Indonesian Journal of Science, Technology and Humanities
Vol. 2, No. 3, Februari 2025
Hal 261-267
E-ISSN: 3024-997X
P-ISSN: 2988-7976
Site: https://jurnal.intekom.id/index.php/ijstech

Perbandingan Akurasi Algoritma Decision Tree dengan Random Forest dalam Diagnosa Penyakit Hepatitis

Muhammad Riansyah[1], Syaiful Bahri[2], Harry Pratama Fiqna[3]
[1,2,3]Program Studi Pendidikan Teknik Informatika, STKIP Al Maksum, Stabat, Indonesia

Article Info

Article history:

Received Februari 15, 2025
Revised Februari 15, 2025
Accepted Februari 19, 2025

Kata Kunci:

Decision Tree,
Random Forest,
Machine Learning,
Dataset,
Hepatitis

Keywords:

Decision Tree,
Random Forest,
Machine Learning,
Dataset,
Hepatitis

ABSTRAK

Penyakit hepatitis merupakan salah satu masalah kesehatan global yang membutuhkan deteksi dini untuk mencegah komplikasi lebih lanjut. Dalam konteks ini, algoritma machine learning dapat memberikan kontribusi penting untuk meningkatkan akurasi diagnosis. Penelitian ini bertujuan untuk membandingkan tingkat akurasi antara dua algoritma machine learning yang populer, yaitu Decision Tree dan Random Forest, dalam memprediksi penyakit hepatitis. Dataset yang digunakan Kaggle.com. Kedua algoritma dievaluasi menggunakan metrik akurasi, klasifikasi eror, presisi, dan recall, setelah dilakukan pembagian data menjadi 80% untuk pelatihan dan 20% untuk pengujian. Hasil pengujian menunjukkan bahwa algoritma Random Forest memberikan tingkat akurasi yang lebih tinggi 90,32% dibandingkan dengan Decision Tree 80,65%. Selain itu, Random Forest juga lebih efektif dalam menangani data yang tidak seimbang, yang sering ditemukan dalam diagnosis penyakit hepatitis. Meskipun Decision Tree memiliki keunggulan dalam interpretasi model, hasil penelitian ini menunjukkan bahwa Random Forest lebih unggul dalam meningkatkan akurasi prediksi.

ABSTRACT

Hepatitis is a global health problem that requires early detection to prevent further complications. In this context, machine learning algorithms can make an important contribution to improving diagnostic accuracy. This research aims to compare the level of accuracy between two popular machine learning algorithms, namely Decision Tree and Random Forest, in predicting hepatitis. Dataset used by Kaggle.com. The second algorithm was evaluated using measurements of accuracy, classification error, precision and recall, after dividing the data into 80% for training and 20% for testing. The test results show that the Random Forest algorithm provides a higher accuracy rate of 90.32% compared to Decision Tree 80.65%. Apart from that, Random Forest is also more effective in handling imbalanced data, which is often found in hepatitis diagnosis. Although Decision Trees have advantages in model interpretation, the results of this study show that Random Forest is superior in improving prediction accuracy.

Corresponding Author:

Muhammad Riansyah
Program Studi Pendidikan Teknik Informatika, STKIP Al Maksum
Langkat, Indonesia
Email: riansyahmuhammad88@gmail.com

1. PENDAHULUAN

Hepatitis adalah suatu penyakit yang muncul akibat infeksi dari berbagai jenis virus. Virus-virus tersebut menyerang organ hati manusia, menyebabkan peradangan dan kerusakan pada sel-sel hati. [1] Menurut data dari World Health Organization (WHO), diperkirakan sekitar 325 juta orang di dunia mengidap Hepatitis B atau C, 38 negara menyumbang hampir 80% infeksi dan kematian hepatitis, dan Indonesia termasuk di antara 10 negara dengan beban tertinggi. Prevalensi hepatitis di Indonesia tinggi, dengan 7,1% penduduk terkena Hepatitis B dan 1% terkena Hepatitis C. [2]

Seiring berjalan nya waktu teknologi semakin berkembang. Perkembangan teknologi sudah masuk juga dibergagai sektor kesehatan. Dengan berkembangnya teknologi saat ini, suatu penyakit bisa terdeteksi dengan lebih cepat melalui gejala-gejala penyakit.

