



Perbandingan Tingkat Akurasi Algoritma *Decision Tree* dan *Random Forest* dalam Klasifikasi pada Dataset Heterogen

Chan Uswatun Khasanah

Politeknik Akbara Surakarta

e-mail: chanuswatunkhasanah@gmail.com

Abstract

Decision Tree and Random Forest are two widely used classification algorithms in data mining. However, most previous studies have evaluated these algorithms using a single dataset, making their findings potentially less generalizable across different data characteristics. This study compares the accuracy of Decision Tree and Random Forest using three datasets with distinct characteristics, namely Iris, which is relatively small and simple; Breast Cancer Wisconsin, which contains a larger number of attributes; and Telco Customer Churn, which has a larger data volume and an imbalanced class distribution. The results show that Random Forest achieves higher accuracy on the Breast Cancer Wisconsin dataset (96.49% compared to 92.40%) and the Telco Customer Churn dataset (79.79% compared to 75.77%), whereas Decision Tree performs better on the Iris dataset (93.33% compared to 91.11%). These findings indicate that the performance of the two algorithms is data-dependent, with the advantage of Random Forest becoming more evident when applied to datasets with greater complexity. The results of this study can serve as a reference for selecting an appropriate classification algorithm based on the characteristics of the dataset.

Keywords: Accuracy, Decision Tree, Random Forest, Heterogeneous Datasets.

Abstrak

Decision Tree dan Random Forest merupakan dua algoritma klasifikasi yang banyak digunakan dalam data mining, namun sebagian besar penelitian terdahulu hanya menguji keduanya pada satu dataset, sehingga kesimpulannya belum tentu berlaku umum. Penelitian ini membandingkan akurasi Decision Tree dan Random Forest pada tiga dataset dengan karakteristik berbeda, yaitu Iris (kecil dan sederhana), Breast Cancer Wisconsin (atribut lebih banyak), dan Telco Customer Churn (volume besar dan tidak seimbang). Hasil menunjukkan Random Forest lebih unggul pada Breast Cancer Wisconsin (96,49% dan 92,40%) dan Telco Customer Churn (79,79% dan 75,77%), sementara Decision Tree lebih unggul pada Iris (93,33% dan 91,11%). Temuan ini menunjukkan bahwa kinerja kedua algoritma bersifat data-dependent, di mana keunggulan Random Forest baru terlihat signifikan pada data yang lebih kompleks. Hasil penelitian dapat menjadi acuan pemilihan algoritma klasifikasi sesuai karakteristik data.

Kata Kunci: Akurasi, Decision Tree, Random Forest, Dataset Heterogen.

PENDAHULUAN

Ketersediaan data dalam jumlah besar menuntut adanya metode analisis yang mampu mengekstraksi pola dan pengetahuan tersembunyi di dalamnya secara efektif sehingga dapat menghasilkan informasi yang bernilai dan mendukung proses pengambilan keputusan. Salah satu pendekatan yang dapat digunakan adalah teknik klasifikasi menggunakan algoritma Decision Tree dan Radom Forest yang mampu mengidentifikasi pola berdasarkan karakteristik data untuk menghasilkan prediksi terhadap kelas tertentu.

Penelitian terdahulu yang membandingkan algoritma Decision Tree dan Random Forest sudah banyak dilakukan. Random Forest menunjukkan nilai akurasi yang lebih unggul dibandingkan Decision Tree dalam melakukan klasifikasi nutrisi pada makanan cepat saji, yaitu 66.67% (Yaman *et al.*, 2024). Random Forest juga lebih unggul dalam melakukan klasifikasi terhadap penyakit kanker (kanker tiroid, usus besar, dan paru-paru), yaitu dengan akurasi 99% (Oktavianto *et al.*, 2024). Penelitian dengan beragam topik dan dataset, seperti deteksi dini kesehatan mental mahasiswa (Jayadi *et al.*, 2026), klasifikasi tingkat stres mahasiswa (Abdah Syakiroh Gustian and Fathoni Mahardika, 2025), diagnosa penyakit hepatitis (Muhammad Riansyah, Syaiful Bahri and Harry Pratama Fiqna, 2025), prediksi kelulusan mahasiswa (Yuspita, Okra and Rezeki, 2025), prediksi perdarahan pascasalin (Sinambela *et al.*, 2023), klasifikasi penjualan produk (Putri Ayu Firnanda *et al.*, 2025), klasifikasi kepuasan masyarakat (Wicaksono and Nuryana, 2024), dan prediksi penyakit jantung (Alfajr and Defiyanti, 2024) juga menunjukkan bahwa akurasi Random Forest lebih unggul dibandingkan Decision Tree.

