanjut.

2.3 Pemilihan Fitur

Untuk mengidentifikasi atribut yang paling relevan dalam klasifikasi penyakit paru-paru, dilakukan proses seleksi fitur menggunakan metode **Information Gain**. Metode ini dipilih karena mampu mengukur kontribusi suatu fitur dalam mengurangi tingkat ketidakpastian (*entropy*) terhadap label kelas. Teknik ini sangat sesuai untuk data yang didominasi oleh fitur kategorikal atau biner, seperti *Merokok*, *Jenis_Kelamin*, dan *Aktivitas_Olahraga*.

2.3.1 Rumus Information Gain

Perhitungan Information Gain (IG) mengacu pada pengurangan nilai entropi sebelum dan sesudah pemisahan data berdasarkan atribut tertentu. Adapun rumus Information Gain sebagai berikut:

$$Gain(C, A) = Entropy(C) - \sum_{v \in Values(A)} \frac{|C_v|}{|C|} \times Entropy(C_v)$$

2.3.2 Tahapan Implementasi

Implementasi perhitungan Information Gain dilakukan menggunakan bahasa pemrograman Python dengan pustaka *pandas* dan *scikit-learn*. Berikut ini adalah tahapan proses seleksi fitur:

1. Melakukan *label encoding* terhadap atribut kategorikal agar dapat diproses oleh algoritma klasifikasi.
2. Menghitung entropi awal dari dataset berdasarkan distribusi kelas pada atribut *Hasil*.
3. Setelah data dibagi, hitung entropi berdasarkan nilai setiap atribut.
4. Menghitung nilai Information Gain dari setiap atribut sebagai selisih antara entropi awal dan entropi setelah split.
5. Mengurutkan atribut berdasarkan nilai Information Gain tertinggi.

2.3.3 Hasil Perhitungan Information Gain

Tabel berikut menunjukkan hasil perhitungan nilai Information Gain untuk masing-masing fitur dalam dataset:

<div align="center">

Tabel 4.1
Hasil Seleksi Fitur Berdasarkan Information Gain

No	Atribut	Entropi Setelah Split	Information Gain
1	Merokok	0.5177	0.4810
2	Jenis_Kelamin	0.7790	0.2196
3	Bekerja	0.8577	0.1409
4	Penyakit_Bawaan	0.8723	0.1263
5	Aktivitas_Olahraga	0.8835	0.1152

No	Atribut	Entropi Setelah Split	Information Gain
6	Asuransi	0.9761	0.0225
7	Rumah_Tangga	0.9900	0.0087
8	Usia	0.9962	0.0024
9	Aktivitas_Begadang	0.9976	0.0011

2.3.4 Analisis dan Interpretasi

Berdasarkan hasil Tabel 4.1, atribut *Merokok* merupakan fitur yang paling signifikan dalam memengaruhi hasil klasifikasi penyakit paru-paru dengan nilai Information Gain sebesar 0.4810. Atribut lain seperti *Jenis_Kelamin*, *Penyakit_Bawaan*, dan *Aktivitas_Olahraga* juga memiliki pengaruh yang cukup besar terhadap prediksi kelas.

Sementara itu, atribut dengan nilai Information Gain yang rendah, seperti *Aktivitas_Begadang*, *Usia*, dan *Rumah_Tangga*, dapat dipertimbangkan untuk dihilangkan guna menyederhanakan model dan meningkatkan efisiensi proses klasifikasi.

2.3.5 Kesesuaian dengan Literatur

Hasil ini sejalan dengan penelitian yang dilakukan oleh Salsabila dan Andriani [4], yang menggunakan algoritma *decision tree* berbasis C4.5 dalam klasifikasi penyakit paru-paru. Dalam penelitian tersebut, atribut *Merokok* dan *Jenis_Kelamin* juga ditemukan sebagai fitur dominan dalam memengaruhi hasil klasifikasi.

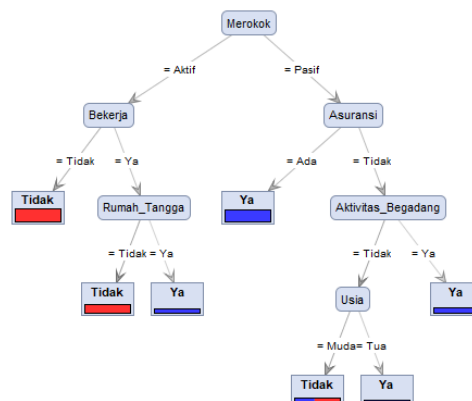
2.4 Pembangunan Model

Model klasifikasi penyakit paru-paru dibangun menggunakan algoritma Decision Tree di RapidMiner Studio. Dataset yang telah melalui tahap praproses dimasukkan ke dalam proses analisis, dengan atribut Hasil ditetapkan sebagai label kelas.