Diagnosis adalah proses identifikasi karakteristik suatu penyakit atau kondisi, serta membedakannya dari penyakit atau kondisi lainnya. Penilaian dapat dilakukan dengan cara pemeriksaan fisik, tes laboratorium, atau teknik-teknik serupa lainnya. Selain itu, proses ini juga dapat didukung oleh "melalui program komputer yang dikembangkan untuk memperbaiki proses pengambilan keputusan. [3] Data mining merupakan proses untuk mengidentifikasi informasi atau pengetahuan yang tersembunyi bermanfaat dari kumpulan data berskala besar. Data mining juga merupakan bagian dari proses Knowledge Discovery in Databases (KDD), yang mencakup beberapa tahapan, seperti pemilihan data, pra-pemrosesan, transformasi, penerapan teknik data mining, dan interpretasi hasil. [4]

Klasifikasi merupakan proses pengelompokkan data ke dalam kategori tertentu dengan tujuan memprediksi label dari objek yang belum diketahui sebelumnya. Proses ini memungkinkan perbedaan antara satu objek dengan objek lainnya berdasarkan atribut atau fitur yang dimilikinya.

Terdapat dua jenis klasifikasi, yaitu *supervised learning* dan *unsupervised learning*. Dalam *supervised learning*, algoritma yang umum digunakan meliputi *Decision Tree*, *Naïve Bayes Classifier*, dan *K-Nearest Neighbors* (K-NN). Sementara itu, dalam *unsupervised learning*, algoritma yang sering diterapkan adalah *K-Means* dan *Hierarchical Clustering*. Dalam menyelesaikan suatu komputasi menggunakan teknik klasifikasi tentunya terdapat beragam algoritma yang dapat digunakan, antara lain ialah algoritma Naïve Bayes, K-Nearst Neighbor (KNN), Support Vector Machine (SVM) namun algoritma yang tergolong cukup populer dalam menangani kasus klasifikasi data ialah Decision Tree. Algoritma *Decision Tree* adalah salah satu metode klasifikasi berbasis pohon keputusan yang sering digunakan, terutama karena memiliki keunggulan utama dibandingkan dengan algoritma-algoritma lainnya.[5]

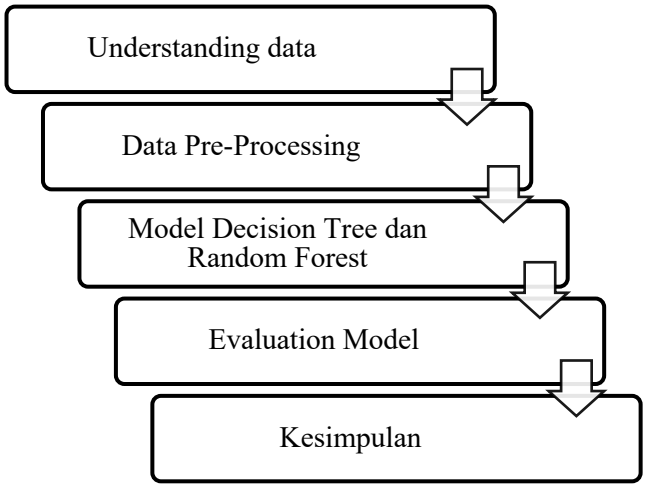
Decision tree mampu menyederhanakan keputusan yang kompleks dengan menguraikannya menjadi lebih sederhana. Selain itu, decision tree mudah dipahami dalam pengolahan data berukuran kecil tanpa mengurangi kualitas hasil yang diperoleh, karena setiap node menggunakan kriteria tertentu dalam proses pengambilan keputusan. [6]

Random forest adalah teknik klasifikasi yang dikembangkan dengan memanfaatkan metode Decision Tree. Dalam metode ini, pemilihan atribut pada setiap node dilakukan secara acak untuk menentukan klasifikasi. Proses klasifikasi ini mengandalkan suara terbanyak dari hasil yang dikembalikan oleh pohon-pohon keputusan. Random forest dapat dibangun dengan memanfaatkan teknik bagging dan pemilihan atribut secara acak. Proses pertumbuhan pohon keputusan dilakukan menggunakan metode CART (Classification and Regression Tree). Pohon-pohon ini dibiarkan tumbuh

hingga mencapai ukuran maksimal tanpa dilakukan pemangkasan. Hasilnya adalah sekumpulan pohon keputusan yang kemudian disebut sebagai hutan (Forest). [7].
Dalam penelitian ini, peneliti mencoba membandingkan tingkat akurasi antara dua algoritma machine learning yang populer, yaitu Decision Tree dan Random Forest, dalam memprediksi penyakit hepatitis.