Penelitian ini berfokus pada penggunaan tiga dataset yang memiliki karakteristik berbeda-beda, dimulai dari dataset sederhana hingga dataset yang kompleks. Hasil penelitian ini diharapkan dapat memberikan pengetahuan mengenai karakteristik kinerja kedua algoritma pada berbagai macam kondisi dataset sehingga dapat menjadi acuan dalam pemilihan algoritma klasifikasi yang sesuai dengan kebutuhan.

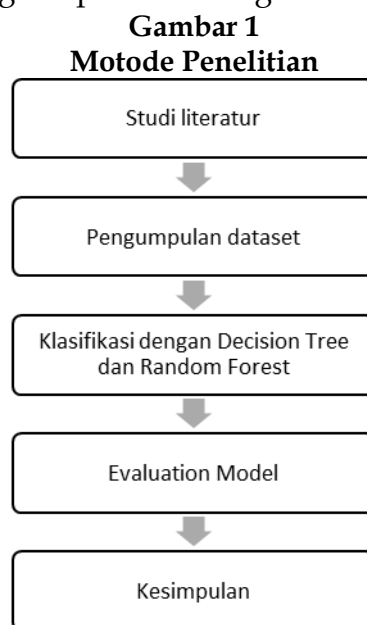
Decision Tree merupakan algoritma supervised machine learning yang digunakan untuk memecahkan masalah klasifikasi, dengan tujuan menghasilkan model prediksi dalam bentuk aturan yang mudah diimplementasikan (Katriani and Mailoa, 2022). Algoritma ini direpresentasikan dalam bentuk struktur pohon, di mana simpul pada pohon dibedakan menjadi tiga jenis, yaitu simpul akar (root node), simpul percabangan/internal (branch/internal node), dan simpul daun (leaf node). Secara lebih rinci, setiap simpul internal berisi pengujian terhadap atribut tertentu, setiap cabang menggambarkan hasil dari pengujian tersebut, sementara simpul daun menyimpan informasi label kelas, dengan simpul

teratas dalam struktur pohon disebut sebagai simpul akar. Pemilihan atribut sebagai pemisah pada setiap simpul ditentukan berdasarkan kriteria seperti Gini impurity atau entropi yang menilai tingkat ketidakhomogenan atau ketidakpastian pada setiap simpul, sehingga proses klasifikasi berlangsung secara bertahap hingga diperoleh kelas prediksi pada simpul daun.

Random Forest adalah salah satu algoritma populer dalam machine learning dan data mining yang digunakan untuk melakukan prediksi baik dalam kasus klasifikasi maupun regresi, dikenal karena tingkat akurasinya yang tinggi serta kemampuannya dalam menangani data yang kompleks (Hazizah, Sharipuddin and Feranika, 2025). Algoritma ini termasuk kategori pembelajaran ensemble dan menggunakan beberapa model sederhana untuk membuat prediksi yang lebih baik secara keseluruhan, dengan menggabungkan bagging dan pemilihan atribut acak. Random Forest menggunakan konsep yang sama dengan Decision Tree, kemudian melakukan agregasi hasil dari setiap pohon keputusan untuk memperoleh hasil klasifikasi, dan bekerja dengan menggabungkan banyak pohon keputusan serta melakukan klasifikasi berdasarkan voting mayoritas. Pendekatan ensemble ini mampu mengurangi risiko overfitting melalui kombinasi pohon keputusan yang dilatih pada subset data yang berbeda.

METODE PENELITIAN

Tahapan atau langkah-langkah penelitian digambarkan pada Gambar 1.



