



Department of Digital Business

**Journal of Artificial Intelligence and Digital Business (RIGGS)**

Homepage: <https://journal.ilmudata.co.id/index.php/RIGGS>

Vol. 4 No. 3 (2025) pp: 4769-4775

P-ISSN: 2963-9298, e-ISSN: 2963-914X

---

## Klasifikasi Dampak dan Kondisi Pasien Hepatitis Menggunakan Metode Decision Tree

Luhur Pambudi, Sefrika

Fakultas Teknik Dan Informatika, Universitas Bina Sarana Informatika

[cihuybuseng@gmail.com](mailto:cihuybuseng@gmail.com)

### Abstrak

Hepatitis masih menjadi salah satu tantangan serius di bidang kesehatan, terutama karena sifatnya yang sering “diam-diam berbahaya”. Hepatitis C, misalnya, kerap tidak menimbulkan gejala nyata pada fase awal, sehingga banyak pasien baru terdeteksi ketika kondisinya sudah cukup parah. Berangkat dari permasalahan tersebut, penelitian ini berupaya menghadirkan pendekatan data mining melalui algoritma Decision Tree tipe C4.5 untuk membantu memprediksi kondisi pasien hepatitis secara lebih sistematis. Dataset yang digunakan berasal dari platform Kaggle dengan total 155 entri data klinis, mencakup variabel seperti usia, jenis kelamin, konsumsi obat, hingga gejala medis tertentu. Tahapan penelitian meliputi preprocessing data, pelatihan model, evaluasi performa dengan metrik akurasi, precision, recall, serta pembuatan visualisasi pohon keputusan menggunakan RapidMiner. Hasil eksperimen menunjukkan bahwa model mampu mencapai akurasi optimal sebesar 80,65% dengan proporsi data latih dan uji 80:20. Selain itu, precision pada kelas “Live” mencapai 91,30% dan recall pada kelas “Die” sebesar 66,67%. Pohon keputusan yang terbentuk juga mudah diinterpretasikan, menjadikannya relevan sebagai sarana pendukung keputusan klinis. Dengan demikian, penelitian ini tidak hanya menawarkan gambaran teknis, tetapi juga membuka peluang pemanfaatan teknologi prediktif dalam deteksi dini hepatitis.

*Kata kunci: Hepatitis, Decision Tree, C4.5, Data Mining, RapidMiner, Klasifikasi Medis*

### 1. Latar Belakang

Hepatitis, khususnya tipe C, masih menjadi tantangan kesehatan global yang sulit diatasi karena sifatnya yang laten pada tahap awal. Berbeda dengan penyakit menular lain yang menampilkan gejala jelas, infeksi hepatitis C sering berlangsung diam-diam hingga pasien berada pada kondisi kronis. Laporan World Health Organization (2024) menyebut lebih dari 350 juta orang hidup dengan hepatitis kronis, menjadikannya penyebab kematian infeksi kedua setelah COVID-19[1]. Di Indonesia, prevalensinya masih tinggi, diperburuk oleh keterlambatan diagnosis, rendahnya kesadaran masyarakat, dan distribusi layanan kesehatan yang belum merata. Kondisi ini mempertegas bahwa deteksi dini bukan lagi sekadar rekomendasi, tetapi kebutuhan mendesak dalam menekan angka morbiditas dan mortalitas. Gejala awal hepatitis C seperti kelelahan, mual, anoreksia, dan nyeri perut kanan atas kerap dianggap ringan sehingga pasien cenderung abai [2]. Akibatnya, sebagian besar penderita baru terdeteksi setelah komplikasi berat seperti sirosis atau kanker hati muncul [3]. Di sisi lain, metode diagnostik berbasis laboratorium yang selama ini menjadi standar memerlukan biaya besar dan infrastruktur yang belum sepenuhnya tersedia di fasilitas kesehatan primer. Kesenjangan inilah yang kemudian memunculkan kebutuhan akan pendekatan teknologi yang lebih cepat, efisien, dan mudah diakses.