2. METODE

Tahapan metode yang diajukan dalam penelitian ini dapat dilihat pada gambaran umum penelitian yang disajikan dalam Gambar 1.



Gambar 1. Gambaran umum penelitian

Penjelasan mengenai Gambaran Umum Penelitian pada Gambar 1. sebagai berikut:
1. *Data Understanding* merupakan tahap pengumpulan dan penelaahan data guna memahami data yang akan digunakan. Selain itu, tahap ini mencakup identifikasi masalah dengan memahami substansi dalam data serta mencari aspek menarik yang dapat digunakan untuk merumuskan hipotesis awal [8]. Dalam penelitian ini, data diperoleh melalui *Kaggle.com.* dataset Hepatitis terdiri dari 155 data dan 20 atribut data, dengan dukungan peangkat lunak Rapid Miner.
2. Dalam penelitian ini, proses *data pre-processing* dilakukan melalui beberapa teknik. Tahap pertama adalah data cleansing, yang bertujuan untuk menganalisis kualitas data. Proses ini mencakup agregasi data, identifikasi *missing values*, penghapusan data duplikat, serta *data imputation* [9]. Tahap kedua adalah Label Encoding, yaitu proses mengubah label pada data kategorial menjadi bilangan bulat unik berdasarkan urutan abjad, menggunakan teknik pengkodean tertentu [10]. Tahap ketiga adalah Feature Selection, yaitu proses pemilihan subset fitur dari keseluruhan fitur dalam dataset untuk meningkatkan efisiensi dan akurasi model [11].
3. Setelah melalui tahap *pre-processing*, langkah berikutnya adalah membangun model prediksi menggunakan algoritma Decision Tree, dan Random Forest.
4. Pada tahap ini, evaluasi performa model algoritma yang telah diterapkan dalam metode pembelajaran klasifikasi dilakukan menggunakan algoritma Decision Tree, dan Random Forest. Penilaian performa model klasifikasi didasarkan pada jumlah prediksi yang benar dan salah dalam mengklasifikasikan objek. Dalam penelitian ini, evaluasi model menggunakan confusion matrix, yang menyajikan hasil klasifikasi aktual dan prediksi model [12]. Tabel 1 menampilkan *confusion matrix* untuk klasifikasi dua kelas.

Tabel 1. *Confusion matrix*

Actual Class	Predicted Class	
	Predicted. Class 1	Predicted. Class 0
Actual. Class 1	TP	FN
Actual. Class 0	FP	TN

Pada Tabel 1, TP (*True Positive*) mengacu pada jumlah prediksi positif yang benar, TN (*True Negative*) adalah jumlah prediksi negatif yang benar, FP (*False Positive*) merupakan jumlah prediksi positif yang keliru, dan FN (*False Negative*) adalah jumlah prediksi negatif yang keliru. Dalam penelitian ini, metrik yang digunakan untuk mengevaluasi performa klasifikasi meliputi accuracy, *Classifikasi Error*, precision, recall,. Persamaan (1), (2), (3), dan (4) secara berturut-turut menunjukkan rumus perhitungan untuk accuracy, precision, dan *Classifikasi Error* [12].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Classificasion Error} = \frac{FP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} = \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Data yang digunakan dalam penelitian ini adalah data hepatitis dari https://www.kaggle.com, jumlah atribut dataset hepatitis 20 (*Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver_Big, Liver_Firm, Spleen_Palpable, Spiders, Ascites, Varices, Bilirubin, Alk_Phosphate, Sgot, Albumin, Protime, Histology, Class*), jumlah instances adalah 155 data, salah satu atribut data set ini adalah label class yaitu live dan Die. Data hepatitis dibagi dua, yaitu, 80% menjadi data training dan 20% menjadi data testing.