Sumber: Data diolah, 2026

Studi literatur dilakukan dengan mencari penelitian terdahulu dengan topik yang serupa untuk menemukan kebaruan dari penelitian yang akan

dilakukan. Pengumpulan dataset dilakukan untuk mendapatkan dataset yang sesuai untuk penelitian ini. Dataset yang digunakan diperoleh dari sumber dataset publik, yaitu Kaggle. Proses klasifikasi dilakukan menggunakan 2 algoritma, yaitu Decision Tree dan Random Forest. Penggunaan dua algoritma ini bertujuan untuk mengetahui kemampuan masing-masing algoritma dalam melakukan klasifikasi dengan dataset yang heterogen. Model Decision Tree dan Random Forest kemudian dievaluasi menggunakan data pengujian (data testing). Evaluasi menggunakan confusion matrix. Hasil evaluasi kedua algoritma tersebut kemudian dibandingkan untuk menentukan model yang memberikan performa klasifikasi terbaik. Tahap terakhir adalah penarikan kesimpulan. Kesimpulan diperoleh berdasarkan hasil pengujian dan perbandingan kinerja Decision Tree dan Random Forest.

Dataset yang digunakan merupakan dataset yang sudah tersedia di internet, dengan rincian sebagai berikut.

Tabel 1
Dataset

Dataset	Karakteristik Pembeda	Alasan Dipilih
Iris (Fisher, 1936)	Kecil, seimbang, atribut sedikit	Baseline (kondisi ideal/ sederhana)
Breast Cancer Wisconsin (William <i>et al.</i> , 1993)	Menengah, atribut banyak (30)	Dimensi data lebih tinggi
Telco Customer Churn (IBM, 2018)	Besar, tidak seimbang, campuran tipe data	Kondisi data kompleks

Sumber: Data diolah, 2026

PEMBAHASAN

Deskripsi Dataset

Penelitian ini menggunakan 3 dataset yang berbeda, yaitu sebagai berikut.

a. Iris Dataset

Dataset tentang bunga Iris yang digunakan untuk mengklasifikasikan jenis/spesies bunga berdasarkan ukuran bagian bunga.

b. Breast Cancer Wisconsin Dataset

Dataset yang digunakan untuk klasifikasi tumor payudara (jinak atau ganas).

c. Telco Customer Churn Dataset

Dataset yang digunakan untuk mengklasifikasikan sikap customer pada perusahaan telekomunikasi (berlangganan atau tidak).

Karakteristik setiap dataset ditunjukkan pada Tabel 2.

Tabel 2
Karakteristik Dataset

Dataset	Jumlah Data	Jumlah Atribut	Jumlah Kelas
Iris	150	4	3
Breast Cancer Wisconsin	569	30	2
Telco Customer Churn	7043	20	2

Sumber: Data diolah, 2026

Hasil Perhitungan Confusion Matrix

Berikut ini adalah hasil confusion matrix pada algoritma Decision Tree dan Random Forest di setiap dataset yang ditunjukkan pada Tabel 3-8.

Tabel 3
Decision Tree pada Dataset Iris

accuracy: 93.33%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	15	0	0	100.00%
pred. Iris-versicolor	0	15	3	83.33%
pred. Iris-virginica	0	0	12	100.00%
class recall	100.00%	100.00%	80.00%	

Sumber: Data diolah, 2026

Tabel 4
Random Forest pada Dataset Iris

accuracy: 91.11%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	15	0	0	100.00%
pred. Iris-versicolor	0	15	4	78.95%
pred. Iris-virginica	0	0	11	100.00%
class recall	100.00%	100.00%	73.33%	

Sumber: Data diolah, 2026

Tabel 5
Decision Tree pada Dataset Breast Cancer Wisconsin

accuracy: 92.40%

	true M	true B	class precision
pred. M	58	7	89.23%
pred. B	6	100	94.34%
class recall	90.62%	93.46%	

Sumber: Data diolah, 2026

Tabel 6
Random Forest pada Dataset Breast Cancer Wisconsin

accuracy: 96.49%

	true M	true B	class precision
pred. M	60	2	96.77%
pred. B	4	105	96.33%
class recall	93.75%	98.13%	

Sumber: Data diolah, 2026

Tabel 7
Decision Tree pada Dataset Telco Customer Churn

accuracy: 75.77%

	true No	true Yes	class precision
pred. No	1230	190	86.62%
pred. Yes	322	371	53.54%
class recall	79.25%	66.13%	

Sumber: Data diolah, 2026

Tabel 8
Random Forest pada Dataset Telco Customer Churn

accuracy: 79.79%

	true No	true Yes	class precision
pred. No	1406	281	83.34%
pred. Yes	146	280	65.73%
class recall	90.59%	49.91%	