Perkembangan machine learning (ML) dalam beberapa tahun terakhir membuka peluang baru dalam diagnosis berbasis data. Algoritma ML memungkinkan analisis data klinis dalam jumlah besar untuk menghasilkan prediksi yang akurat sekaligus objektif. Salah satu algoritma yang menonjol adalah Decision Tree, yang secara konsisten menunjukkan kinerja baik dalam klasifikasi penyakit karena model yang dihasilkan transparan dan mudah dipahami. Studi oleh Gunawan et al. (2024) membuktikan bahwa Decision Tree mampu memanfaatkan atribut sederhana seperti kelelahan dan pembesaran hati untuk menghasilkan klasifikasi hepatitis yang valid[4]. Demikian pula, Nugroho et al. (2024) menggunakan algoritma CART dan mencapai akurasi tinggi dalam diagnosis awal[5]. Fernando et al. (2021) bahkan secara khusus membandingkan berbagai algoritma machine learning untuk klasifikasi pasien hepatitis dan menemukan bahwa Decision Tree tetap kompetitif dibanding metode lain[6]. Kelebihan Decision Tree juga ditegaskan oleh Gunawan dan Pratama (2024), yang membandingkannya dengan algoritma lain dan menemukan keunggulan signifikan dalam hal interpretabilitas. Dalam konteks medis, interpretabilitas ini penting karena setiap keputusan harus dapat dijelaskan kepada pasien maupun tenaga

kesehatan. Parkhan et al. (2022) secara khusus mengaplikasikan algoritma C4.5 untuk hepatitis C dan menunjukkan hasil stabil dengan tingkat akurasi yang menjanjikan[7].

Namun, penelitian lain juga memperlihatkan bahwa model berbasis ensemble seperti Random Forest dan XGBoost mampu memberikan akurasi tinggi. Ciobica et al. (2023), misalnya, melaporkan kombinasi LASSO, Random Forest, dan XGBoost dapat mengidentifikasi faktor penting hepatitis dengan akurasi 81% [8]. Saied et al. (2023) bahkan mencapai akurasi lebih dari 98% menggunakan Naïve Bayes dan Random Forest pada kasus hepatitis B dan C[9]. Koh dan Kim (2022) turut memperkuat temuan ini dengan menunjukkan bahwa pendekatan Tree Ensemble efektif dalam menghubungkan faktor gaya hidup dengan risiko kanker hati pada pasien hepatitis B [10]. Sementara itu, Lu et al. (2023) membuktikan bahwa algoritma Decision Tree tidak hanya bermanfaat untuk diagnosis hepatitis C, tetapi juga dapat digunakan untuk memprediksi kanker hati pada pasien pasca eradikasi virus[11]. Meskipun demikian, kompleksitas tinggi dari model tersebut membuatnya sulit diinterpretasikan, terutama oleh tenaga medis non-teknis maupun masyarakat umum. Dengan kata lain, meski akurasi meningkat, manfaat praktis berkurang jika hasil tidak dapat dipahami dengan jelas.

Dari berbagai penelitian tersebut, terlihat adanya kesenjangan: fokus mayoritas studi adalah akurasi teknis, sedangkan aspek keterpahaman dan aksesibilitas hasil klasifikasi masih terabaikan. Padahal, dalam konteks kesehatan masyarakat, transparansi model dan kemampuan untuk menjelaskan keputusan menjadi elemen kunci. Sebuah sistem prediksi yang mampu menunjukkan jalur logika secara visual tidak hanya memudahkan dokter, tetapi juga meningkatkan literasi kesehatan pasien. Penelitian ini hadir untuk mengisi celah tersebut dengan mengembangkan model klasifikasi hepatitis menggunakan algoritma Decision Tree tipe C4.5. Algoritma ini dipilih karena mampu menangani atribut kategorikal dan numerik sekaligus, serta menghasilkan model yang dapat divisualisasikan dalam bentuk pohon keputusan yang intuitif. C4.5 juga menggunakan metode Gain Ratio untuk mengatasi bias atribut dengan banyak nilai unik sehingga hasil klasifikasi lebih stabil dan dapat diandalkan.

Kebaruhan penelitian ini terletak pada orientasinya yang tidak hanya menekankan akurasi teknis, tetapi juga keterpahaman hasil bagi berbagai kalangan. Dengan memanfaatkan dataset terbuka dari Kaggle yang berisi 155 data pasien hepatitis, penelitian ini menekankan pentingnya penyajian hasil dalam bentuk visual yang sederhana, transparan, dan logis. Pendekatan ini berbeda dari sebagian besar studi sebelumnya yang cenderung hanya menyoroiti performa model tanpa mempertimbangkan aspek keberterimaan pengguna. Selain itu, penelitian ini memanfaatkan RapidMiner sebagai alat visualisasi sehingga proses analisis dapat dijangkau oleh pengguna tanpa latar belakang teknis pemrograman.