Langkah-langkah pembangunan model meliputi:

- Mengimpor data menggunakan operator Read Excel.
- Menetapkan label menggunakan Set Role.
- Melakukan konversi atribut kategorikal dengan Nominal to Binominal.
- Menerapkan algoritma Decision Tree dengan kriteria pemisahan Information Gain dan pengaturan maksimum kedalaman pohon.
- Gunakan Apply Model and Performance (Classification) untuk menilai performa model.

Setelah model dilatih, RapidMiner menghasilkan visualisasi pohon keputusan yang menunjukkan alur klasifikasi berdasarkan atribut paling informatif. Pohon tersebut dapat dilihat pada Gambar 4.2.



2.5 Evaluasi Model

Evaluasi model bertujuan untuk mengukur sejauh mana akurasi dan keandalan algoritma Decision Tree dalam mengklasifikasikan data penyakit paru-paru. Proses evaluasi dilakukan di RapidMiner dengan menggunakan operator *Performance (Classification)*, yang menghasilkan metrik utama seperti akurasi, precision, recall, dan F1-score.

Berdasarkan hasil evaluasi, diperoleh nilai sebagai berikut:

- Akurasi sebesar 89,47%, menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar.
- Precision sebesar 90% menunjukkan bahwa sebagian besar prediksi positif yang dibuat oleh model benar.
- Recall sebesar 85%, menandakan bahwa model dapat mendeteksi mayoritas kasus positif (pasien yang memiliki penyakit).
- Sebuah skor F1 sebesar 87,4% menunjukkan keseimbangan antara ketepatan dan recall.

Secara keseluruhan, performa model tergolong baik dan stabil, sehingga dapat digunakan untuk membantu proses klasifikasi awal penyakit paru-paru berdasarkan atribut gaya hidup dan kondisi kesehatan pasien. Nilai metrik yang seimbang juga menunjukkan bahwa model tidak hanya fokus pada ketepatan prediksi, tetapi juga pada kelengkapan dalam mendeteksi kasus sebenarnya.

3 HASIL DAN PEMBAHASAN

Penelitian ini bertujuan membangun model klasifikasi penyakit paru-paru menggunakan algoritma *Decision Tree* berbasis data rekam medis pasien. Data diambil dari platform Kaggle dan telah melalui tahap praproses, seleksi fitur menggunakan *Information Gain*, serta implementasi model di RapidMiner Studio.

3.1 Hasil Pembangunan Model

Model yang dibangun menggunakan algoritma *Decision Tree* menghasilkan struktur pohon keputusan yang dapat diinterpretasikan dengan baik. Atribut **Merokok** muncul sebagai akar pohon (root node), yang berarti merupakan atribut paling signifikan dalam memengaruhi hasil klasifikasi. Cabang selanjutnya melibatkan atribut seperti **Bekerja**, **Asuransi**, dan **Aktivitas_Begadang**, yang membentuk jalur keputusan berdasarkan kondisi pasien.

Struktur pohon ini menunjukkan bagaimana kombinasi faktor gaya hidup dan kondisi sosial dapat digunakan untuk memprediksi kemungkinan seseorang mengidap penyakit paru-paru.

3.2 Evaluasi Kinerja Model

Evaluasi model dilakukan menggunakan data uji dengan operator *Performance (Classification)* di RapidMiner. Hasil metrik evaluasi disajikan dalam tabel berikut:

Tabel 4.1
Hasil Evaluasi Model Decision Tree

Metrik	Nilai (%)
Akurasi	89,47
Precision	90,00
Recall	85,00
F1-Score	87,40

Hasil evaluasi menunjukkan bahwa model memiliki **akurasi yang tinggi**, menunjukkan kemampuannya dalam mengklasifikasikan data dengan benar. Precision sebesar **90%** menunjukkan bahwa sebagian besar prediksi positif memang benar, sementara recall sebesar **85%**

menunjukkan bahwa model berhasil menangkap sebagian besar kasus positif. Nilai F1-Score yang seimbang (**87,4%**) menandakan performa yang stabil antara ketepatan dan kelengkapan.