Tabel 2. Dataset hepatitis

Data Set	Atribut	Type	Kelas	Total Data
Hepatitis	20	*Int, Float, Bool*	2	155

3.2 Hasil Perhitungsn *Confusion Matrix*

Berikut perhitungan table confusion matri untuk mengevaluasi tingkat akurasi hasil klasifikasi yang dihasilkan oleh algoritma Decision TreeTabel 3. Hasil *confusion matrix* algoritma Decision Tree

	true live	true die	class precision
pred. live	22	3	88.00%
pred. die	3	3	50.00%
class recall	88.00%	50.00%	

a. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ = 80,65%

b. $Classifikasi\ Error = \frac{FP+FN}{TP+TN+FP+FN} = 19{,}35\%$

c. $Precision = \frac{TP}{TP+FP} = 88\%$

d. $Recall = \frac{TP}{TP+FN} = 88\%$

Tabel 4. Hasil *confusion matrix* algoritma Random Forest

	true live	true die	class precision
pred. live	25	3	89.29%
pred. die	0	3	100.00%
class recall	100.00%	50.00%	

a. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 90{,}32\%$

b. $Classifikasi\ Error = \frac{FP+FN}{TP+TN+FP+FN} = 9{,}68\%$

c. $Precision = \frac{TP}{TP+FP} = 100\%$

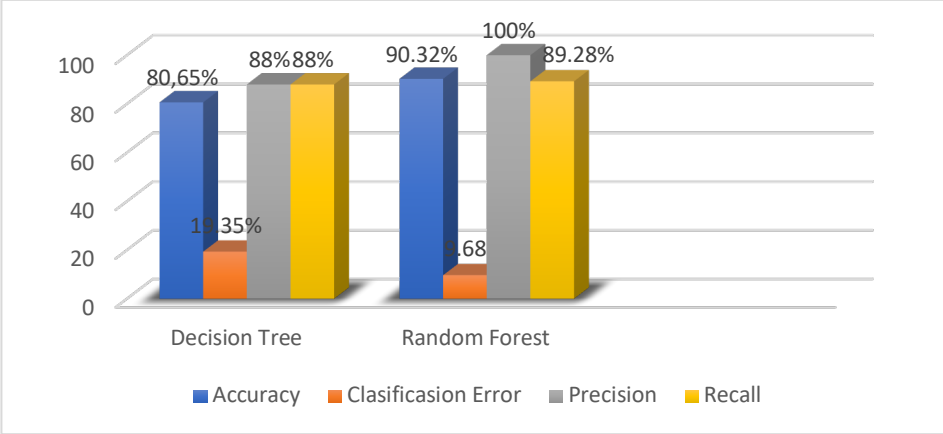
d. $Recall = \frac{TP}{TP+FN} = 89{,}29\%$

3.3 Hasil Akurasi

Algoritma *Decision Tree* memiliki akurasi sebesar 80,65% dengan *classification error* 19,35%, *Precision* 88%, *Recall 88%*. Sementara itu, algoritma Random Forest menampilkan kinerja yang lebih unggul dengan tingkat akurasi yang lebih tinggi 90,32% dan *classification error* 9,68%, *Precision* 100%, *Recall* 89,28%. Hal ini mengindikasikan bahwa *Random Forest* lebih efektif dalam menangani variabilitas data dan menghasilkan prediksi yang lebih akurat. Keunggulan Random Forest terletak pada kemampuannya dalam menerapkan ensemble learning, di mana beberapa pohon keputusan digabungkan untuk memperoleh prediksi yang lebih stabil. Sebaliknya, Decision Tree cenderung lebih rentan terhadap overfitting, terutama pada dataset dengan variabilitas tinggi.

Tabel 5. Hasil Perbandingan algoritma *Decition Tree* dengan *Random Forest*

	Dataset Hepatitis	
	Decision Tree	Random Forest
Accuracy	80,65%	90,32%
Clasificasion Error	19,35%	9,68%
Precision	88%	100%
Recall	88%	89,28%



Gambar 2. Grafik Tingkat Akurasi

4. KESIMPULAN

Penelitian ini telah berhasil mengevaluasi dan membandingkan kinerja dari algoritma Decision Tree dan Random Forest dala mengklasifikasikan penyakit Hepatitis dengan menggunakan dataset dari *Kaggle.com*. Hasil penelitian menunjukkan bahwa *Random Forest* memiliki performa yang lebih baik dibandingkan *Decision Tree*, terutama dalam aspek akurasi dan nilai error. Algoritma *Random Forest* mencapai akurasi 90,32% dengan klasifikasi eror 9,68%, sedangkan *Decision Tree* hanya memperoleh akurasi 80,65**%** dengan klasifikasi eror 19,35%.