Pertanyaan utama yang ingin dijawab adalah: dapatkah algoritma C4.5 tidak hanya memberikan prediksi akurat, tetapi juga menyajikan hasil yang dapat diinterpretasikan dengan mudah sehingga mendukung deteksi dini hepatitis C bagi tenaga medis dan masyarakat? Pertanyaan ini sekaligus menegaskan tujuan penelitian, yaitu mengembangkan sistem klasifikasi berbasis data klinis yang andal dan inklusif.

Dengan demikian, penelitian ini diharapkan mampu memberikan kontribusi pada dua ranah sekaligus. Pada ranah akademik, hasilnya memperkaya literatur mengenai implementasi Decision Tree C4.5 dalam diagnosis medis dengan menekankan aspek interpretabilitas. Pada ranah praktis, penelitian ini berupaya menjembatani kompleksitas teknologi dengan kebutuhan nyata di lapangan, yakni menyediakan alat bantu deteksi dini yang cepat, sederhana, dan transparan. Upaya ini bukan hanya bagian dari pengembangan teknologi kesehatan, tetapi juga langkah konkret menuju sistem pelayanan kesehatan yang lebih responsif, modern, dan berorientasi pada masyarakat.

## 2. Metode Penelitian

Data utama berasal dari platform terbuka Kaggle yang memuat 155 rekam medis pasien hepatitis. Setiap entri mencakup atribut numerik dan kategorikal, mulai dari usia, jenis kelamin, penggunaan steroid maupun antivirus, hingga gejala klinis seperti kelelahan, malaise, anoreksia, serta kondisi hati. Variasi atribut ini memberi ruang yang cukup untuk membangun model klasifikasi yang representatif dan mendukung generalisasi.

Sebelum membentuk model, data diproses melalui tahap pra-pemrosesan. Nilai kosong diperiksa dan ditangani, entri duplikat dihapus, sementara variabel kategorikal dikonversi menjadi representasi numerik agar kompatibel dengan perangkat analisis. Langkah-langkah ini memastikan bahwa pola yang terbentuk bukan hasil dari bias atau ketidakaturan data.

Dataset kemudian dipisahkan menjadi data latih dan data uji dengan rasio 80:20. Teknik *stratified sampling* digunakan untuk menjaga keseimbangan proporsi kelas pasien hidup dan meninggal, sehingga hasil evaluasi tidak dipengaruhi oleh distribusi kelas yang timpang.

Model dibangun menggunakan algoritma Decision Tree tipe C4.5. Proses pembentukan pohon mengikuti perhitungan entropy untuk mengukur ketidakmurnian, dilanjutkan dengan perhitungan *information gain* dan *gain ratio* untuk memilih atribut pemisah terbaik. Perhitungan manual pada beberapa atribut kunci, seperti *Fatigue* dan *LiverBig*, dilakukan sebagai validasi awal logika model sebelum seluruh dataset diproses. Metode C4.5 dipilih karena kemampuannya menangani atribut campuran serta menghasilkan pohon keputusan yang lebih stabil dan adil terhadap atribut dengan banyak nilai unik [4].

Kinerja model dievaluasi menggunakan metrik akurasi, precision, recall, dan *confusion matrix* (Riliandhita, 2024). Untuk memastikan reliabilitas, pengujian dilengkapi dengan teknik *k-fold cross validation* dengan nilai  $k=10$ . Evaluasi ini memberikan gambaran yang lebih menyeluruh, sekaligus menghindari bias dari satu kali pembagian data. Hasil klasifikasi divisualisasikan dalam bentuk pohon keputusan menggunakan perangkat lunak RapidMiner. Visualisasi ini menampilkan jalur klasifikasi dari akar hingga simpul daun, sehingga interpretasi keputusan dapat ditelusuri dengan mudah oleh tenaga medis maupun pengguna non-teknis.

### 3. Hasil dan Diskusi

#### 3.1. Pengumpulan Data

Data penelitian ini diperoleh dari situs Kaggle berupa data sekunder yang berisi informasi klinis pasien penderita hepatitis. Dataset tersebut mencakup 155 pasien dengan atribut yang merepresentasikan kondisi kesehatan dan riwayat medis masing-masing pasien. Atribut yang digunakan meliputi usia (age), jenis kelamin (gender), penggunaan steroid, penggunaan antivirus, gejala klinis seperti kelelahan (fatigue), rasa tidak enak badan (malaise), hilangnya nafsu makan (anorexia), serta kondisi hati yang dilihat dari pembesaran hati (liverBig) dan kekakuan hati (liverFirm). Atribut target berupa status pasien dengan dua kemungkinan nilai, yaitu live (pasien bertahan hidup) dan die (pasien meninggal dunia). Seluruh data ini digunakan sebagai dasar dalam membangun model klasifikasi berbasis algoritma Decision Tree untuk memprediksi risiko kematian pasien hepatitis.