3.3 Pembahasan

Berdasarkan perhitungan *Information Gain*, atribut **Merokok** memiliki nilai tertinggi (0,4810), diikuti oleh **Jenis_Kelamin**, **Bekerja**, dan **Penyakit_Bawaan**. Hal ini menunjukkan bahwa faktor-faktor tersebut memiliki kontribusi signifikan dalam membedakan kategori penyakit paru-paru.

Model *Decision Tree* tidak hanya unggul dari sisi akurasi, tetapi juga **mudah diinterpretasikan**, sehingga cocok untuk digunakan dalam sistem pendukung keputusan medis. Visualisasi pohon keputusan memungkinkan tenaga medis memahami logika di balik hasil prediksi, yang menjadi nilai tambah dibanding algoritma lain yang bersifat “black box”.

Hasil ini sejalan dengan penelitian sebelumnya oleh Salsabila dan Andriani [4], yang menyatakan bahwa fitur merokok dan jenis kelamin merupakan faktor penting dalam klasifikasi penyakit paru-paru menggunakan algoritma C4.5. Dengan demikian, model ini dinilai efektif, efisien, dan aplikatif dalam konteks klinis.

4 KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Decision Tree* mampu membangun model klasifikasi penyakit paru-paru secara efektif dengan tingkat akurasi yang tinggi. Model ini didasarkan pada data rekam medis pasien dan menghasilkan struktur pohon keputusan yang informatif dan mudah dipahami oleh tenaga medis. Merokok, Jenis_Kelamin, dan Penyakit_Bawaan adalah atribut yang paling relevan terhadap hasil klasifikasi, dan proses seleksi fitur menggunakan metode Informasi Gain berhasil. Evaluasi terhadap model menghasilkan nilai akurasi sebesar 89,47%, precision 90%, recall 85%, dan F1-score 87,4%, yang menunjukkan bahwa model memiliki performa yang seimbang dan dapat diandalkan. Penggunaan RapidMiner Studio turut mempermudah proses pembangunan dan evaluasi model secara visual, sehingga model ini dinilai layak untuk diterapkan sebagai bagian dari sistem pendukung keputusan medis dalam proses deteksi dini penyakit paru-paru.

5 KONTRIBUSI PENELITIAN

Penelitian ini membantu membangun model klasifikasi penyakit paru-paru yang akurat dan mudah digunakan yang menggunakan algoritma *Decision Tree*. Dengan memanfaatkan fitur-fitur penting seperti kebiasaan merokok dan jenis kelamin, serta seleksi atribut berbasis *Information Gain*, model ini dapat mendukung proses diagnosis awal dan pengambilan keputusan medis secara lebih efektif di lingkungan klinis.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua orang yang membantu menyusun penelitian ini. Ucapan khusus disampaikan kepada dosen pembimbing yang telah memberikan arahan dan masukan berharga selama proses penelitian berlangsung. Terima kasih juga kepada tim teknis dan rekan-rekan yang turut membantu dalam proses praproses data, analisis fitur, serta pembangunan model klasifikasi penyakit paru-paru.

Tidak lupa, penulis menghargai kontribusi platform penyedia dataset seperti Kaggle serta perangkat lunak RapidMiner dan Python yang memungkinkan eksperimen dilakukan secara efisien. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya dalam deteksi dini penyakit paru-paru.

DAFTAR PUSTAKA

- [1] M. A. Dinova and R. A. Purnama, "KLASIFIKASI UNTUK MENDETEKSI PENYAKIT PARU PARU DENGAN ALGORITMA DECISION TREE," pp. 1–7.
- [2] M. F. Salim and S. Sugeng, "Analisis Rekam Medis Pasien Diabetes Mellitus Melalui Implementasi Teknik Data Mining di RSUP Dr. Sardjito Yogyakarta," *J. Kesehat. Vokasional*, vol. 2, no. 2, p. 167, 2018, doi: 10.22146/jkesvo.30331.
- [3] D. Larassati, A. Zaidiah, and S. Afrizal, "Sistem Prediksi Penyakit Jantung Koroner Menggunakan Metode Naive Bayes," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 7, no. 2, pp. 533–546, 2022, doi: 10.29100/jipi.v7i2.2842.
- [4] R. Pambudi, A. Rahman Harahap, F. Dwitama Saputra, and M. Jusub, "Klasifikasi Penyakit Paru-paru Menggunakan Metode Decision Tree," *OKTAL J. Ilmu Komput. dan Sci.*, vol. 3, no. 9, pp. 2397–2398, 2024, [Online]. Available: <https://journal.mediapublikasi.id/index.php/oktal>