ULIL ALBAB : Jurnal Ilmiah Multidisiplin
Vol.5, No.12, November 2026

Prediksi Penyakit Liver Menggunakan Algoritma Naive Bayes

Meyshe Putri[1], Viovi Bunga Harsono[2], Ade Fitria Lestari[3]
[1,2,3]Program Studi Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
E-mail: meysheputri05@gmail.com[1], viovibungaharsono@gmail.com[2], ade.afr@bsi.ac.id[3]

Article History:
Received: 06 September 2026
Revised: 15 Oktober 2026
Accepted: 29 Oktober 2026

Keywords: *Data Mining, Naïve Bayes, Penyakit Liver, Prediksi, RapidMiner.*

Abstrak: *Penyakit liver merupakan salah satu masalah kesehatan yang sulit dideteksi pada tahap awal karena gejalanya tidak spesifik. Penelitian ini bertujuan menerapkan algoritma Naïve Bayes untuk memprediksi penyakit liver, mengevaluasi kinerja model, serta mengidentifikasi atribut yang paling berpengaruh. Dataset diperoleh dari Kaggle dengan jumlah 1.700 data pasien dan 11 atribut. Analisis dilakukan menggunakan RapidMiner melalui tahap preprocessing, seleksi fitur menggunakan Weight by Information, pembagian data dengan rasio 40:60, 70:30, 80:20, dan 90:10, pembentukan model Naïve Bayes, serta evaluasi menggunakan Accuracy, Precision, Recall, F1-Score, dan Area Under Curve (AUC). Rasio 90:10 menghasilkan performa terbaik dengan Accuracy 82,35%, Precision 84,78%, Recall 82,98%, F1-Score 83,87%, dan AUC 0,895. Atribut Liver Function Test menjadi atribut paling berpengaruh, diikuti Alcohol Consumption dan Smoking. Hasil penelitian menunjukkan bahwa Naïve Bayes mampu memberikan performa yang baik dalam memprediksi penyakit liver pada dataset yang digunakan.*

PENDAHULUAN

Penyakit liver merupakan salah satu permasalahan kesehatan yang perlu mendapat perhatian karena hati memiliki peran penting dalam berbagai proses fisiologis tubuh, termasuk metabolisme, detoksifikasi, dan pengaturan zat-zat penting dalam darah. Gangguan fungsi hati dapat berkembang secara bertahap dan pada tahap awal sering kali tidak menunjukkan gejala yang spesifik. Kondisi tersebut menyebabkan proses identifikasi penyakit liver menjadi tidak mudah dilakukan hanya berdasarkan gejala klinis, sehingga diperlukan pendekatan yang dapat membantu mengidentifikasi pola penyakit berdasarkan karakteristik pasien dan hasil pemeriksaan klinis (Ibrahim et al., 2025). Deteksi dan identifikasi secara lebih awal diperlukan untuk mendukung pengambilan keputusan medis serta meningkatkan peluang penanganan yang tepat.

Perkembangan teknologi informasi telah menghasilkan ketersediaan data kesehatan dalam jumlah besar yang dapat dimanfaatkan untuk mendukung proses analisis penyakit. Salah satu pendekatan yang dapat digunakan adalah data mining, yaitu proses pengolahan dan analisis data untuk menemukan pola atau informasi yang memiliki nilai dalam mendukung pengambilan keputusan. Dalam bidang kesehatan, data mining dapat diterapkan untuk menganalisis karakteristik pasien, mengidentifikasi pola tertentu, serta membangun model klasifikasi dan prediksi penyakit

berdasarkan sejumlah atribut klinis (World Health Organization, 2025). Pemanfaatan pendekatan tersebut memberikan peluang untuk mengembangkan sistem prediksi yang mampu mengolah data pasien secara lebih sistematis.