#### 3.2 Preprocessing Data

Tahapan preprocessing dilakukan untuk menyiapkan data sebelum digunakan dalam pembangunan model klasifikasi. Proses ini meliputi beberapa langkah. Pertama, dilakukan pemeriksaan dan pembersihan data (data cleaning) untuk mengidentifikasi adanya nilai kosong, duplikasi, maupun data yang tidak konsisten. Nilai kosong ditangani dengan metode imputasi sederhana, seperti pengisian berdasarkan modus, atau penghapusan baris apabila data yang hilang terlalu banyak.

Langkah berikutnya adalah penyandian atribut kategorikal (encoding). Atribut seperti gender, steroid, antivirals, fatigue, malaise, anorexia, liverBig, liverFirm, dan target dikonversi ke bentuk numerik dengan teknik label encoding. Misalnya, nilai yes dikodekan sebagai 1, sedangkan no menjadi 0. Untuk atribut target, pasien live dikodekan sebagai 1, sementara die sebagai 0. Setelah itu, seluruh data disesuaikan dalam format numerik agar dapat diproses oleh perangkat lunak pemodelan. Karena algoritma Decision Tree tidak memerlukan normalisasi data, maka tahap tersebut tidak dilakukan secara eksplisit, tetapi pemeriksaan terhadap outlier tetap dilaksanakan untuk memastikan tidak ada nilai ekstrem yang memengaruhi hasil pemodelan.

Tahap selanjutnya adalah pembagian data (data splitting). Untuk mengevaluasi kinerja model secara menyeluruh, dataset dibagi menjadi data latih dan data uji dalam beberapa skema, yaitu 90:10, 80:20, 70:30, dan 60:40. Proses pembagian dilakukan dengan metode shuffled sampling untuk memastikan distribusi kelas target merata dan acak. Evaluasi model kemudian dilakukan pada setiap skema untuk membandingkan kinerja berdasarkan metrik akurasi, presisi, dan recall.

#### 3.3. Pemodelan Algoritma Decision Tree

Setelah melalui tahapan preprocessing, data yang telah bersih dan terstruktur selanjutnya diproses untuk membangun model klasifikasi menggunakan algoritma Decision Tree. Pemodelan ini dilakukan dengan memanfaatkan perangkat lunak RapidMiner, yang dipilih karena memiliki antarmuka visual yang memudahkan perancangan alur kerja pemodelan sekaligus menyediakan berbagai operator yang mendukung proses analisis data.

Tahapan awal dimulai dengan operator *Read Excel*, yang berfungsi untuk membaca dan memuat dataset ke dalam sistem. Dataset yang dimuat kemudian dihubungkan dengan operator *Split Data*, di mana data dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk membangun dan melatih model, sedangkan data

uji digunakan untuk menilai performa model terhadap data baru yang belum pernah dilihat sebelumnya. Proporsi pembagian data dilakukan dalam beberapa skema, yakni 60:40, 70:30, 80:20, dan 90:10, untuk mengetahui pengaruh variasi rasio pembagian terhadap kualitas model yang dihasilkan.

Setelah data dibagi, operator *Decision Tree* diaplikasikan pada data latih. Algoritma ini dipilih karena kemampuannya dalam menghasilkan model yang mudah dipahami, bersifat interpretable, serta dapat menangani variabel kategorikal maupun numerik. Dalam proses pembentukan model, algoritma *Decision Tree* bekerja dengan membagi dataset ke dalam cabang-cabang keputusan berdasarkan atribut yang paling informatif, sehingga menghasilkan struktur pohon yang dapat digunakan untuk klasifikasi. Setiap node pada pohon keputusan merepresentasikan sebuah atribut, cabangnya menunjukkan nilai atribut, dan daun pohon (leaf) menunjukkan hasil klasifikasi akhir, yaitu status pasien hepatitis yang *live* atau *die*.