Hasil ini membuktikan bahwa *Random Forest* lebih efektif dalam mengidentifikasi kasus positif serta mampu mengurangi jumlah prediksi positif yang keliru, yang sangat penting dalam diagnosis penyakit. Keunggulan *Random Forest* berasal dari penerapan teknik *ensemble learning*, di mana beberapa pohon keputusan digabungkan untuk meningkatkan stabilitas dan ketepatan prediksi. Selain itu, algoritma ini lebih mampu menghadapi variabilitas data dan mengurangi risiko *overfitting* dibandingkan dengan *Decision Tree*.

Penelitian ini memberikan wawasan berharga bagi praktisi medis dan peneliti dalam memilih metode klasifikasi yang tepat untuk diagnosis penyakit, terutama dalam kondisi di mana akurasi sangat penting. Mengingat tingginya angka kejadian penyakit tiroid serta dampaknya terhadap kesehatan masyarakat, penerapan algoritma *machine learning* seperti *Random Forest* dapat berkontribusi secara signifikan dalam deteksi dini serta pengobatan yang lebih optimal.

Selain itu, penelitian ini menekankan pentingnya pemanfaatan teknik analisis data yang lebih maju dalam bidang kesehatan untuk meningkatkan hasil klinis. Sebagai rekomendasi untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan beragam, serta mengeksplorasi algoritma lain yang berpotensi meningkatkan akurasi serta keandalan dalam klasifikasi penyakit tiroid.

REFERENSI

[1] H. Syahputra, and D. M. Syafindy, "Sistem Pakar Diagnosa Penyakit Hepatitis Dengan Menggunakan Metode Certainty Factor. *Jurnal Sains Informatika Terapan*, vol. 2. no. 1. pp. 45-50, Feb. 2023, doi: 10.62357/jsit.v2i1.186

[2] T. R. P. Lestari, "Kolaborasi Global Dalam Penanganan Hepatitis: Posisi Dan Peran Indonesia," *Pusat Analisis Keparlemenan Badan Keahlian DPR RI*., vol. XVI, no. 14, II, pp. 21-25, Juli. 2024.

[3] J. Adler, "Diagnosa Penyakit dengan Gejala Demam pada Manusia Berbasis Mobile: Knowledge Based System" Komputika: Jurnal Sistem Komputer, Vol. 6, No. 2, pp. 51-58, 2017, doi: 10.34010/komputika.v6i2.1607

[4] Ihsan. Reduksi Atribut Pada Algoritma K-Nearest Neighbor (KNN) Dengan Menggunakan Algoritma Genetika. Tesis., Universitas Sumatera Utara., Medan, 2018.

[5] Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2nd ed. Athens, Greece: Academic Press, 2020.

[6] Hendra., Azis, M.A. and Suhardjono, "Analisis Prediksi Kelulusan Mahasiswa Menggunakan Decission Tree Berbasis Particle Swarm Optimization" *Jurnal SISFOKOM (Sistem Informasi dan Komputer),* pp. 102-107, 2022.
[7] L. Ratnawati and D. R. Sulistyaningrum. "Penerapan random forest untuk mengukur tingkat keparahan penyakit pada daun apel". *Jurnal Sains dan Seni ITS*, *8*(2), pp. A71-A77. 2020.
[8] R. Ordila, R. Wahyuni, Y. Irawan, dan M. Y. Sari, "Penerapan Data Mining untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit dengan Algoritma Clustering (Studi Kasus: Poli Klinik Pt. Inecda)," *Jurnal Ilmu Komputer*, vol. 9, no. 2, pp. 148–153, 2020.
[9] D. Darwis, "Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional," *Jurnal Tekno Kompak*, vol. 15, no. 1, pp. 131-145, 2021.
[10] F. Ardiansyah. "Sistem Prediksi Harga Sewa Kost Dengan Menggunakan Random Forest Analytics (Studi Kasus: Kost Eksklusif di Daerah Istimewa Yogyakarta)," Tugas Akhir., Universitas Islam Indonesia Yogyakarta., Yogyakarta, 2020.
[11] A. Rahmansyah *et al.*, "Membandingkan Pengaruh Feature Selection Terhadap Algoritma Naïve Bayes dan Support Vector Machine," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, 2018.
[12] M. F. Naufal, "Analisis Perbandingan Algoritma SVM, KNN dan CNN untuk Klasifikasi Citra Cuaca," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 8, no. 2, pp. 311-317, 2021.