Sejumlah penelitian sebelumnya telah menunjukkan bahwa metode data mining dapat digunakan untuk melakukan klasifikasi penyakit liver berdasarkan berbagai atribut pasien. Aldana dan Wibowo (2024) menunjukkan bahwa karakteristik yang terdapat dalam dataset pasien dapat dimanfaatkan sebagai dasar dalam membangun model klasifikasi penyakit liver. Meskipun demikian, perbedaan karakteristik dataset, tahapan preprocessing, pemilihan atribut, pembagian data, serta algoritma yang digunakan dapat menghasilkan kinerja model yang berbeda. Oleh karena itu, diperlukan pengujian terhadap algoritma klasifikasi tertentu untuk mengetahui kemampuannya dalam menghasilkan prediksi penyakit liver secara akurat.

Salah satu algoritma klasifikasi yang dapat diterapkan adalah Naïve Bayes. Algoritma ini menggunakan pendekatan probabilistik berdasarkan Teorema Bayes untuk menentukan kelas suatu data berdasarkan probabilitas karakteristik yang dimilikinya. Naïve Bayes memiliki kelebihan berupa proses komputasi yang relatif sederhana dan kemampuan dalam menangani dataset klasifikasi dengan sejumlah atribut. Dalam konteks prediksi penyakit, algoritma ini dapat digunakan untuk menentukan apakah data pasien termasuk ke dalam kelompok terindikasi penyakit liver atau tidak berdasarkan atribut yang tersedia.

Penelitian ini menerapkan algoritma Naïve Bayes untuk memprediksi penyakit liver menggunakan dataset sebanyak 1.700 data pasien yang terdiri atas 11 atribut. Tahapan penelitian meliputi preprocessing data, seleksi fitur menggunakan metode Weight by Information, pembagian dataset menggunakan beberapa rasio data training dan testing, pembentukan model klasifikasi menggunakan algoritma Naïve Bayes, serta evaluasi kinerja model. Penggunaan seleksi fitur dilakukan untuk mengidentifikasi atribut yang memiliki kontribusi terhadap proses klasifikasi sehingga dapat mendukung pembentukan model yang lebih efektif.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengevaluasi kinerja algoritma Naïve Bayes dalam memprediksi penyakit liver berdasarkan data pasien. Kinerja model dievaluasi menggunakan beberapa indikator, yaitu Accuracy, Precision, Recall, F1-Score, dan Area Under the Curve (AUC). Hasil penelitian diharapkan dapat memberikan gambaran mengenai kemampuan algoritma Naïve Bayes dalam melakukan klasifikasi penyakit liver serta menjadi referensi dalam pemanfaatan teknik data mining untuk mendukung analisis dan prediksi penyakit berbasis data klinis.

LANDASAN TEORI

Penyakit Liver

Penyakit liver merupakan kondisi yang memengaruhi struktur atau fungsi organ hati sehingga berbagai proses penting dalam tubuh dapat terganggu. Hati memiliki peranan penting dalam proses metabolisme, penyaringan zat beracun, penyimpanan energi, serta pembentukan protein yang dibutuhkan tubuh. Penyakit liver dapat dipengaruhi oleh berbagai faktor, antara lain konsumsi alkohol, obesitas, diabetes, faktor genetik, serta penggunaan obat-obatan tertentu (Sodikin et al., 2024). Gangguan pada hati dapat berkembang secara bertahap apabila tidak mendapatkan penanganan yang tepat dan dapat berkembang menjadi fibrosis, sirosis, hingga gagal hati (Ginting et al., 2022). Oleh karena itu, identifikasi penyakit liver berdasarkan karakteristik pasien diperlukan untuk membantu mengetahui kondisi pasien dan mendukung proses deteksi secara lebih dini.