Tahap selanjutnya adalah evaluasi kinerja model yang dilakukan dengan menggunakan operator *Apply Model* dan *Performance*. Operator *Apply Model* berfungsi untuk menerapkan model *Decision Tree* yang telah terbentuk pada data uji, sementara operator *Performance* menghitung metrik evaluasi, yaitu akurasi, presisi, dan recall. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data, presisi menunjukkan sejauh mana model mampu mengklasifikasikan pasien yang diprediksi meninggal sesuai dengan kenyataan, sedangkan recall mengukur kemampuan model dalam menemukan semua pasien yang benar-benar meninggal dari seluruh data.

Hasil pengujian menunjukkan bahwa dari keempat skema pembagian data, proporsi 80:20 memberikan nilai akurasi tertinggi dibandingkan dengan skema lainnya. Hal ini menunjukkan bahwa pembagian dengan porsi 80% data latih dan 20% data uji memberikan keseimbangan yang optimal antara kemampuan model dalam belajar pola dari data latih serta menguji generalisasi pada data uji. Oleh karena itu, skema 80:20 dipilih sebagai konfigurasi terbaik dalam penelitian ini.

Melalui alur pemodelan tersebut, data latih dimanfaatkan untuk menghasilkan struktur pohon keputusan yang merepresentasikan hubungan antar atribut klinis pasien hepatitis dengan status akhirnya. Sementara itu, data uji dipakai sebagai dasar evaluasi untuk menilai sejauh mana model dapat menggeneralisasi pola yang telah dipelajari dan memprediksi status pasien baru dengan tingkat akurasi yang tinggi. Dengan cara ini, diperoleh model klasifikasi *Decision Tree* yang tidak hanya efektif dalam mengelompokkan pasien ke dalam kategori bertahan hidup atau meninggal dunia, tetapi juga memberikan interpretasi yang jelas mengenai faktor-faktor klinis yang paling berpengaruh terhadap kondisi akhir pasien.

### 3.4. Hasil dan Pembahasan

Keluaran utama dari proses ini berupa *Performance Vector*, yakni hasil komputasi yang merangkum performa model klasifikasi. Di dalamnya terdapat *Confusion Matrix* yang menjadi dasar untuk mengukur tingkat ketepatan model. Melalui matriks ini, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, serta visualisasi pohon keputusan dapat ditelaah secara mendalam. Analisis ini tidak hanya memberikan gambaran numerik, tetapi juga membantu memahami pola prediksi yang dihasilkan model, apakah lebih cenderung akurat pada kelas tertentu atau justru memiliki kelemahan dalam mendeteksi kasus lain.

#### *Perhitungan RapidMiner (Performance Vector)*

Sebagai salah satu metode evaluasi yang paling sering digunakan, *Confusion Matrix* menampilkan informasi detail mengenai perbandingan antara hasil prediksi model dengan kondisi aktual pada data uji. Matriks ini memungkinkan peneliti melihat jumlah kasus yang benar dan salah diprediksi, baik untuk kelas positif maupun negatif. Dengan adanya visualisasi ini, kelemahan dan keunggulan model dapat teridentifikasi dengan lebih jelas. Hasil visualisasi *Confusion Matrix* ditunjukkan pada gambar berikut:

	true LIVE	true DIE	class precision
pred. LIVE	21	2	91.30%
pred. DIE	4	4	50.00%
class recall	84.00%	66.67%	

Gambar 1. Confusion Matrix

Berdasarkan Gambar 1, dari total 124 data pasien yang digunakan dalam pengujian, terdapat 21 pasien yang berhasil diprediksi hidup secara benar (*true positive*). Namun, terdapat pula 4 pasien yang salah diklasifikasikan sebagai meninggal padahal sebenarnya masih hidup (*false negative*). Di sisi lain, 4 pasien berhasil diprediksi

meninggal sesuai dengan kondisi sebenarnya (*true negative*), sedangkan 2 pasien keliru diprediksi hidup meskipun kenyataannya telah meninggal (*false positive*). Dari hasil ini, nilai *precision* untuk kelas LIVE tercatat sebesar 91,30%, sementara kelas DIE hanya mencapai 50,00%. Untuk metrik *recall*, kelas LIVE memperoleh nilai 84,00% dan kelas DIE 66,67%. Angka-angka tersebut memberikan gambaran bahwa model lebih handal dalam mendeteksi pasien yang bertahan hidup, meskipun masih menunjukkan keterbatasan dalam mengenali pasien dengan status meninggal. Temuan ini penting karena memberikan masukan tentang area yang perlu ditingkatkan dari sisi algoritma maupun kualitas data latih.