Sumber: Data diolah, 2026

Hasil Akurasi

Berdasarkan hasil pengujian yang ditunjukkan pada Tabel 3-8, diperoleh nilai akurasi yang bervariasi antara algoritma Decision Tree dan Random Forest pada ketiga dataset yang digunakan (Tabel 9). Pada dataset Iris, algoritma Decision Tree memperoleh akurasi sebesar 93,33%, lebih tinggi dibandingkan Random Forest yang hanya mencapai 91,11%, dengan selisih 2,22%. Hasil ini menunjukkan bahwa pada dataset berukuran kecil dengan jumlah atribut sedikit dan kelas yang seimbang seperti Iris, satu pohon keputusan tunggal sudah cukup mampu melakukan klasifikasi dengan baik, sehingga algoritma Random Forest justru berpotensi menurunkan akurasi akibat variasi antar-pohon yang tidak diperlukan pada dataset yang sederhana.

Pada dataset Breast Cancer Wisconsin, Random Forest menunjukkan keunggulan signifikan dengan akurasi 96,49%, dibandingkan Decision Tree

yang hanya 92,40%, dengan selisih 4,09%. Hal ini sejalan dengan karakteristik dataset ini yang memiliki jumlah atribut jauh lebih banyak (30 atribut) dibandingkan Iris. Pada kondisi dataset dengan karakteristik yang lebih kompleks, Random Forest lebih mampu mengatasinya.

Pada dataset Telco Customer Churn, Random Forest kembali unggul dengan akurasi 79,79% dibandingkan Decision Tree 75,77%, dengan selisih 4,02%. Meskipun keunggulan Random Forest tetap konsisten seperti pada Breast Cancer Wisconsin, akurasi kedua algoritma pada dataset ini tergolong paling rendah di antara ketiga dataset yang diuji. Hal ini dipengaruhi oleh jumlah data yang jauh lebih besar (7.043 data) serta sifat data Telco Customer Churn yang cenderung tidak seimbang antar kelas (jumlah pelanggan yang berhenti berlangganan/churn lebih sedikit dibandingkan yang tidak churn), sehingga kedua algoritma lebih sulit mengenali pola secara akurat dibandingkan pada dataset dengan distribusi kelas yang lebih seimbang seperti Iris dan Breast Cancer Wisconsin.

Tabel 9
Hasil Akurasi pada Algoritma Decision Tree dan Random Forest

Dataset	Accuracy	
	Decision Tree	Random Forest
Dataset Iris	93.33%	91.11%
Dataset Breast Cancer Wisconsin	92.40%	96.49%
Dataset Telco Customer Churn	75.77%	79.79%

Sumber: Data diolah, 2026

Dari tiga dataset yang diuji, Random Forest unggul secara konsisten pada dua dataset dengan jumlah atribut dan/atau volume data yang lebih besar (Breast Cancer Wisconsin dan Telco Customer Churn), sementara Decision Tree unggul pada dataset yang berukuran kecil dan sederhana (Iris). Hal ini menunjukkan bahwa kinerja kedua algoritma bersifat data-dependent, yaitu keunggulan pendekatan ensemble learning pada Random Forest baru terlihat signifikan seiring meningkatnya kompleksitas data, baik dari sisi jumlah atribut maupun volume data, sedangkan pada data yang relatif sederhana, Decision Tree sudah cukup baik dalam melakukan klasifikasi.

KESIMPULAN

- Random Forest unggul secara konsisten pada dataset dengan atribut dan/atau volume data lebih besar (Breast Cancer Wisconsin dan Telco Customer Churn), sedangkan Decision Tree lebih unggul pada dataset kecil dan sederhana (Iris), menunjukkan bahwa kinerja kedua algoritma bersifat data-dependent.

- b. Decision Tree lebih cocok digunakan pada data berskala kecil dan sederhana karena efisien dan mudah diinterpretasikan, sementara Random Forest lebih direkomendasikan untuk data berskala besar atau kompleks karena menghasilkan akurasi yang lebih tinggi dan stabil.
- c. Penelitian selanjutnya dapat menambahkan jumlah dan variasi dataset dengan karakteristik yang lebih beragam untuk memperkuat generalisasi temuan mengenai hubungan antara karakteristik data dan kinerja algoritma.