RapidMiner

RapidMiner merupakan perangkat lunak berbasis Java yang dapat digunakan dalam proses pengolahan data dan mendukung berbagai teknik analisis, termasuk prediksi. RapidMiner juga menyediakan berbagai operator yang dapat digunakan dalam tahapan preprocessing, pemodelan, dan evaluasi data (Sofyan et al., 2023). RapidMiner mendukung berbagai teknik machine learning, seperti klasifikasi, regresi, dan clustering, serta menyediakan fasilitas evaluasi model (Setyawan et al., 2025). Dalam penelitian ini, RapidMiner digunakan untuk melakukan preprocessing data, seleksi fitur menggunakan Weight by Information, pembagian data, pemodelan Naïve Bayes, dan evaluasi kinerja model.

Algoritma Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema Bayes untuk menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu berdasarkan atribut yang dimilikinya (Riany & Testiana, 2023). Naïve Bayes bekerja berdasarkan probabilitas dan frekuensi kemunculan data pada proses pelatihan. Algoritma ini memiliki proses klasifikasi yang relatif sederhana dan cepat sehingga dapat diterapkan pada berbagai permasalahan klasifikasi (Insan et al., 2023). Dalam penelitian ini, algoritma Naïve Bayes digunakan untuk melakukan klasifikasi terhadap data pasien dan menentukan kelas Diagnosis pada dataset penyakit liver. Model dibangun berdasarkan data training, kemudian digunakan untuk memprediksi kelas Diagnosis pada data testing.

Normalisasi Min-Max

Normalisasi Min-Max merupakan salah satu metode preprocessing yang digunakan untuk menyesuaikan nilai atribut numerik ke dalam rentang tertentu. Pada penelitian ini, normalisasi dilakukan dengan mengubah nilai atribut ke dalam rentang 0 sampai 1. Normalisasi Min-Max diterapkan pada atribut BMI, Alcohol Consumption, Physical Activity, dan Liver Function Test untuk mempersiapkan data numerik sebelum digunakan dalam proses pemodelan.

Weight by Information

Weight by Information merupakan metode yang digunakan untuk mengetahui tingkat kepentingan atau bobot suatu atribut terhadap atribut target. Pembobotan atribut dilakukan berdasarkan kontribusi informasi yang dimiliki masing-masing atribut terhadap kelas target. Dalam penelitian ini, Weight by Information digunakan untuk menghitung bobot setiap atribut terhadap target Diagnosis. Hasil pembobotan menunjukkan bahwa Liver Function Test memiliki bobot tertinggi, kemudian diikuti oleh Alcohol Consumption dan Smoking. Hasil tersebut digunakan untuk mengetahui atribut yang memiliki kontribusi informasi terbesar terhadap proses klasifikasi penyakit liver.

Evaluasi Kinerja Model

Evaluasi dilakukan untuk mengetahui kemampuan model Naïve Bayes dalam melakukan prediksi penyakit liver. Pengukuran kinerja model menggunakan Accuracy, Precision, Recall, F1-Score, dan Area Under the Curve (AUC). Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data. Precision menunjukkan ketepatan model dalam mengidentifikasi data yang diprediksi sebagai kelas positif, sedangkan Recall menunjukkan kemampuan model dalam menemukan seluruh data positif yang sebenarnya. F1-Score merupakan ukuran yang menggabungkan Precision dan Recall, sehingga dapat memberikan gambaran keseimbangan antara kedua metrik tersebut. Sementara itu, AUC digunakan untuk menggambarkan kemampuan model dalam membedakan kelas positif dan negatif pada berbagai nilai ambang klasifikasi.