### Hasil Accuracy RapidMiner

Akurasi merupakan metrik yang mengukur proporsi prediksi benar dari keseluruhan prediksi yang dihasilkan oleh model. Dalam penelitian ini, akurasi dihitung menggunakan data uji yang sebelumnya telah dipisahkan dari data latih, sehingga hasil yang diperoleh benar-benar mencerminkan kemampuan generalisasi model terhadap data baru. Visualisasi hasil akurasi dapat dilihat pada gambar berikut:

accuracy: 80.65%

	true LIVE	true DIE	class precision
pred. LIVE	21	2	91.30%
pred. DIE	4	4	50.00%
class recall	84.00%	66.67%	

Sumber: Hasil Penelitian (2025)

Gambar 2. Accuracy

Dari Gambar 2, terlihat bahwa model *Decision Tree* menghasilkan akurasi sebesar 80,65%. Nilai ini menunjukkan bahwa sekitar delapan dari sepuluh prediksi yang dilakukan model sesuai dengan kondisi sebenarnya. Capaian tersebut dapat dianggap cukup baik mengingat data yang digunakan memiliki variasi atribut yang kompleks. Selain itu, angka ini juga menegaskan bahwa algoritma *Decision Tree* mampu memberikan klasifikasi yang relatif stabil meskipun tidak sempurna. Dengan tingkat akurasi tersebut, model dapat digunakan sebagai alat bantu awal dalam proses diagnosis, meskipun tetap diperlukan pertimbangan medis lebih lanjut untuk memastikan hasil yang lebih presisi.

### Hasil Precision RapidMiner

Untuk mengukur ketepatan prediksi pada kelas tertentu, digunakan metrik *Precision*. Metrik ini sangat penting karena menekankan pada proporsi prediksi positif yang benar-benar sesuai dengan kenyataan. Dalam konteks penelitian ini, *Precision* menilai seberapa akurat model dalam mengklasifikasikan pasien ke dalam kategori LIVE maupun DIE. Visualisasi hasil perhitungan dapat dilihat pada gambar berikut:

precision: 50.00% (positive class: DIE)

	true LIVE	true DIE	class precision
pred. LIVE	21	2	91.30%
pred. DIE	4	4	50.00%
class recall	84.00%	66.67%	

Sumber: Hasil Penelitian (2025)

Gambar 3. Precision

Berdasarkan Gambar 3, diketahui bahwa nilai *precision* untuk kelas DIE hanya mencapai 50%, yang berarti setengah dari prediksi pasien meninggal ternyata keliru. Sebaliknya, untuk kelas LIVE, tingkat *precision* jauh lebih tinggi, yakni 91,30%, sehingga sebagian besar prediksi pasien hidup terbukti benar. Selain itu, jika dilihat dari aspek *recall*, model mampu mengenali 84% dari total kasus sebenarnya yang berstatus LIVE, serta 66,67% dari total kasus sebenarnya yang berstatus DIE. Kombinasi nilai presisi dan *recall* ini memperlihatkan bahwa model lebih dapat diandalkan untuk mengidentifikasi pasien yang bertahan hidup dibandingkan dengan mereka yang berisiko meninggal. Analisis seperti ini penting karena dalam konteks medis, kesalahan mendeteksi pasien berisiko meninggal bisa berdampak serius pada keputusan klinis.

DOI: <https://doi.org/10.31004/riggs.v4i3.2692>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

### Hasil Recall RapidMiner

*Recall* adalah metrik yang menekankan pada kemampuan model dalam menemukan seluruh kasus positif yang ada dalam data. Dengan kata lain, *recall* menjawab pertanyaan: dari semua pasien yang sebenarnya meninggal atau hidup, berapa banyak yang berhasil dikenali dengan benar oleh model? Nilai ini divisualisasikan pada gambar berikut:

recall: 66.67% (positive class: DIE)

	true LIVE	true DIE	class precision
pred. LIVE	21	2	91.30%
pred. DIE	4	4	50.00%
class recall	84.00%	66.67%	

Sumber: Hasil Penelitian (2025)  
Gambar 4. Recall

Berdasarkan Gambar 4, terlihat bahwa nilai *recall* untuk kelas DIE mencapai 66,67%, yang berarti dua pertiga pasien yang benar-benar meninggal dapat diprediksi dengan tepat oleh model. Sementara itu, nilai *recall* untuk kelas LIVE lebih tinggi, yaitu 84%, menunjukkan bahwa sebagian besar pasien yang hidup berhasil dikenali [12]. Jika ditinjau dari sisi presisi, model mencatat 50% untuk kelas DIE dan 91,30% untuk kelas LIVE. Hal ini semakin menegaskan bahwa model lebih optimal ketika berhadapan dengan data pasien yang bertahan hidup, meskipun masih cukup lemah dalam mendeteksi pasien yang berstatus meninggal.