DAFTAR PUSTAKA

- Abdah Syakiroh Gustian and Fathoni Mahardika (2025) "Analisis Klasifikasi Risiko Dropout Mahasiswa Menggunakan Algoritma Decision Tree dan Random Forest," *Jupiter: Publikasi Ilmu Keteknikan Industri, Teknik Elektro dan Informatika*, 3(4). Available at: <https://doi.org/10.61132/jupiter.v3i4.980>.
- Alfajr, N.H. and Defiyanti, S. (2024) "Prediksi Penyakit Jantung Menggunakan Metode Random Forest Dan Penerapan Principal Component Analysis (PCA)," *Jurnal Informatika dan Teknik Elektro Terapan*, 12(3S1). Available at: <https://doi.org/10.23960/jitet.v12i3s1.5055>.
- Fisher, R. (1936) *Iris*, *UCI Machine Learning Repository*. Available at: <https://doi.org/10.24432/C56C76> (Accessed: August 11, 2026).
- Hazizah, N., Sharipuddin and Feranika, A. (2025) "Implementasi Algoritma Random Forest Dalam Klasifikasi Risiko Gagal Bayar Kartu Kredit Pada Nasabah Bank," *Jurnal Manajemen Teknologi Dan Sistem Informasi (JMS)*, 5(1). Available at: <https://doi.org/10.33998/jms.2025.5.1.2384>.
- IBM (2018) *Telco Customer Churn*, *Kaggle*. Available at: <https://www.kaggle.com/datasets/blaschar/telco-customer-churn> (Accessed: August 11, 2026).
- Jayadi, J. *et al.* (2026) "Deteksi Dini Kesehatan Mental Mahasiswa Dengan Machine Learning: Perbandingan Algoritma Decision Tree Dan Random Forest," *Technologia: Jurnal Ilmiah*, 17(1). Available at: <https://doi.org/10.31602/tji.v17i1.21251>.
- Katriani, N. and Mailoa, E. (2022) "Klasifikasi Bahasa Daerah Menggunakan Decision Tree Dan Gradient Boots," *Jurnal Teknik Informatika dan Sistem Informasi*, 9(2).
- Muhammad Riansyah, Syaiful Bahri and Harry Pratama Fiqna (2025) "Perbandingan Akurasi Algoritma Decision Tree dan Random Forest dalam Diagnosa Penyakit Hepatitis," *Indonesian Journal of Science, Technology and Humanities*, 2(3). Available at: <https://doi.org/10.60076/ijstech.v2i3.1108>.
- Oktavianto, H. *et al.* (2024) "Analisis Komparasi Kinerja Metode Decision Tree dan Random Forest dalam Klasifikasi Teks Data Kesehatan," *Bina Insani ICT Journal*, 11(1).
- Putri Ayu Firnanda *et al.* (2025) "Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket,"

- Emerging Statistics and Data Science Journal*, 3(1). Available at: <https://doi.org/10.20885/esds.vol3.iss.1.art2>.
- Sinambela, D.P. *et al.* (2023) "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *Jurnal Informasi dan Teknologi*, 5(3). Available at: <https://doi.org/10.60083/jidt.v5i3.393>.
- Wicaksono, S.A. and Nuryana, I.K.D. (2024) "Perbandingan Klasifikasi Hasil Indeks Kepuasan Masyarakat Terhadap Upt Balai Latihan Kerja Surabaya Menggunakan Algoritma Decision Tree, Random Forest, K-Nearest Neighbor," *Journal of Informatics and Computer Science (JINACS)*, 5(04). Available at: <https://doi.org/10.26740/jinacs.v5n04.p552-228>.
- William, W. *et al.* (1993) *Breast Cancer Wisconsin (Diagnostic)*, *UCI Machine Learning Repository*. Available at: <https://doi.org/10.24432/C5DW2B> (Accessed: August 11, 2026).
- Yaman, N.I. *et al.* (2024) "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," *Jurnal Algoritma*, 21(2). Available at: <https://doi.org/10.33364/algoritma/v.21-2.1649>.
- Yuspita, Y.E., Okra, R. and Rezeki, M. (2025) "Penerapan Algoritma Klasifikasi Untuk Prediksi Tingkat Kelulusan Mahasiswa Menggunakan Rappidminer," *Djtechno: Jurnal Teknologi Informasi*, 6(1). Available at: <https://doi.org/10.46576/djtechno.v6i1.6169>.