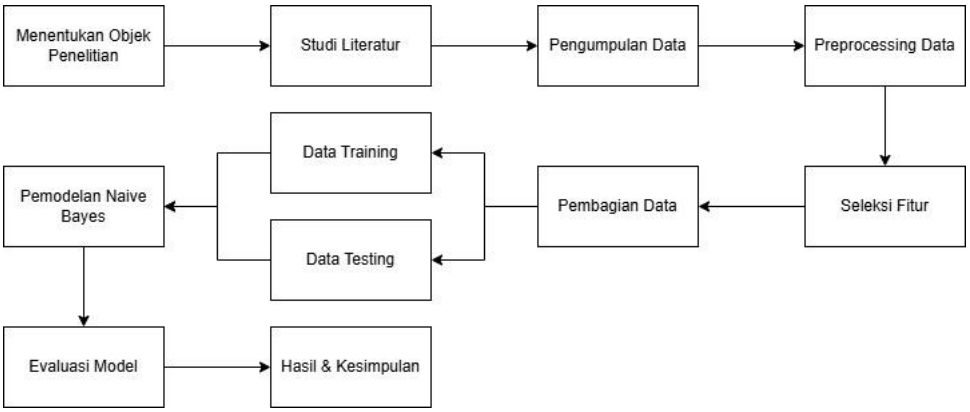
Berdasarkan kajian tersebut, penelitian ini menggunakan algoritma Naïve Bayes sebagai metode klasifikasi dan Weight by Information sebagai metode seleksi fitur pada dataset penyakit

liver. Kombinasi kedua metode tersebut digunakan untuk mengetahui kemampuan model dalam melakukan prediksi penyakit liver serta mengevaluasi kinerjanya berdasarkan Accuracy, Precision, Recall, F1-Score, dan AUC.

METODE PENELITIAN

Penelitian ini menggunakan dataset penyakit liver yang diperoleh dari Kaggle dengan jumlah 1.700 data pasien dan 11 atribut, yaitu Age, Gender, BMI, Alcohol Consumption, Smoking, Genetic Risk, Physical Activity, Diabetes, Hypertension, Liver Function Test, dan Diagnosis. Kelas Diagnosis terdiri atas 1 yang menunjukkan pasien terindikasi penyakit liver dan 0 yang menunjukkan pasien tidak terindikasi penyakit liver.

Pengolahan data dilakukan menggunakan RapidMiner. Tahapan penelitian meliputi pengumpulan data, preprocessing, seleksi fitur, pembagian data, pemodelan menggunakan algoritma Naïve Bayes, dan evaluasi model. Kerangka penelitian disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Pada tahap preprocessing dilakukan pemeriksaan missing value, data duplikat, tipe data, dan outlier. Hasil pemeriksaan menunjukkan tidak terdapat missing value maupun data duplikat. Deteksi outlier menghasilkan 10 data atau sekitar 0,59% dari 1.700 data. Data tersebut masih berada pada rentang yang logis sehingga tetap digunakan. Selanjutnya, atribut numerik BMI, Alcohol Consumption, Physical Activity, dan Liver Function Test dinormalisasi menggunakan Min-Max. Normalisasi Min-Max dilakukan untuk mengubah nilai atribut ke dalam rentang 0 sampai 1. Persamaan normalisasi yang digunakan adalah:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Sumber: (Tambunan et al., 2026)

Seleksi fitur dilakukan menggunakan Weight by Information untuk menghitung bobot setiap atribut terhadap target Diagnosis. Hasil pembobotan digunakan untuk mengidentifikasi atribut yang paling berpengaruh sebelum proses pemodelan. Data kemudian dibagi menjadi data training dan testing dengan empat skenario, yaitu 40:60, 70:30, 80:20, dan 90:10. Model Naïve Bayes dibangun menggunakan data training dan diterapkan pada data testing melalui Apply Model. Perhitungan probabilitas pada algoritma Naïve Bayes didasarkan pada Teorema Bayes sebagai berikut:

Persamaan dasar Teorema Bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)}$$

Sumber: (Damayanti & Testiana, 2023)

Evaluasi dilakukan menggunakan Accuracy, Precision, Recall, F1-Score, dan AUC. Evaluasi kinerja model dilakukan menggunakan Confusion Matrix melalui nilai Accuracy, Precision, Recall, dan F1-Score. Rumus yang digunakan adalah:

1. Rumus *accuracy* adalah:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Sumber: (Hakim et al., 2025)

2. Rumus *precision* adalah:

$$Precision = \frac{TP}{(TP + FP)}$$

Sumber: (Witaradiani et al., 2025)