### Pohon Keputusan

Visualisasi struktur pohon keputusan hasil pemrosesan data pasien hepatitis menggunakan algoritma *Decision Tree* disajikan pada Gambar III.10. Pohon ini berfungsi untuk memprediksi status pasien, apakah masuk kategori hidup (LIVE) atau meninggal (DIE), dengan memanfaatkan sejumlah atribut klinis yang telah tersedia.



Sumber: Hasil Penelitian (2025)

Gambar 5. Visualisasi Pohon Keputusan (*Decision Tree*)

Pohon keputusan yang dihasilkan melalui RapidMiner ini dibangun berdasarkan perhitungan *entropy*, *gain*, dan *gain ratio* sebagaimana dijelaskan sebelumnya pada bagian metode. Pada percobaan manual menggunakan 10 data pertama, atribut *Fatigue* dipilih sebagai akar pohon karena memiliki nilai *gain* tertinggi. Namun, ketika seluruh dataset digunakan, atribut *Age* terpilih sebagai akar pohon dengan nilai *gain ratio* tertinggi. Perbedaan ini dapat dipahami karena jumlah data yang lebih besar mampu mengubah distribusi informasi, sehingga atribut yang paling berpengaruh juga bisa berubah.

Secara keseluruhan, hasil ini membuktikan bahwa proses otomatis yang dilakukan RapidMiner sesuai dengan prinsip algoritma C4.5. Setiap simpul pohon dipilih berdasarkan nilai *gain ratio* terbaik, lalu bercabang hingga terbentuk simpul daun dengan kelas akhir, yaitu Live atau Die [13]. Dengan cara ini, sistem yang

diimplementasikan benar-benar menggambarkan langkah kerja algoritma C4.5 sebagaimana teori yang telah dibahas.

Sebagai contoh, pada pasien perempuan dengan atribut *liverBig* = Yes, model akan memprediksi status meninggal (DIE). Namun, jika *liverBig* = No, atribut *Fatigue* menjadi penentu berikutnya. Pasien tanpa gejala kelelahan diklasifikasikan hidup (LIVE), sementara pasien dengan gejala kelelahan perlu ditinjau dari faktor usia. Pasien berusia lebih dari 62 tahun diprediksi hidup (LIVE), usia antara 56,5–62 tahun diprediksi meninggal (DIE), sedangkan usia di bawah 56,5 tahun kembali diprediksi hidup (LIVE). Untuk pasien laki-laki dengan kondisi *liverBig* = Yes, model secara otomatis memberikan prediksi hidup (LIVE).

Dari keseluruhan pola ini, terlihat bahwa atribut yang paling berpengaruh dalam menentukan klasifikasi adalah usia, jenis kelamin, kondisi hati (*liverBig*), serta gejala kelelahan (*Fatigue*). Pohon keputusan ini tidak hanya menyajikan hasil prediksi, tetapi juga memberikan gambaran yang mudah dipahami bagi tenaga medis. Dengan visualisasi yang jelas, dokter atau analis kesehatan dapat menggunakan model ini sebagai referensi awal dalam memprediksi risiko pasien hepatitis berdasarkan data klinis yang tersedia.

#### 4. Kesimpulan

Berdasarkan hasil penelitian mengenai klasifikasi risiko pasien hepatitis menggunakan algoritma *Decision Tree*, diperoleh kesimpulan bahwa atribut klinis seperti usia, jenis kelamin, kondisi hati, gejala kelelahan, dan penggunaan antivirus berperan penting dalam menentukan status pasien apakah bertahan hidup atau meninggal. Model yang dibangun menunjukkan tingkat akurasi sebesar 80,65%, dengan nilai *precision* dan *recall* yang relatif baik meskipun masih terdapat keterbatasan dalam mendeteksi pasien berisiko meninggal. Visualisasi dalam bentuk pohon keputusan memberikan kejelasan jalur logika klasifikasi, sehingga hasilnya mudah dipahami dan dapat dimanfaatkan sebagai alat bantu dalam proses triase serta pengambilan keputusan medis secara lebih cepat dan objektif. Penelitian ini memiliki implikasi bahwa algoritma *Decision Tree* berpotensi diterapkan sebagai sistem pendukung keputusan awal dalam skrining pasien hepatitis. Namun, efektivitas model masih dapat ditingkatkan dengan penggunaan dataset yang lebih besar dan bervariasi agar kemampuan generalisasi menjadi lebih kuat. Untuk penelitian selanjutnya, disarankan dilakukan perbandingan dengan algoritma lain seperti *Random Forest* atau *Support Vector Machine*, serta pengujian pada data klinis nyata di fasilitas kesehatan. Selain itu, pengembangan sistem berbasis web atau aplikasi yang mengintegrasikan hasil model ke dalam praktik klinis dapat menjadi langkah penting guna mempercepat proses identifikasi risiko pasien dan meningkatkan efisiensi pelayanan kesehatan.