3. Rumus *recall* adalah:

$$Recall = \frac{TP}{(TP + FN)}$$

Sumber: (Rafi et al., 2025)

4. Rumus F1-score adalah:

$$F1 - Score = \frac{2\ x\ (Precision\ x\ Recall)}{(Precision + \ Recall)}$$

Sumber: (Permatasari & Hardiani, 2025)

HASIL DAN PEMBAHASAN

Berdasarkan hasil pengujian algoritma Naïve Bayes dengan empat skenario pembagian data, diperoleh hasil evaluasi sebagai berikut:

Tabel 1. Hasil pengujian *Naïve Bayes*

Rasio	*Accuracy*	*Precision*	*Recall*	*F1-Score*	AUC
40:60	81,18%	83,15%	82,56%	82,85%	0,871
70:30	80,39%	81,98%	82,56%	82,26%	0,884
80:20	80,88%	83,15%	81,82%	82,48%	0,883
90:10	82,35%	84,78%	82,98%	83,87%	0,895

Berdasarkan hasil pengujian pada Tabel 1, algoritma Naïve Bayes menghasilkan performa yang relatif konsisten pada seluruh skenario pembagian data. Skenario 90:10 memberikan hasil tertinggi pada seluruh metrik, dengan Accuracy 82,35%, Precision 84,78%, Recall 82,98%, F1-Score 83,87%, dan AUC 0,895. Hasil tersebut menunjukkan bahwa proporsi data latih yang lebih besar pada skenario 90:10 memberikan model lebih banyak informasi untuk mempelajari pola dari data pasien, sehingga menghasilkan kinerja klasifikasi yang lebih baik dibandingkan skenario lainnya.

Nilai Precision sebesar 84,78% menunjukkan bahwa model cukup baik dalam menghasilkan prediksi positif yang sesuai, sedangkan Recall sebesar 82,98% menunjukkan kemampuan model dalam mengenali pasien yang termasuk kelas penyakit liver. Keseimbangan kedua nilai tersebut tercermin pada F1-Score sebesar 83,87%. Sementara itu, AUC sebesar 0,895 menunjukkan kemampuan model yang baik dalam membedakan pasien yang terindikasi penyakit liver dan yang tidak. Hasil ini juga mendukung penggunaan Naïve Bayes pada data kesehatan karena karakteristiknya yang sederhana dan berbasis probabilitas memungkinkan model melakukan klasifikasi berdasarkan pola atribut pasien yang tersedia.

Penggunaan Weight by Information dalam penelitian ini turut membantu mengidentifikasi atribut yang memiliki kontribusi terhadap proses prediksi. Liver Function Test memiliki bobot tertinggi, diikuti oleh Alcohol Consumption dan Smoking, sehingga ketiga atribut tersebut memiliki informasi yang lebih besar terhadap kelas Diagnosis pada dataset yang digunakan. Meskipun demikian, hasil penelitian ini masih terbatas pada karakteristik dataset yang digunakan, sehingga performa model belum dapat digeneralisasikan pada seluruh populasi pasien.lainnya.

KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Naïve Bayes mampu digunakan untuk memprediksi penyakit liver dengan performa terbaik pada rasio pembagian data 90:10, dengan nilai Accuracy sebesar 82,35%, Precision 84,78%, Recall 82,98%, F1-Score 83,87%, dan AUC 0,895.

Hasil seleksi fitur menggunakan Weight by Information menunjukkan bahwa Liver Function Test, Alcohol Consumption, dan Smoking merupakan atribut dengan bobot tertinggi pada dataset yang digunakan. Keterbatasan penelitian ini terletak pada jumlah dan keberagaman data yang masih terbatas. Penelitian selanjutnya dapat menggunakan dataset yang lebih besar dan beragam serta membandingkan algoritma klasifikasi lain atau menerapkan optimasi model untuk meningkatkan performa prediksi

DAFTAR REFERENSI

Aldana, S., & Wibowo, J. S. (2024). Penerapan data mining terhadap klasifikasi pasien penderita penyakit liver menggunakan metode K-Nearest Neighbor. *Progresif: Jurnal Ilmiah Komputer, 20*(1), 124–132.

Damayanti, A., & Testiana, G. (2023). Penerapan data mining untuk prediksi penyakit Hepatitis C menggunakan algoritma Naive Bayes. *Jurnal Manajemen Informatika Jayakarta, 3*, 177–186.

Ginting, L. B., Lestari, R. A., Ginting, F. A., Hosada, D., & Harapan, K. (2022). Melaksanakan pelatihan tentang perawatan pasien penyakit hepatitis (liver) di Desa Paluh Sibaji. *Jurnal Mitra Keperawatan dan Kebidanan Prima, 4*(1), 39–43.

Hakim, L., Sobri, A., Sunardi, L., & Nurdiansyah, D. (2025). Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode K-NN. *Jurnal Digital Teknologi Informasi, 7*(2), 14–20.

Ibrahim, T., Wahyudi, I., Pratama, V. J., Sumanto, Budiawan, I., & Pakpahan, R. (2025). Penerapan metode Logistic Regression untuk memprediksi potensi penyakit liver pada pasien. *Jurnal Teknik Informatika dan Teknologi Informasi, 5*(3), 11.

Insan, M. K., Hayati, U., & Nurdiawan, O. (2023). Analisis sentimen aplikasi Brimo pada ulasan pengguna di Google Play menggunakan algoritma Naive Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(1), 478–483.

Permatasari, M., & Hardiani, T. (2025). Implementation of the K-Nearest Neighbor algorithm for low sodium food recommendations in hypertension patients. *Jurnal Sistem Informasi dan Teknologi Informasi, 7*(3), 867–879.

Rafi, R. R., Ramadhani, R. A., & Sanjaya, A. (2025). Pemanfaatan algoritma Cosine Similarity untuk mengkoreksi ujian esai. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(3), 2242–2254.

Riany, A. F., & Testiana, G. (2023). Penerapan data mining untuk klasifikasi penyakit stroke menggunakan algoritma Naïve Bayes. *Jurnal Saintekom: Sains, Teknologi, Komputer, dan Manajemen, 9*(1), 42–54.

Setyawan, A., Fitriani, A., & Rilvani, E. (2025). Klasifikasi kemiskinan di Indonesia dengan Decision Tree menggunakan RapidMiner. *Jurnal Media Akademik (JMA), 3*(7), 1–12.

Sodikin, D. A., Is'ad, E. T., Prayoga, R., & Purwanto, A. N. I. (2024). Analisis metode Decision Tree dan Naive Bayes pada pasien penyakit liver. *Prosiding Seminar Nasional Rekayasa, Sains dan Teknologi, 3*(1), 12–18.

Sofyan, F. M. A., Voutama, A., & Umaidah, Y. (2023). Penerapan algoritma C4.5 untuk prediksi penyakit paru-paru menggunakan RapidMiner. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(2), 1409–1415.

Tambunan, E., Limbeng, Y. B., & Sipayung, S. (2026). Implementasi algoritma K-Means dengan normalisasi Min-Max pada analisis data ketidakbersekolahan anak. *JDMIS: Journal of Data Mining and Information Systems, 4*(1), 26–32. https://doi.org/10.54259/jdmis.v4i1.7064

Witaradiani, A. A. A., Wibawa, I. G. A., & Santika, P. P. (2025). Klasifikasi genre buku berdasarkan sinopsis menggunakan Naïve Bayes dan Logistic Regression. *JNATIA (Jurnal Nasional Teknologi Informasi dan Aplikasinya), 3*(4), 835–844.

World Health Organization. (2025). *Hepatitis*. World Health Organization.