#### Referensi

- [1] W. H. Organization, "Global hepatitis report 2024." [Online]. Available: <https://www.who.int/publications/i/item/9789240091672>
- [2] M. Alharthi, H. Alghamdi, and A. Dabbagh, "Machine learning for predicting hepatitis B or C virus infection in diabetic patients," *Sci. Rep.*, vol. 13, no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-49046-9.
- [3] I. Ofri, N. Peleg, M. Leshno, and A. Shlomai, "A decision-making model for prediction of a stable disease course in chronic hepatitis B patients," *Sci. Rep.*, vol. 13, no. 1, pp. 1–8, 2023, doi: 10.1038/s41598-023-50460-2.
- [4] R. G. Gunawan and M. I. Pratama, "Analisa kinerja algoritma machine learning untuk prediksi virus hepatitis C," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 4, no. 3, pp. 772–777, 2024, doi: 10.37859/coscitech.v4i3.6513.
- [5] R. G. Gunawan, M. I. Pratama, F. D. Prasetya, H. W. Nugroho, J. Triloka, and C. Romadhon, "Analisa data mining untuk prediksi penyakit hepatitis C menggunakan algoritma decision tree C4.5 dengan particle swarm optimization," *Semin. Nas. Has. Penelit. dan Pengabd. Masy.*, vol. 4, no. 3, pp. 772–777, 2024, doi: 10.37859/coscitech.v4i3.6513.
- [6] D. Fernando, F. Huwaidi, M. H. Ananto, and S. Pramadya, "Comparison of machine learning algorithms in the role of hepatitis patient disease classification," *SEICT*, vol. 4, no. 2, pp. 159–170, 2021, doi: 10.17509/seict.v2i1.34253.
- [7] A. Parkhan, M. Sugarindra, C. Romadhon, F. D. Prasetya, H. W. Nugroho, and J. Triloka, "Analisa data mining untuk prediksi penyakit hepatitis C menggunakan algoritma decision tree C4.5 dengan particle swarm optimization," *J. Disprotek*, vol. 13, no. 2, pp. 199–209, 2022, doi: 10.34001/jdpt.v12i2.
- [8] A. Ciobica, A. Cristea, and M. Sandu, "Machine learning approaches for the prediction of hepatitis B and C seropositivity," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, 2023, doi: 10.3390/ijerph20032380.
- [9] T. M. Saied, M. A. Sherif, and M. Hammad, "Hepatitis C virus prediction based on machine learning framework: A real-world case study in Egypt," *Knowl. Inf. Syst.*, vol. 65, no. 6, pp. 2595–2617, 2023, doi: 10.1007/s10115-023-01851-4.
- [10] E. Koh and Y. Kim, "Risk association of liver cancer and hepatitis B with tree ensemble and lifestyle features," *Int. J. Environ. Res. Public Health*, vol. 19, no. 22, 2022, doi: 10.3390/ijerph192215171.
- [11] M. Y. Lu *et al.*, "Artikel asli algoritma pohon keputusan memprediksi karsinoma hepatoseluler di antara pasien hepatitis C kronis setelah pemberantasan virus," *Nama J. Tidak Dicantumkan*, vol. 13, no. 1, pp. 190–203, 2023.
- [12] R. E. Nugroho, W. Y. Pamungkas, and J. H. Jaman, "Pendeteksi penyakit hepatitis menggunakan," *Nama J. Tidak Dicantumkan*, vol. 12, no. 3, pp. 3690–3696, 2024.
- [13] M. Azzani, E. Ba-Alawi, W. M. Atroosh, and H. Yadav, "Awareness of cervical cancer and its associated socio-demographic factors among Yemeni immigrant women in Malaysia," *BMC Women's Heal.*, vol. 23, no. 1, 2023, doi: 10.1186/s12905-023-02172-y.